博士論文

Towards a digital environment for passive exposure to books based on online social media

(SNSを用いた受動的図書遭遇環境の構築に向けて)

矢田　竣太郎

Contents

List of Tables

List of Figures

Part I

Background and research questions

1 Introduction

1.1 Research purpose

This section provides a brief overview of the purpose of this study. We will elaborate on it in succeeding sections in this chapter.

As we will see in Section 1.2.1, international organisations for education consider reading books to be highly important for fostering literacy. These days, the meaning of literacy has come to extend from reading and writing texts to understanding and utilising mathematics, science, and information technology. Since reading books is regarded as the basis of all these, many organisations promote reading. Other than actual action of reading, *passive exposure to books*—for example, the number of household books—plays an important role to improve literacy (in its extended definition) in a hidden manner (see Section 1.2.2) Traditionally, exposure to books has been supported by physical environments such as bookshelves in homes, bookshops, and libraries. Even seeing people reading on the train or in parks can inspire us to read.

Exposure to books in physical environments is decreasing in many developed countries (see Section 1.3.2). The number of local bookshops has been significantly declining in major developed countries and regions worldwide. This leads to fewer opportunities for the casual browsing of books when one passes by local bookshops. We also have fewer chances to visit bookshops incidentally, for example, simply by accompanying friends. Besides, e-books are gaining growing popularity. Unlike physical books, however, e-books remain limited to their owner's devices, so no one else is exposed to them by chance. E-books are read with generic smart devices, which do not show book titles or contents to anyone except the reader, while reading paper books in public can tell others what books are being read.

Online digital environments, the standard entry point of information seeking and consumption, do not support book exposure as well as the physical world (see Section 1.4.1). Search engines and recommendation systems can provide better accessibility to books than physical environments. However, this is only applicable to those who have the proactive intention to find books regularly. Whereas such *frequent* readers can insert themselves into the rich circumstances of book information, it is very hard for ***infrequent*** readers to be exposed to books online. Since infrequent readers show more book-unrelated activities online, their personalised search results and recommendation items tend not to include books, which causes less exposure to books (see Section 1.4.2). Furthermore, digital environments afford few channels for frequent readers to entice infrequent readers into book exposure. For instance, the experience of online bookshops is personalised and not shareable with acquaintances, unlike local bookshops, which we can visit accompanied. E-books have a similar issue of shareability, since a user's library is strongly tethered to their account. Meanwhile, the number of infrequent readers is increasing worldwide (see Section 1.3.2).

Unintentional encounters with books have not been supported well online, while physical environments, which have traditionally provided such encounters, are shrinking. This is not just 'paper book nostalgia'—given the importance of reading books and exposure to books, the current situation could potentially cause a serious educational gap in the future. As online digital environments are becoming ubiquitous, we need *a digital surrogate* for book encounters that take place in the physical world. It can work like a complement to existing reading-promotion programs in the physical world.

We believe that online social media can provide such digital exposure to books, since its sharing functionalities and the absence of explicit information needs of its users make it ready for unexpected encounters (see Section 1.5.1). Social media posts about books can work as an entry point to book information even for infrequent readers, much like random chats in the physical world, which occasionally convey accidental mentions of unfamiliar topics not only to the chat participants but also to passersby. In this context, casual mentions of books are more suitable than book reviews mainly for those who are already interested in books. We can utilise social media to promote better exposure to books in digital environments. A straight-forward way can be proposed: propagating mentions of books in social media.

We are concerned, nevertheless, that the current personalisation algorithms for social media feeds may prohibit infrequent readers from exposure to mentions of books, e.g. since they would have been presenting fewer book-related activities on the platforms. Many studies claim that online social media is a major place to form filter bubbles (see Section 1.5.2). In order to make use of social media for a digital surrogate for physical exposure to books, we need to alleviate the effect of current personalisation techniques.

One solution we propose is building a system to deliver social mentions of books to users' digital online environments, independently from the built-in personalisation algorithms of social media platforms. We should be aware of what types of such book mentions are delivered and how, so as not to annoy users but still be able to attract their attention positively. This problem can be formulated into *how to attract users, including infrequent readers, to mentions of books.* From the user's perspective, it is regarded as the degree to which the user experience (UX) of our system is *inspiring* for users to read the books proposed. In Section 1.6, we name this degree *inspiring-ness* (or *inspiringness*) and introduce its key components in advance of their formal definitions provided in a later chapter (i.e. Chapter 3).

We design the principal architecture of the system: it first collects social mentions of books, then organises them from the perspective of inspiringness, and finally delivers them to users' digital environments. The first two modules comprise non-trivial tasks related to natural language processing (NLP) that require careful research and development. We thus set the goal of this research as implementing *the core NLP modules* towards the realisation of the overall system (see Section 1.7).

The rest of this chapter is organised as follows. From Sections 1.2 to 1.6, we elaborate on the situations and concepts we mentioned above. Section 1.7 clarifies the goal and scope of this research. In Section 1.8, we provide overviews of the remaining chapters.

1.2 Passive exposure to books

This section confirms the importance of books and reading activities. We first summarise the benefit of reading activities, and second pay attention to the importance of passive exposure to books.

1.2.1 The value of reading books

In the context of education, reading books has been regarded as a key component across many regions where books are available. Major international organisations concerned with education recognise its importance. For instance, the United Nations Educational, Scientific and Cultural Organization (UNESCO)'s literacy project attaches great importance to reading books.[1] The Organization for Economic Cooperation and Development (OECD) incorporates reading skills into *global competence*, i.e. the abilities to 'examine local, global and intercultural issues, [to] understand and appreciate different perspectives and world views, [to] interact successfully and respectfully with others, and [to] take responsible action toward sustainability and collective well-being' (OECD, 2018, p. 4).

This recognition is supported by evidence. For instance, reading to children has a significant effect on traditional literacy, i.e. textual reading and writing skills (Bus, van

[1] https://en.unesco.org/themes/literacy

IJzendoorn and Pellegrini, 1995). Wolf (2008), who summarises the neurological mechanism in children's development of literacy, explains that reading (written text) is not an innate human ability and requires enormous efforts of repetitive reading experience in younger ages for children to acquire—their brains need to establish a bridge between linguistic comprehension and image recognition. Books are a good resource for children to read written text with pleasure. While reading to children is one suitable method to expose children to text, its frequency may result in a huge gap in the number of words to which children are exposed. Logan et al. (2019) estimated the difference in the frequency of parents' reading sessions and revealed that a maximum 100 million word gap by the age of five may occur in between children whose parents almost never read to them and children to whom five books are read per day. This difference in reading frequency may well cause a large gap in literacy in the end.

The concept of 'literacy' has been extended to mathematics, science, and information-and-communications technology (ICT) skills. Amongst these skills, it is actually shown that reading for pleasure links to mathematics as well as vocabulary (Sullivan and M. Brown, 2015). The effect on numeracy skills has also been confirmed in a survey report by a Japanese company (Benesse Educational Research and Development Institute, 2018). Other skills of extended literacy have been demonstrated by Sikora, M.D. Evans and Kelley (2019), the details of which will be introduced in the next section (Section 1.2.2).

Other than literacy, it is known that reading books can also provide the following benefits to humans:

- enhancing empathy (Mar, Oatley and Peterson, 2009)
- helping to express one's identity in social interactions (Kaiser and Quandt, 2016)
- supporting inter-cultural understanding, decreasing solitude, and providing insights for life planning and decision making (Clark and Rumbold, 2006)

1.2.2 The value of passive exposure

Studies have shown that passive exposure to books has statistically significant effects on literacy. Passive exposure can enhance reading, which further benefits literacy as we have just seen in the previous section. In addition to this lucid causation, moreover, literacy can also be developed directly through passive exposure. M.D. Evans et al. (2010) studied the effect of the number of books at home towards the level of education, using around 70 thousand cases from 27 nations. They found that children growing up with 500 books at home gain at most three more schooling years than children without books at home, regardless of their parents' incomes and education. This trend was independent from nations and periods too. M.D.R. Evans, Kelley and Sikora (2014) next analysed the data from the

OECD's Programme for International Student Assessment, widely known as PISA, which contains 200 thousand cases of 15-year-old students across 42 nations. As a result, the larger number of books in the family home is an idicator of better academic performance among children,[2] and this effect is stronger than their parents' education, occupational status, and wealth. Furthermore, Sikora, M.D. Evans and Kelley (2019) showed that the number of books in one's household during adolescence has a significant effect (at the $p = 0.01$ level) directly on extended literacy (reading/writing, numeracy, and ICT skills) in later years, regardless of individual education level or reading frequency as an adult. Their path analysis also illustrates that the independent (direct) effect of the number of household books on extended literacy was still statistically significant, even when indirect pathways (e.g. via reading habits) were taken into consideration.

The mechanism of the standalone effect of passive exposure still remains unclear. A study (van Bergen et al., 2017), which hypothesised that parents' reading skills are inherited by children, found that the size of home library had a statistically significant influence on the children's reading skills even after controlling for parents' reading skills. We suppose that one's cognitive skills are developed not only by reading through whole books, but also by brief experiences related to reading, such as leafing through books, reading just one chapter or the table of contents, or talking about books at home. While Wade and Kidd (2019) revealed that learning is driven by one's objective perception of the amount of one's own knowledge (i.e. a self assessment of what we know and what we do not), accessible shelves of books available for browsing could afford such perspectives of the boundary of one's own present knowledge.

These studies above were inspired by the *scholarly culture* theory: reading activities develop cognitive skills, and furthermore, educational achievements (Bus, van IJzendoorn and Pellegrini, 1995; Dronkers, 1992). In this theory, the key foundation is home environments putting importance on books. The famous *cultural reproduction* theory (Bourdieu, 1986) argues that cultural capital, which tends to be stored more in high-class environments, gives higher benefits to people in such environments. Although this argument is similar to the scholarly culture theory, it negatively expects that people's classes or statuses will be fixated for generations; that is, that the economic and educational gap would be increased between upper and lower classes by cultural capital. However, by analysing the result of PISA data, Andersen and Jæger (2015) revealed that the effect of cultural capital tends to be higher in low-achieving schooling environments than in high-achieving ones. This result is also supported by many studies including the ones introduced above (Benesse

[2] This corresponds to PISA's combined reading scale. It consists of information retrieval, text interpretation, and reflection-and-evaluation.

Educational Research and Development Institute, 2018; M.D.R. Evans, Kelley and Sikora, 2014; M.D. Evans et al., 2010). These results rather support the *cultural mobility* theory: cultural capital will produce more for people in disadvantaged environments (DiMaggio, 1982).

1.3 Physical environments for reading

As investigated in the research above, exposure to books has traditionally been supported by physical environments, such as bookshelves in homes. We review the roles of these environments first, then look at how and how much they have been decreasing in recent decades.

1.3.1 Roles of physical environments in book exposure

Other than household stores of books, we can also find many public bookshelves in bookshops, libraries, and even waiting rooms of hospitals, where they serve local communities. While we have already seen the evidence for household books, local bookshops can be regarded as 'public bookshelves' for the community. In fact, several Japanese surveys show that bookshops are the main place to find books to read (Japan Publishing Industry Foundation for Culture, 2009; Mainichi Shimbun, 2013; National Institution For Youth Education, 2013). The value of the physical existence of public libraries has been discussed with a theme called *the library as place* (Council on Library and Information Resources, 2005; Waxman et al., 2007): physical libraries share communities filled with books opened up to local citizens regardless of their economic status. Considering their effect on education, these facilities that give passive exposure to books can be considered *social common capital* (Uzawa, 2005; 2010), which 'provides members of a society with those services and institutional arrangements that are crucial in maintaining human and cultural life' (Uzawa, 2009, p. 7).

Neuman and Knapczyk (2018) studied the effect of a book distribution programme that installs free vending machines of children's books in a community. They found that the physically closer to these machines were, the more they encouraged reading behaviour in children. Japan MEXT (2019)'s survey also agrees that the reachability to libraries and bookshops is positively associated with the proportion of children who read books. One possible theoretical explanation is that physical exposure should cause mere-exposure effects (Zajonc, 1968), i.e. repetitive exposure to a certain stimulus leads to one's positive acceptance to that stimulus.

Social interaction in the physical world must be also considered as a form of physical exposure to books. The book vending-machine experiment we referred to above (Neuman and Knapczyk, 2018) also showed that greater support by adults around children yielded greater enhancement of reading behaviour in children. That is, the people around you, who read books, are an important source of book encounters. Booknet Canada, a Canadian nonprofit organisation (NPO) for the book industry, reported that 46% of people chose their books to read from word-of-mouth (WoM), i.e. casual mentions from consumers to consumers.[3]

Several other surveys agree on the effect of physical exposure to books mentioned above. Scholastic (2019), a biennial survey of 1,718 pairs of parents and children across the US, revealed that:

- frequent readers (those who read books for fun 5–7 days per week) tend to be surrounded more by those who enjoy reading than infrequent readers (those who read books for fun less than 1 day per week)
- frequent readers possess twice as many books for their children as infrequent readers

According to Japan MEXT (2016), a survey result of Japanese students living in municipalities that promote reading, the top five frequent triggers of reading were as follows:

1. reading promotion at schools, such as morning reading time
2. friends telling about or lending their recommended books
3. books displayed at accessible places in school
4. books displayed at accessible places at home
5. family members often reading books to children

These triggers of reading seem extrinsic and environmental. Besides, a higher number of household books is associated with a larger proportion of readers and number of books read.

1.3.2 Decrease of physical passive exposure to books

We showed international evidence of the decline of physical environments for book exposure. Although this evidence comes from major developed countries, we suppose that the trend is shared across most developed countries, since we can roughly estimate that the countries we will mention represent different types of economic and information-technological stages.

[3] 27 April 2018 (https://www.booknetcanada.ca/blog/2018/4/27/canadians-and-their-reading-habits — accessed on 26 September 2019)

Decline of bookshops

US Although small independent bookshops have been gradually increasing since 2009 in reaction to a massive decrease between the mid-90s and 2009,[4][5] the total number of 'brick-and-mortar' bookshops in 2016 has fallen into less than half that of 1992.[6]

China Similar to the US, the number of bookshops in China was decreasing from 2000 until 2017—in 2012, it was almost half that of 2000—but the number has increased 2.33% year-on-year from 2017.[7]

UK As British news articles repeatedly report,[8][9] the number of bookshops in the UK is also decreasing. In 2017, it had halved in comparison to 1995 (from 1,894 to 897 bookshops) while 2018 and 2019 observed only a slight increase (+16 stores).[10]

Singapore Luyt and Heok (2015) mention the decrease of several bookshop chains in Singapore, such as Borders, Page One, and Sunny Bookshops.

Japan In 1999, there were 22,296 bookshops in Japan, and that number had fallen to 12,026 by 2018 according to a survey by Almedia, Co., Ltd. (Tokyo, Japan). Note that these numbers include offices without stores and stands without salespeople. It is estimated that there are even fewer brick-and-mortar bookshops, or around 8,800, if we only count the stores adopting book-voucher adjustment machines, according to Nippon Tosho Fukyu Co., LTD. (Tokyo, Japan), which licences those machines in Japan. As a result of this decline, more than 20% of municipalities (= 420/1741) in Japan did not have any bookshops

[4] The New York Times, 26 February 2015 (https://www.nytimes.com/2015/02/26/arts/international/assessing-the-health-of-independent-bookshops.html — accessed on 26 September 2019)
[5] CBS News, 23 November 2018 (https://www.cbsnews.com/news/small-bookstores-are-booming-after-nearly-being-wiped-out-small-business-saturday/ — accessed on 26 September 2019)
[6] American Academy of Arts and Sciences analysed from data published by US Census Bureau (https://www.humanitiesindicators.org/content/indicatordoc.aspx?i=11095 — accessed on 26 September 2019)
[7] CGTN, 19 February 2019 (https://news.cgtn.com/news/3d3d774d33677a4e32457a6333566d54/index.html — accessed on 26 September 2019)
[8] The Telegraph, 2 September 2011 (https://www.telegraph.co.uk/culture/books/booknews/8738701/Internet-and-supermarkets-kill-off-2000-bookshops.html — accessed on 26 September 2019)
[9] Independent, 21 February 2014 (http://ind.pn/2U5nlk9 — accessed on 26 September 2019)
[10] The Guardian, 7 January 2019 (https://www.theguardian.com/books/2019/jan/07/independent-bookshops-grow-for-second-year-after-20-year-decline — accessed on 26 September 2019)

in 2017,[11] up from 325 in 2015.[12] This situation is causing a huge geographical gap in book distribution between rural and urban areas in Japan; recently, even urban areas suffer from zero-bookshop regions.[13]

The rise of e-books

E-books are convenient for users to read, but their current implementation does not support passive exposure very well. Consider the situation of home libraries; purchasing e-books does not add to physical books in household bookshelves, and purchased e-books are difficult to share with others, unlike physical books due to copyright issues. After all, children never know what their parents buy and read unless they are explicitly informed about it. In addition, since e-books enable us to find/obtain/read books without visiting local bookshops and libraries, they may prevent us from encountering unexpected books arranged in/across physical bookshelves.

The popularity of e-books is increasing worldwide. Morder Intelligence (2019) reported that the e-book market share grew 25.8% in 2018 from 12.3% in 2013. Detailed statistics for some countries are also available.

US The percentage of US adults who have read e-books in the previous 12 months is increasing, up from 17% in 2011 to 28% in 2016.[14]

China The Chinese e-books market is quickly growing year-by-year: from 2.7 billion yuan to 12 billion yuan during the period of 2011 to 2016.[15] A 2019 survey revealed that the number of the readers of e-books reached 430 million in 2018 as a result of a 14.49% increase from the previous year (China Audio-video and Digital Publishing Association, 2019).

Singapore According to 2018 survey data, the percentage of those who read e-books increased from 2016 data, i.e. from 41% to 55% (National Library Board Singapore, 2019).

[11] According to a survey by Almedia, there were 332 zero-bookshop municipalities (< 420 of 2017) in 2015.

[12] 「書店ゼロの街、2 割超 420 市町村・行政区」(= ‘more than 20% of zero-bookshop cities: 420 municipalities’), Asahi Shimbun (Tokyo, Japan), 24 August 2017.

[13] Urban Commerce Research, 17 July 2019 (https://hbol.jp/197179 — accessed on 26 September 2019)

[14] Pew Research Center, 1 September 2016 (https://www.pewinternet.org/2016/09/01/book-reading-2016/ — accessed on 26 September 2019)

[15] Jiangsu Metropolitan Network, 20 December 2017 (http://news.jsdushi.cn/2017/1220/130626.shtml — accessed on 26 September 2019)

Japan Impress Corporation's research institute (Tokyo, Japan) reported the growing market share of e-books in Japan: fiscal year 2018 was 126.1% of the previous fiscal year.[16] Also, questionnaire results from MyVoice Communications, Inc. (Tokyo, Japan) in July 2018 showed that those who had read e-books in the previous year constituted around 40%, consisting more of younger generations.[17]

Increase of infrequent readers

In this research, we refer to the term *infrequent readers* as those who do not read books regularly. We roughly define the degree of 'regularity' as the existence of a reasonably regular reading habit, e.g. reading one book per month, reading books one day per week, and so on. This flexible definition allows us to integrate different data, and to help us to understand such data.

As we have already seen, being surrounded by frequent readers is also regarded as physical exposure to books. The global trends show, however, that the population of infrequent readers is growing.

US Twenge, G.N. Martin and Spitzberg (2019)'s analysis of a large-scale survey, carried out over more-than 100 million 8th, 10th, and 12th graders across the US, showed that the percentage of adolescents who had not read any books for pleasure in the last year had risen from 10% to more than 30% between 1976 and 2016, and that the percentages of adolescents reading books sometimes or regularly are dropping steeply.

The New Yorker also worried that 'fewer people were reading at all, a proportion falling from 26.3 per cent of the population in 2003 to 19.5 per cent in 2016' based on American Time Use Survey by US Department of Labor.[18]

Netherlands The Netherlands also faces a sharp fall in readers. The proportion of those who read books weekly or more declined to 72% from 90% during the period of 2006 to 2016 (Netherlands Institute for Social Research, 2018). This decreasing trend is stronger in younger ages, or those under 35 years old.

South Korea According to the national reading-behaviour report published by the South Korean Ministry of Culture, Sports and Tourism in 2017, 40% of adults in South Korea do

[16] 23 July 2019 (https://research.impress.co.jp/topics/list/ebook/566 — accessed on 26 September 2019)
[17] https://myel.myvoice.jp/products/detail.php?product_id=24005 — accessed on 26 September 2019
[18] 14 June 2018 (https://www.newyorker.com/culture/cultural-comment/why-we-dont-read-revisited — accessed on 26 September 2019)

not read any books in a year.[19] This proportion increased 5.4 points from the previous survey in 2015, and 75% of those who have read one or more books in a year read less than one book per week.

Japan According to the Statistics Bureau of Japan (Statistics Bureau of Japan, 2016), the average number of people reading books for pleasure has declined from 39.5% to 38.7% and the averag days per year spent reading has decreased from 94.6 to 79.7 days during the period from 2011 to 2016. Furthermore, Agency for Cultural Affairs of Japan reported that the amount of reading books is decreasing in around five years (Japanese Agency for Cultural Affairs, 2019). Infrequent readers among younger generations are also growing in Japan. According to an inter-college survey,[20] half of Japanese college students spend zero minutes reading on the average day, as a result of this increasing trend. In a 2014 survey (Japan MEXT, 2015), 51.4% of high-school students in Japan did not read any books in a month. During the last 15 years, however, the proportion of such non-readers has stayed steady (Mainichi Shimbun, 2019).

1.4 Digital environments for reading

1.4.1 Status of online digital environments

We are surrounded by digital environments that provide information services through digital devices. In the current stage of digital environments, information is usually displayed on screens of digital devices such as personal computers (PC), smartphones, or tablets. Most of them are able to connect to the internet, i.e. digital devices are connected to and communicate with each other through the internet all over the world. For convenience, we use the terms ‘digital environments’ and ‘online environments’ as interchangeably.

Online digital environments have become the primary place for information seeking and consumption. International Telecommunication Union estimated that 51.2% of the global population (= 3.9 billion people) was using the internet at the end of 2018.[21] The estimate says that internet users in developed countries constitute 80.9% of the population, and users in developing countries 45.3%.

[19] 5 February 2018 (http://www.mcst.go.kr/kor/s_notice/press/pressView.jsp?pSeq=16550 — accessed on 26 September 2019)

[20] National Federation of University Co-operative Associations of Japan, 26 February 2018 (https://www.univcoop.or.jp/press/life/report53.html — accessed on 26 September 2019)

[21] 7 December 2018 (https://www.itu.int/en/mediacentre/Pages/2018-PR40.aspx — accessed on 26 September 2019)

According to Twenge, G.N. Martin and Spitzberg (2019), where 40 years of survey data on 1 million US adolescents was analysed, the 12th graders of 2016 spent 6 hours a day on online activities, such as browsing, texting, and using social media, which is more than twice as much time as in 2006.

Kitamura, Hashimoto et al. (2018) made an international survey of information behaviour across five countries in 2016. Across the five countries, the internet was perceived as the best medium for obtaining the following information types: latest news, credible information,[22] and useful information for work and research. A large-scale survey of information behaviour in Japan also showed that the internet was used as the primary source of information seeking for hobbies, and as the secondary source for news consumption (Hashimoto, 2016, pp. 58–62).

1.4.2 Issues with digital environments

A growing number of people spend long amounts of time in online environments. This, in turn, substitutes for the time that people spend paying attention to physical environments. In terms of exposure to books, this situation brings two drawbacks: on-demand UX and personalisation.

On-demand user experience

We can easily find book-oriented digital environments. The major places are online bookshops, digital libraries, and book-review sites. They can provide better experiences than physical bookshops and libraries in some aspects, such as discovering books, learning about book reviews, and managing purchase/loan history.

One problem here is that almost all of these functionalities provided by online environments for reading work only on-demand, i.e. if the users have the proactive intentions to seek for books. Unlike physical bookshops, for example, online bookshops will never show up in front of your eyes unless you so require, such as by typing the store name into the query box of general search engines. In other words, they provide less support for passive exposure to books. As Neuman and Knapczyk (2018) found, a physical availability of books motivates reading behaviours (cf. Section 1.3.1). The appearance of the physical local library close to the local government office, for instance, reminds you of books even when your primary purpose is to obtain a certain legal document. The current accessibility to book-oriented digital environments like this prevents infrequent readers

[22] Except for Japan; the results of the other two types agreed among all of the five countries.

from encountering books, since they rarely have proactive intentions towards books, nor visit/use such services.

Consumers use online bookshops and brick-and-mortar bookshops in different ways. According to a survey in 2015 by the Japan Direct Marketing Association,[23][24] users of online bookshops tend to buy a set of books in bulk, as well as sequels of books they have already read, whereas physical-bookshop users more often buy books by authors whom they have never read. This survey suggested that physical stores were used to encounter new books and online stores were utilised to buy specific books without effort.

Personalisation

Another big issue with digital environments is *filter bubbles* caused by personalisation. E-commerce sites often provide feeds of 'recommended' items based on the user's activity, such as purchasing and browsing within the site (Knijnenburg et al., 2012). Major search engines also show personalised search results by incorporating various users information.[25] Pariser (2011) coined the term 'filter bubble' to describe the situation caused by the personalisation ubiquitously embedded in online environments. Personalisation basically tries to show 'what the user wants', which leads to filtering out 'what the user did not explicitly or actively require' from her/his search results or feeds. This issue is gaining more and more attention, especially in terms of political context (e.g. democracy), since it may result in extremism (Kessler and Jena, 2002) or partisan selective exposure (Prior, 2013). We can find a line of research to measure filter-bubble effects on search engines (Courtois, Slechten and Coenen, 2018) and recommendation systems (Geschke, Lorenz and Holtz, 2019; Haim, Graefe and Brosius, 2018; Möller et al., 2018; Nguyen et al., 2014; O'Hara and Stevens, 2015), although the results are not yet in agreement on the existence of filter-bubble effects in some cases.

Online filter bubbles could be more serious than those in physical environments. As a similar concept, *echo chambers* have also been found in the physical world (Jamieson and Cappella, 2008): people tend to form homogeneous circles of opinions. The one big difference is that physical environments can expose people to 'what they do not want' more easily. An online environment can be designed as a field where 'unrelated' information is completely filtered out.

[23] https://www.jadma.or.jp/tsuhan-kenkyujo/

[24] Impress Internet Watch, 26 January 2015 (https://internet.watch.impress.co.jp/docs/news/685441.html — accessed on 26 September 2019)

[25] E.g. Google started showing personalised search results from 2009. (https://googleblog.blogspot.com/2009/12/personalized-search-for-everyone.html — accessed on 26 September 2019)

1.5 Online social media

While general online environments do not support passive exposure well, their use is expanding all over the world. This is the reason why we need a digital surrogate for unexpected encounters with books that take place in physical environments. An obvious way to implement of such surrogates could be to increase the appearance of book information online, regardless of users' intentions. Although electronic advertisements embedded in web pages have been intended for such unintentional information display, many internet users perceive them as simply annoying (Adobe, 2013). This is likely because digital ads are too conspicuous regardless of the context of the users' information needs, even though the content of the ads is often personalised to be 'relevant'. Somewhere else in the digital environments where users are open-minded would be preferable.

We believe that online social media (OSM) fits this scenario. OSM is a growing online environment that 80% of the internet users in the world actively use (Kemp, 2019). Time spent using OSM is increasing year-by-year (Kemp, 2019; Twenge, G.N. Martin and Spitzberg, 2019). OSM platforms allow users to connect to each other to keep updated on friends' activities. OSM users browse what happens within/around the user's social network without specific information needs (Kitamura, Sasaki and Kawai, 2016). In other words, OSM feeds are consumed passively rather similar to watching TV, unlike traditional online information-seeking behaviours, e.g. using search engines.

To amplify social media posts about books can be a promising digital surrogate to physical passive exposure to books. As introduced in Section 1.3.1, people's choice of books tends to be inspired by social suggestions and recommendations, and this also applies to the online behaviour as well. Booknet Canada reported that 31% of people in Canada chose what books to read next via OSM; tied with browsing online bookshops, OSM was one of the third most frequent reasons, next to WoM (46%) and browsing in physical bookshops (36%).

We explain a positive reason to support this idea, but also state that OSM platforms with their default setup are difficult to use for this purpose.

1.5.1 Pros: Support for unexpected encounters

OSM platforms support sharing functions, e.g. *Retweet* (Twitter[26]) and *Share* (Facebook[27]). The sharing behaviour is common and popular among OSM users; information sharing constitutes one of the significant motivations of users (e.g. in Twitter: I.L. Liu, Cheung and

[26] https://twitter.com

[27] https://facebook.com

M.K. Lee, 2016). In this way, users encounter unintended information through their social networks, including from sources outside their direct friends lists (i.e. friends of friends). On Facebook, viral pieces of information can spread to networks more than 100 deep (Dow, Adamic and Friggeri, 2013). Weng and Takaku (2019) similarly reports that viral hashtags on Twitter can disseminate fast and far across networks. The fact that Twitter users often follow hundreds of different topics (Bhattacharya et al., 2014) suggests that diverse topics are highly mixed in one's social media feed. OSM allows people to keep in touch with friends of varying levels of ties, from colleagues at their current workplace to old friends from school days, and users with whom they have weaker ties are an important carrier of unexpected information (Bakshy et al., 2012).[28] Fletcher and Nielsen (2018) conducted an empirical study of intentional news consumption on OSM comparing several platforms (Facebook, YouTube, and Twitter) across four countries (Italy, Australia, UK, and US). It revealed that unexpected exposure had a stronger effect of increasing the usage of online news sources, especially on younger people and on those with a low interest in news. This could accelerate this topical diversity in the stream of social mentions, since around 30% of topics distributed in social media come from external, out-of-network sources (Myers, Zhu and Leskovec, 2012).

1.5.2 Cons: Strong filter bubbles

We acknowledge that filter bubbles appear as a more obvious and serious issue in OSM than in search engines and recommendation systems. Many studies reveal filter-bubble effects in a range of OSM platforms (Dagoula, 2019; Halberstam and Knight, 2016; Himelboim, McCreery and M. Smith, 2013). This fact stems from *social homophily*, i.e. the general tendency for people to connect similar individuals (McPherson, Smith-Lovin and Cook, 2001). OSM allows users to form a large homogeneous network more easily and quickly than in the physical world. This social filter and algorithmic personalisation work synergistically to form a filter bubble (Geschke, Lorenz and Holtz, 2019).

Social media is basically driven by homophily, and therefore provides a homogeneous information flow with users. As we saw in Section 1.3.2, in physical environments, infrequent readers are likely to be surrounded more by other infrequent readers. This phenomenon would also apply to online social networks, resulting in fewer encounters with mentions of books amongst infrequent readers' OSM environments.

Nevertheless, Flaxman, Goel and Rao (2016) showed that online social media also provide substantial chances for users to face 'unwanted' information (e.g. opposing polit-

[28] Bakshy et al. (2012) defined tie strength as the frequency of interaction (e.g. messaging and commenting) between mutually followed users.

ical opinions) even though filter bubbles were actually formed by users' homogeneous social networking and algorithmic personalisation.

1.6 Necessary attributes of the digital surrogate

So far, we have observed the current situation of the physical and digital environments in terms of passive exposure to books. Physical environments are losing their traditional benefit, whereas the current digital environments offer less support. Considering the growing importance of digital environments in our daily lives, we need a digital surrogate for physical passive exposure to books. One promising digital environment is OSM, where users are open to unintentional information encounters on the one hand, but they tend to form filter bubbles on the other.

One solution is to build a digital environment where users are exposed more to social mentions of books regardless of personalisation of OSM. Such a system should be able to deliver social mentions in a way that physical environments have traditionally inspired people to read. The key perspective is *infrequent readers*, those who rarely read books. They potentially suffer most from the decreasing physical exposure to books, but leading them to reading is harder than with frequent readers, who are basically more sensitive to book-related information. For surrogate physical environments for book exposure, the system must have the capability to inspire infrequent readers to read.

Note that we should never try to match users' interests to delivered book information as with book recommendation systems. What we need to build is a digital surrogate to passive exposure to books that is not personalised in nature (though it may have biases and regional gaps). However, showing information that is completely irrelevant to users' explicit information needs may well be perceived as annoying, or simply ignored, much like online advertisements (Adobe, 2013). We should at least curate social mentions so that they can gently and subtly attract users' notice without discomfort, as in physical environments.

We generically name this necessary attribute of the system *inspiringness* and define its components from the user's perspective in the later chapter (i.e. Chapter 3).

1.7 Research overview

We claimed the necessity of building a digital surrogate system for physical passive exposure to books and mentioned the key concepts, infrequent readers and inspiringness, which we should take into account for the system. To emulate the functionality of physical environments for book exposure, we make use of social mentions of books found in OSM. Amongst the types of physical exposure to books, social mentions can be smoothly translated into the digital world, i.e. online social mentions.[29] Online users are surrounded by a stream of social mentions on a daily basis, and leaving them open to unexpected information encounters.

The fundamental functionality of the system is to expose users to social mentions of books, which is reasonable according to the mere-exposure effect (see Section 1.2.2). In order to build such a system, what should we achieve? The three principal research questions (RQs) are defined as follows:

- how do we formulate the desirable attributes of a digital surrogate system that exposes users (including infrequent readers) to social mentions of books (i.e. inspiringness)?
- what is the feasible design for a digital surrogate system with inspiringness embedded?
- how can the system modules be implemented?

While we devote Chapter 3 to answering these three RQs, we will quickly give the overview here, since it determines the overall goals of this research.

First, based on observations of infrequent readers and related studies around concepts similar to inspiringness, we formulate the four key components of inspiringness that should be implemented in the book exposure provided by the system.

Daily-ness: to what extent the exposure of the social mention happens within the user's daily digital behaviours

Proximity: the closeness to or influence on the user that the author of the exposed social mention possesses

Pleasantness: how pleasant/attractive the exposed social mention is

Considerateness: to what extent the exposure of the social mention is moderate

Second, we design the system with inspiringness as three consecutive parts:

1. to retrieve mentions about books from the social network *proximate* to the user
2. to score the *pleasantness* and *considerateness* of the social mentions
3. to deliver inspiring social mentions to the user's *daily* digital environment (in a *considerate* manner)

Third, we summarise the requirements to implement these parts. The first retrieval part requires the capability to identify social mentions of books from general post feeds

[29] Although the representative form of book exposure in physical environments is bookshelves, simulating them in digital environments (a straight-forward implementation could be a web-browser extension that randomly puts book cover images on the margin of web pages that users browse) may well annoy users, much like online advertisements.

of OSM, while their *proximity* to users can be statically calculated from network statistics. The second scoring part should be able to understand *pleasantness* and *considerateness* from the text of social mentions. The third delivery part needs to adjust the exposure to social mentions of books to the user's preferences so that it is perceived as *considerate*.

We noticed that, amongst the above requirements, the techniques related to NLP must be developed from substantial efforts of careful research due to the difficulty of the tasks: the retrieval of social mentions of books and scoring *pleasantness* and *considerateness*. Henceforth, we refer to these technical modules as the *core NLP modules*. Solving tasks for these modules is essential to build the overall system, and their performances should reach a sufficiently pragmatic level because it directly affects the quality of the succeeding (downstream) modules such as the third delivery part.

We thus set the goal of this research to implement the core NLP modules with *practical* performance.

1.7.1 Technical scopes

In this research, we focus on specific situations and conditions in building the digital surrogate to book exposure. First, we focus on **Twitter**, amongst popular OSM platforms such as Facebook, YouTube, and Instagram. The social mentions of books we deal with are now, more specifically, *tweets that mention books* (TMBs). This is because the nature of posts on Twitter (tweets) seems to correspond best with that of random chats in physical environments in terms of unintentional information encounters. Tweets can be regarded as a mixed stream of individual assertions weakly interacting with one another. It resembles the situation of wandering around small groups of people chatting in a public park near one's home. Of course, posts and comments on other popular platforms are still worth addressing, but they behave like topic/event/content-oriented threads rather than like random chats. Furthermore, Twitter's nature of short text makes the problem harder than handling other OSM's mentions, such as Facebook posts. In other words, we aim to solve more challenging NLP tasks in which we cannot rely on rich textual information within text. The technical outcome obtained by tackling tweet text processing can be smoothly applied to other OSM with longer text, while the reverse is difficult.

Second, we target **Japanese** and Japan for the language and region. As introduced earlier, the situation of Japan is critical in terms of the decrease of passive exposure to books. In particular, an emerging huge geographical gap (i.e. more than 20% municipalities without bookstores) caught our attention. Additionally, Twitter is popular in Japan; the second most frequent language on Twitter is Japanese (Carter, Weerkamp and Tsagkias, 2013).

Third, for the range of ‘**books**’, we adopt UNESCO’s definition of books:

> A book is a non-periodical printed publication of at least 49 pages, exclusive of the cover pages, published in the country and made available to the public. (UNESCO, 1965, p. 144)

Some, especially those who feel that comics are not suitable resources for education, may argue that comics should be excluded from this research if it pays attention to the educational functionality of reading books. According to research on comics and education (Nakazawa, 2005, e.g.,), the frequency of reading manga or comics is correlated to achievement in language arts (particularly sentence comprehension) and a preference for social sciences. ‘How comics work’ in cognition has been studied (Cohn, 2014). We thus do not exclude comics from the definition of books.

1.7.2 Contributions

This research contributes to library-and-information science (LIS) and educational technology by achieving the following original outcomes:

- based on the current situation of reading environments, we revealed the necessity of a digital surrogate system for physical passive exposure to books to deal with a potential educational gap in the future
- we identified and confirmed the requirements for inspiring infrequent readers online to read
- we designed a feasible system architecture and formulated novel tasks that must be solved in order to build the system
- we solved the tasks for the core NLP modules of the system with practical performance, starting from compiling datasets and carefully analysing, ending with thorough error analyses
- we summarised the future paths for building the UI/UX of the system by showing the tasks and problems remaining to be solved

In Section 1.8, we further provide the particular contributions of each chapter.

1.8 Thesis structure

This section lists the abstracts of the succeeding chapters. We also mention what RQs they aim to answer. Some of them are based on the author’s previously published work. The original work is declared if its text is reused in the chapter.

Part I: Background and research questions

The first part consists of two chapters including the present chapter and reveals our concern about passive exposure to books and its relationship with neighbouring topics.

Chapter 1: Introduction

The present chapter declared the background of the problem, the problem definition, and research questions of this thesis. A certain amount of the text was borrowed from the *Introduction* section of the published paper below:

> [Yada et al., 2019] Yada, S., Kageura, K., and Paris, C., 2019 (online first). Identification of tweets that mention books. *International Journal on Digital Libraries*. DOI: 10.1007/s00799-019-00273-4

Chapter 2: Related work

We summarise related work from the following perspectives.

Promoting reading in digital environments: to review existing services and activities of promoting reading in digital environments related to our problems with this thesis.

Book information systems and tasks: to clarify the differences between the digital surrogate system we proposed and existing book information applications such as book search engines and book recommendation systems. This part is also based on the *related work* section of [Yada et al., 2019].

Concepts related to inspiringness: to survey close concepts to inspiringness, which aims to attract those who have not yet shown active interest in certain entities.

Part II: Conceptual framework

We make our concept firm in this second part. Careful designs of the required components of inspiringness and the digital surrogate system are provided. Furthermore, we empirically evaluate our concept with human data.

Chapter 3: Requirements for digital exposure to books

This chapter answers these questions:

- What features are required for the digital surrogate system in order to inspire infrequent readers to read?
- How can we design the system architecture with inspiringness embedded?

- What tasks and modules will be solved and implemented in this research?

We first identify the four necessary components of inspiringness, i.e. *dailiness*, *proximity*, *pleasantness*, and *considerateness*. Then, we design the architecture of the digital surrogate system with these four components embedded. The system comprises three steps: TMB collection, inspiringness TMB (iTMB) scoring, and iTMB exposure. These steps further consist of several technical modules. We finally declare the overall objectives of this thesis: to confirm the necessity of inspiringness components and to implement the **core NLP modules** which constitute the beginning part of the system and involve complex textual-data handling.

While the base concept in this chapter comes from Yada (2014) and the author's master thesis (in Japanese), the text is completely written from scratch.

Chapter 4: Validation of the inspiringness components

This chapter assesses whether the four components of inspiringness do contribute to inspiring infrequent readers, i.e. answering whether the components of inspiringness we identified works as we expected or not. We explain that some of these components have already been shown as effective theoretically and empirically by related work. As for the rest, the combination of pleasantness and considerateness within text of social mentions of books should be confirmed. Thus, we study it through a psychological experiment because of the lack of existing research on this perspective. The statistical analyses show that our hypothesis is valid: considerate recommendation messages enhance the desire to read books; forceful messages, even if they are positively framed, decrease the effect of persuasion (or recommendation). The study can contribute to the related fields, i.e. eWoM research and psychological reactance theory (S.S. Brehm and J.W. Brehm, 1981).

Part III: TMB corpora

Before making use of social mentions of books, we should analyse real data to observe the current state of passive exposures to books online. Since we focus on tweets on Twitter, our target data is tweets that mention books (TMBs).

Chapter 5: TMB corpus creation

We investigate real-world TMBs to reveal the answer of the question: do TMBs exist in Twitter, and if so, how many TMBs are found there? First, to create corpora of TMBs from Twitter, we proposed two heuristic criteria to search/filter TMBs with relatively higher precision, though a huge number of irrelevant tweets still remain. The annotation guideline feasible to scalable human labour is designed to label fundamental linguistic attributes related to inspiringness, i.e. pleasantness and considerateness, as well as TMB or not. We also label the purpose of mentioning books to understand the behaviour of users. We thus obtain two annotated TMB datasets corresponding to the heuristic criteria, one of which is publicly available (Yada, 2019).

Chapter 6: Descriptive analysis of TMB corpora

This chapter scrutinises both of the TMB corpora in terms of purpose, pleasantness, and considerateness. Through this analysis, we answer what characteristics are found in TMBs, which also provide insights for how to retrieve TMBs and how to score the inspiringness of TMBs automatically. The descriptive statistics show the distributions of the three labels and their correlations, such as diverse purposes, a majority of pleasant opinions towards books, and the existence of active book-recommending activities in the wild. In addition to this quantitative analysis, we conduct a qualitative analysis of considerateness by looking into the phrases that recommend books. Its outcome, a categorisation of recommendation phrases, provides detailed linguistic patterns for casual recommendations in reality.

Part IV: Core NLP modules of the system

In this part, we tackle the novel NLP tasks to retrieve and organise TMBs based on the observations in Chapter 6. The chapters include a clear definition of the new tasks, proposals for the machine learning (ML) models to solve them, and experiments to measure their performance.

Chapter 7: TMB identification

As the core NLP module constituting the first part of the system, TMBs need to be identified from general tweets. We should reject not only irrelevant tweets that do not mention any books, but also reject mechanically generated tweets that appear frequently on Twitter, because our ultimate aim is to amplify the mentions around online users' social networks and our system concept relies on social relationships between real users. We thus define the *TMB identification* task as ternary text classification: TMB, noise, and bot. Our model solves this task with a two-step pipeline, i.e. bot filtering followed by noise rejection. We show that our model outperforms baselines and achieves practical performance. Careful error analyses direct the ways to improve our model further. Note that this chapter is a reorganisation of [Yada et al., 2019].

Chapter 8: iTMB scoring

After retrieving TMBs, the core NLP module of the second part of the system organises them in terms of inspiringness. This intends to find inspiring TMBs (iTMBs) from TMBs by scoring inspiringness. From the linguistic perspective, we focus on the pleasantness and considerateness of the TMB text. We define pleasantness scoring as ternary text classification into positive, negative, and neutral, while considerateness scoring is formed as forceful phrase detection. Based on the findings of these two attributes in Chapter 6, we design different methods to obtain the values (or scores) of pleasantness and considerateness.

Part V: Conclusions

Three chapters in this part conclude the thesis by offering a summary (Chapter 9) and future directions (Chapter 10).

2 Related Work

The purpose of this research is to increase exposure to books in digital environments. To this end, we design a system which delivers social mentions of books to users' digital environments. As a surrogate for physical exposure to books, the system takes into account inspiringness to entice users, including infrequent readers, into reading.

In this chapter, we organise related studies from these perspectives stated in Introduction (Chapter 1), i.e. the purpose, the approach, and the key concept of this research. First, we introduce digital services and campaigns related to our purpose of exposing online users to books. Second, we refer to book information systems that retrieve or recommend books or book-related information, which resemble our system in their approach to deliver social mentions of books (secondary information to books) with certain criteria. Third, we compare neighbouring concepts with the key concept of this research, i.e. inspiringness. This chapter thus shows where our theoretical background and the ultimate goal are placed around associated fields.

The work related to other parts of this research, e.g. the components of inspiringness, NLP techniques to process TMBs, and tweet corpus analyses, are referred to in corresponding chapters afterwards. In particular, we will review the concepts and findings related to the four components of inspiringness in Section 3.1. Known characteristics of (tweet) corpora, with regard to inspiringness components, are summarised in Section 6.2, where we analyse the inspiringness-annotated TMB corpora. Section 7.2 elaborates on text classification and tweet-processing techniques for identifying TMBs from general tweets. In Sections 8.1.4 and 8.2.4, we mention relevant techniques for iTMB scoring.

2.1 Promoting reading in digital environments

This research aims to increase digital exposure to books. As we saw in Section 1.3.2, physical environments that traditionally support encounters with books are decreasing. Some programmes and promotions try to enrich physical exposure to books, such as:

[1] https://www.booktrust.org.uk/what-we-do/programmes-and-campaigns/bookstart/

Bookstart:[1] A programme that gives free books to new parents to encourage reading to children. The programme also helps parents with how to read books together with children. After starting in the UK, it expanded to Japan and other countries. Children with the experience of reading books together with their parents have greater literacy than the children whose parents did not read to them often (Japan MEXT, 2019; Scholastic, 2019).

Morning book time:[2] This Japanese campaign sets a 10 to 15 minutes time slot every (weekday) morning for students to read books. Before the start of classes, students may read any kinds of book they want in their classroom during that time slot. This campaign is widely adopted by Japanese schools, especially elementary schools.

Little Free Library:[3] By installing public bookcases in communities, this programme promotes book sharing activities by neighbours. It encourages people to put books for sharing into the community, and allows others to borrow these books freely. Launched in 2009, this programme has been adopted by 91 countries, and improves access to books in local communities.

BookCrossing:[4] This book-sharing campaign started on 2001. Its users put their own books in arbitrary public spaces with unique identifying numbers, and let them be used by random people. On the campaign's web page, those who take such books can register the location information, which enables users to track how far their books travel.

Book subscription boxes: In these services a few selected physical books are regularly delivered to subscribers' home per month. Monthly costs span from 15 to 50 US dollars depending on the service, such as Book of the Month,[5] OwlCreate,[6] and shelff.[7]

While these programmes help increase book exposure for all, including infrequent readers, few digital counterparts exist due to the difficulty we explained in Section 1.4.2.

Nevertheless, there are a few web-based services and applications that attempt to provide books or book information to infrequent readers.

ACADEMIC THEATER:[8] In order to create incidental encounters with books, this system recommends books stored in the academic library of Kindai University (Osaka, Japan) that match users' recent social media posts in terms of the big five personal-

[2] https://www.mediapal.co.jp/asadoku/
[3] https://littlefreelibrary.org
[4] https://www.bookcrossing.com
[5] https://www.bookofthemonth.com
[6] https://www.owlcrate.com/
[7] https://shelff.jp/
[8] https://act.kindai.ac.jp/

ity traits. The personality traits are scored based on relevant words, and scores for books were calculated by their review texts in advance.

Bungomail (great writers mail):[9] This Japanese service sends users a serialised snippet of a book every morning so that they can read the whole book in a month. Books are selected from the public domain (i.e. Aozora Bunko[10]).

As mentioned earlier in Section 1.3.2, one of the difficulties for infrequent readers is to determine what books to read. These services choose books to read for them, and are available for infrequent readers whose information needs to read a book are concretely formed.

Our interest is directed more to motivating infrequent readers to such an active stage of reading. We believe that digital exposure to books can fill the gap in the era of advanced information society. In terms of this viewpoint, a couple of instances are worth mentioning because they promote social mentions of books in OSM.

Sharing-to-OSM functions of social reading services Book review sites like GoodReads allow users to share their activity on the sites (e.g. publishing a review, following other users, and liking reviews) to the OSM feeds they are using.

ePuB Viewer for Twitter[11] On this website, users can try reading the first several pages of a collection of ebooks. The site provides an ebook-sharing function that embeds the first several pages for other users of OSM to read. Although the service is rarely used,[12] it increases exposure to the chance to come into contact with books within digital environments.

We are going to include these instances in our target tweets, i.e. TMBs.

2.2 Book information systems

The system we proposed aims to amplify digital encounters with books by delivering social mentions of books to users. While this application may look similar to existing book information retrieval (IR) or recommendation systems, their goals are different from ours. Existing systems in general aim to provide or recommend books that are relevant to readers' needs, while we aim to provide non-readers with an environment that exposes them to books. We elaborate on this difference for each task/system below.

[9] https://bungomail.com/

[10] https://www.aozora.gr.jp/

[11] https://epub-tw.com/

[12] Searching Twitter for its official hashtag `#epub_tw` results almost entirely in tweets posted by the official account (`@epub_tw`), which means almost no user-generated tweets.

2.2.1 Book information retrieval systems

Book IR is the task of returning book information relevant to queries given by users. This task is categorised as a book-specific version of IR. Typical studies try to improve search results of books in various types of bibliographic databases (Willis and Efron, 2013; M. Wu, Scholer and Thom, 2009). The Conference and Labs of the Evaluation Forum (CLEF)'s Initiative for the Evaluation of XML retrieval (INEX) held the Social Book Search Track series from 2011 to 2014 (Koolen et al., 2014), where book IR techniques were investigated in relation to book-oriented social network sites, such as LibraryThing.[13] The main shared task of the series is called the *suggestion task*: to retrieve relevant book information that can meet a user's request posted as a thread in the LibraryThing forum. In other words, the task was to implement an automated version of the reference service found in physical libraries. BooksOnline, organised from 2008 to 2014, was another book IR workshop, covering a broader range of book-seeking activities than the INEX's shared task: 'starting from the act of deciding what to read, through the exploration and interpretation of a book's content, to sharing the overall experience' (Kazai et al., 2012, p. 2764).

In general, IR is intended to provide relevant information only after users construct search queries. Recently, there is growing attention on proactive information retrieval,[14] in which IR systems try to satisfy users' information needs before they begin to search. However, both traditional and proactive IR assumes that users already have information needs. All work in the book IR workshops above also sets as the starting point users with the existing desire to find books. In contrast, we seek to expose infrequent readers to books, even though they have not expressed a desire to read.

2.2.2 Book recommendation systems

Recommendation systems (or recommender algorithms) are also widely used in various online information systems in order to provide users with personalised information, even when users do not actively seek it. In fact, there are substantial studies of book-specific recommendation systems for book readers (Alharthi, Inkpen and Szpakowicz, 2018). Their purpose is to automatically construct a list of relevant items for the user's profile, and the final goal is to make users consume these items (Ponnusamy, Degife and Alemu, 2018; Sharma and Mann, 2013). Although such algorithms may help people encounter novel information items online, they also generate filter bubbles that can potentially trap users in their original interests (see Section 1.4.2). This will result in, for our context, a Matthew-

[13] https://www.librarything.com/

[14] ProActive Information Retrieval (ProActIR) workshop (https://sites.google.com/site/proactir/) is an example of the popularity of this field.

effect like situation, which frequent readers can be exposed to more book information, whereas infrequent readers are kept further away from books online.

In order to tackle this issue, serendipitous recommendation has been investigated actively (Bogers and Björneborn, 2013; de Gemmis et al., 2015; Izyan et al., 2018; R. Jiang et al., 2016; Lex, Wagner and Kowald, 2018; Pandey, Kotkov and Semenov, 2018; Reviglio, 2017; 2019). Serendipitous recommendation systems aim to deliver novel, unexpected, but still relevant information to users (Kotkov, S. Wang and Veijalainen, 2016). This definition is slightly different from the original concept of *Serendipity*, which we will contrast with inspiringness later in this chapter (i.e. Section 2.3.2). This line of research often handles product-specific recommendation systems such as songs and films, where serendipity tends to be translated into content diversity of items within the same product group (Kaminskas and Bridge, 2016). Serendipity inside book-specific recommendation systems is thus orthogonal to our goal i.e. helping infrequent readers to encounter book information.

Another important difference between our motivation and that of recommendation systems is the optimisation target. As mentioned earlier, recommendation systems, including serendipitous ones, seek relevance of items to recommend to users. Exposure to books itself is independent from the concept of relevance between books and user profiles or needs. In other words, a digital surrogate for physical exposure to books is not necessarily conditioned on the constraint that books are relevant to users' interests or preferences. What we intend to construct is an online environment within which people are exposed to book information through daily conversations, irrespective of whether they have interest in books or not.

2.3 Concepts related to inspiringness

This section covers concepts and ideas similar to inspiringness, which we defined as the necessary attributes of the digital surrogate for passive exposure to books that takes place in physical environments. As we also pointed out, inspiringness should be capable of attracting users without discomfort, e.g. by gentle or subtle appearance.

We name four ideas from the studies interested in environmental characteristics that passively affect human will. The first one is *affordances*, defined as such environmental characteristics themselves. Second, we introduce *information encountering* and *serendipity*, both of which are process-oriented concepts whose requirements may correspond to inspiringness. Third, we mention *attitude or behaviour change* studies that focus on the results of certain stimuli on human perception. The *word-of-mouth effect* is discussed fourth, because our system shares the same material, i.e. social media posts.

2.3.1 Affordances

Inspiringness is attributed to the system requirements for establishing incidental encounters. This perspective of how the system functionalities invoke users' behaviour is conceptualised as affordances. It originated in Gibson (1979) from the ecological point of view, and the idea was spread widely by Norman (1988). As a brief definition, affordances are the action possibility that an environment or object offers to its users. For digital environments, human computer interaction (HCI) research adopts affordances to design UI (McGrenere and Ho, 2000). For instance, descriptive or theoretical research on affordances explores what affordances are identified in the system. According to Hafezieh and Eshraghian (2017) for instance, OSM, which we utilise as the source for our presentation, embraces the following affordances: collaboration, information sharing, socialisation, navigability, association, ubiquitous communication, and personalisation. Combinations thereof are associated with outcomes such as knowledge sharing and reuse, fostering collaborative learning, new forms of publishing, and crisis management.

Unlike inspiringness, however, affordance theory mainly focuses on on human actions involving the system or object as a consequence of specific characteristics or functionalities built into the environment, system, or object. An inspiringness-embedded environment simply makes allusions to the availability of certain books to users, but it does not immediately activate concrete actions, such as actual reading. Although the effects of affordances on human perception are also studied, the centre of attention is still resulted actions (Pozzi, Pigni and Vitari, 2014). Also note that the scope of the outcomes is different between affordances and inspiringness. While affordance theory does not limit the type of actions caused by the affordance, inspiringness is defined as making information perceptible without annoyance so that recipients can remain exposed to the information.

2.3.2 Information encountering and serendipity

Passive exposure can sometimes be perceived as unexpected encounters from receivers. There is a research field dealing with such experience, i.e. information encountering, in which people happen to face the unexpected discovery of useful or interesting information (Erdelez, 1999). It is one model of information-seeking behaviours where people acquire different information from their purposes of information seeking. Especially for infrequent readers, exposure to books works like information encountering on occasions.

The desirable conditions for information encountering would overlap with the constructs of inspiringness. A study showed that type (e.g. news, gossips, and ads), relevance, quality (e.g. authenticity, accuracy, and timeliness), visibility, and sources of information are identified as influencing factors of information encountering online, while user and environmental factors also interplay (T. Jiang, F. Liu and Chi, 2015).

Yet another similar concept is serendipity, often referred to in notable scientific and engineering achievements (Erdelez et al., 2016). While the finding of penicillin or the development of Post-it Notes are known as iconic examples of serendipity, the term has several different definitions and interpretations. A review work (McCay-Peet and Toms, 2017) tackles that chaotic situation to understand serendipity research towards conceptualising serendipity for digital environments. According to the review, common usage of the term more-or-less describes anomalous observations such that they later turn into valuable outcomes due to receivers' careful attitude and readiness against it.[15] Although inspiringness by definition does not include the fascinating consequence right after the exposure to books, we can learn from such literature how to make infrequent readers aware of unfamiliar information (i.e. books). McCay-Peet and Toms (2017) suggest three requirements to facilitate serendipity by computer-aided environments (shortened extraction from p. 39 of the work):

1. enabling a chance encounter to trigger an event
2. supporting the user in identifying the relationship between the cue and their knowledge
3. supporting the user in reaching a significant surprise outcome

Among these, the first requirement fits our system that provides exposure to books. Existing attempts to achieve chance encounters are listed, such as randomness, dissimilarity, and diversification, some of which are also adopted by creative professionals who try to increase serendipity for their work (Makri et al., 2014). We will take these aspects into account when identifying the components of inspiringness in Section 3.1.

2.3.3 Attitude and behaviour changes

The system we propose ultimately aims to influence users to read (or at least, pick up) books. Users include infrequent readers who don't often read books. In social psychology, the concepts of *attitude change* and *behaviour change* correspond to this attempt: changing the original attitude or behaviour in terms of certain targets (Jhangiani and Tarry, 2014).

[15] This is why we stated that the definition of serendipity among recommendation system research was different from serendipity-centred studies.

Inspiringness can be represented as a set of necessities for an attitude change to let passive users perceive target information (i.e. books).

Persuasion is representative of the attitude/behaviour change methods. It is characterised as a linguistic communication-based social action intentionally aiming to change an attitude or behaviour without coercivity (J. O'Keefe, 2012). Information systems supporting behaviour or attitude change are called behaviour change support systems or persuasive systems (Karppinen and Oinas-Kukkonen, 2013; Oinas-Kukkonen and Harri, 2013). For instance, J. Lee, E. Walker et al. (2017) has developed a behaviour change support system for health issues, such as sleep difficulties, through self-experimentation in which users 'formulate, test, and iterate on hypotheses related to how well behavio[u]r plans can produce desired behavio[u]ral outcomes' (p. 6840). Affordances we introduced earlier also play the role as a theoretical framework for such research (Weiser et al., 2015).

To use these systems, the user's will to change her/his behaviour is ethically required because of its functionality of operating mind (Oinas-Kukkonen and Harri, 2013). This prerequisite prevents the systems from being applied to or reaching those who do not desire to change their own behaviour. While our research targets both those who are willing to establish reading behaviour and those who are not, the system we propose only weakly exposes them to book information by making use of the situation under which users are open to unexpected information encounters. That is, our exposure system does not intervene with personal free will or autonomy even when users do not yet desire to make reading a habit since it does not explicitly persuade users to read books.

2.3.4 Word-of-mouth effect

Whereas Word of Mouth (WoM) is information that passes from person to person via oral communication, WoM that appears in web and electronic communication tools is called electronic Word of Mouth (eWoM). Unlike advertisements, WoM and eWoM are generated voluntarily by consumers.[16] Since TMBs can be regarded as eWoM about books on Twitter, using them to inspire reading like our system can make use of their attributes.

Cheung and Thadani (2012) reviewed and integrated the previous findings on eWoM. According to their review, while traditional WoM had already been well recognised as influential in consumer literature, the internet extended WoM into eWoM by adding four characteristics: (i) making it more scalable and diffusive, (ii) more persistent and accessible, and (iii) more measurable, and (iv) adding the ability to be received from unknown senders.

[16] Although sellers can encourage consumers to make (positive) comments on their own products or services by giving them financial incentives, which look like WoM or eWoM, they are categorised separately as stealth marketing (A.M. Kaikati and J.G. Kaikati, 2004).

Based on the integrated model of eWoM effectiveness, four variables of eWoM messages turned out to be related to recipients' actual purchases or purchase intention: argument quality, valence, sidedness, and volume. Argument quality means the plausibility of the statement. Valence is the polarity of the review towards products or services i.e. positive and negative. Sidedness refers to whether the message describes both pros and cons of the product/service. Volume simply corresponds to the number of reviews. Furthermore, external factors such as the eWoM author's expertise and attribution, the platform on which the eWoM appears, and recipients' prior knowledge have direct or indirect influence on the purchase activity. We will incorporate these views into a more detailed way to apply towards inspiringness.

From the theoretical point of view, one of the popular models to explain why and how (e)WoM works is the elaboration likelihood model (ELM) (Petty and Cacioppo, 1988). It hypothesises the existence of two routes for information processing in humans: central and peripheral routes. The central route incorporates critical thinking on the given information using one's own knowledge and investigation; in the peripheral route, evaluation of the given information is instantly run by secondary information like other people's opinions. Basically, the central route is adopted for expertised matters whereas the peripheral route is used for non-professional matters. While WoM messages could be handled via both routes, they are often utilised in the peripheral route of non-experts due to their nature of secondariness (P. Gupta and Harris, 2010). This formulation also applies to our methodology of exposing TMBs to users, including infrequent readers who are not specialised in reading: supporting passive non-expert users in processing given information via the peripheral route.

Part II

Conceptual framework

3 Requirements for digital exposure to books

This chapter identifies the necessary properties of a digital surrogate system for passive exposures to books. First, in Section 3.1, we define the concrete components of inspiringness for the system in order to make the system feasible. Then, we design the system's architecture by embedding these inspiringness components in Section 3.2. Finally, Section 3.3 declares the scope of this research: what objectives we solve in this thesis.

3.1 Components of inspiringness

In Section 1.6, we emphasised the importance of *inspiringness*: the necessary attribute of a digital surrogate system for book exposures. The exposure should be able to provide target information as physical environments have done so far. Physical environments can draw people's attention into the target information even if they are not yet strongly interested in or actively seeking it. In our case, the target information is social mentions of books and people including *infrequent readers*. The key perspectives of which we should be aware of are 'no content personalisation' and 'long-term exposure'.

Although the system needs to attract users, we should not personalise the information content delivered to users, even if 'relevant' contents are intriguing to users (T. Jiang, F. Liu and Chi, 2015). In general, physical environments do not personalise the experience to users,[1] which is a significant driver for the encounter with unfamiliar information.

The exposure system will be used in a long term as physical environments exist around users. Digital environments can be controlled by users far more easily than physical environments. The exposure system should be durable to users acceptance in order to work properly in a long period. A similar counterpart, online advertisements, is not often accepted pleasantly by users due to its annoyance (Adobe, 2013). Our system should keep a good distance from users to obtain their acceptance.

[1] Although advertisements and marketing in physical environments try to target certain demographics of people, they are basically not optimised to the personal level.

In order to define concrete components of inspiringness with incorporating the above points, we followed a general framework for persuasion/eWoM factors, which are broadly grouped into four categories: message sources (senders), message recipients, message contents, and contexts (Cheung and Thadani, 2012; Crano and Prislin, 2006; Fukada, 2002). Putting aside recipient factors due to short of our control, we extracted the four properties from each of the rest categories:

Dailiness how daily the user uses the digital environment

Proximity the closeness between the user and the message author

Pleasantness the joyfulness of the message author towards the mentioned target (i.e. books) perceivable from the message

Considerateness the moderateness of the exposure for the user

Proximity corresponds to the source, while pleasantness and considerateness constitute content features; dailiness and some aspects of considerateness belong to context factors. We will polish each of the concepts further in the following subsections. Note that inspiringness can be defined for more general domains, while we focus on reading books as the target domain.

3.1.1 dailiness

In order to let users encounter information *passively*, the information needs to be delivered or placed into users' *daily* behaviours. That is, book exposures must happen in certain digital environments that users usually use in a daily basis. If an exposure system requires users to do a specific behaviour different from their day-to-day routines, the system may well not be adopted. Unlike popular web services, the system we propose just exposes subtle information to users and does not aim to bring random impulses to please users. One of the requirements is to make users encounter social mentions of books without enforcing any special actions for them. To this end, the system should be able to embed such mentions to users' daily digital environments, such as web browsers, smartphones, or social media platforms. We thus clarify that daily-ness, or *dailiness* constitutes inspiringness as its major premise: how daily the user uses the digital environment.

3.1.2 Proximity

A major challenge of designing inspiring exposures is the requirement to attract even those who are not strongly directed to reading without content matching. *Proximity* comes from the intuition that, when we reflect on our experience to open up new interests, it was often brought by our friends or acquaintances. In other words, those who we know could have been serving as a good entrance to novel ideas. In attitude change and eWoM, which have similar aspects to inspiringness as we summarised in Section 2.3, the factors of information sources (senders) are one of their main topics. Persuasion research, a branch of attitude change studies, has been investigating the following source factors:

- credibility (expertise or trustworthiness; Chaiken, 1980; Priester and Petty, 2003)
- attractiveness (facial or body appearance; Z. Li and Yin, 2018)
- closeness (intimacy; A. Aron, E.N. Aron and Smollan, 1992)
- similarity (memberships, demographic properties, etc.; Fleming and Petty, 2000; Mackie, Worth and Asuncion, 1990)
- power relationships (control over/by others; Briñol et al., 2007)

Literature of these fields suggested that information brought via strong-tie relationships (i.e. strong-tie information) superior to weak-tie information in persuasion because familiar sources (senders) can increase perceived credibility of the information (Aral and D. Walker, 2014; J.J. Brown and Reingen, 1987).

A traditional WoM (not eWoM) study showed that those who feel difficulty with assessing the information are inclined to adopt strong-tie recommendations (Duhan et al., 1997). According to ELM (Petty and Cacioppo, 1988, see Section 2.3.4), those who do not have sufficient knowledge on a target are likely to rely on peripheral information such as (e)WoM. While the Duhan et al. (1997)'s result is a empirical support of ELM, weak-tie eWoM was utilised more than direct WoM under the situation where message receivers are skilful or forced to contemplate on the target (Steffes and Burgee, 2009). Since infrequent readers we are taking into account have less knowledge on books, we can assume that strong-tie information works for inspiring them.

We take a relaxed definition for proximity so as to include several different properties like the above source factors examined in persuasion research. The social distance among users can be represented not only by intimacy among real friends or acquaintances, but also by one-way admiration to, e.g. pop stars or entrepreneurs. In OSM platforms, we can safely regard the other accounts who the user follows (i.e. 'friends', 'followings', etc.) as her/his mentally 'close' people, which may include trusted experts, real-world friends, following celebrities, etc. Thus, we define *proximity* as the closeness, in terms of this relaxed context, between the user and the message author.

From the message-sender's point of view, it is known that consumers are more likely to generate promotional messages intended for strong ties (Choi, Seo and S. Yoon, 2017; Kitamura, Sasaki and Kawai, 2016). A certain amount of social mentions, which we target, are expected to exist in OSM platforms. One caveat may be that, if a user is an infrequent reader, the accounts whom the user follows are also tend to be infrequent readers, according to social homophily (McPherson, Smith-Lovin and Cook, 2001). This might result in fewer TMBs to retrieve around the user. We will alleviate this point by expanding source accounts via social network connectivity, i.e. taking two-hop friends into account (see Section 3.2.2).

3.1.3 Pleasantness

Another key attribute in persuasive messages or eWoM is *valence*: whether the target of an utterance is described positively or negatively. The integrated model of eWoM effectiveness which Cheung and Thadani (2012) proposed, in fact, contains a path in which the valence (e.g. positively framed vs. negatively framed; also known as *polarity*) of eWoM messages is positively associated with purchase intentions of message receivers, via increasing eWoM credibility followed by enlarging eWoM adoption. A recent study also showed that textual sentiment in reviews is a better predictor of product sales than rating scores (K. Chen, Luo and H. Wang, 2017). Standing, Holzweber and Mattsson (2016) suggested that sentiment-bearing eWoM would diffuse faster.

In contrast, negative mentions have a strong effect on decreasing the message adoption, which is more powerful than the effect of positive massages on increasing the adoption. (Liebrecht, Hustinx and van Mulken, 2019; Peeters and Czapinski, 1990) Our system targets infrequent readers who have only slight attentions to books although they like reading and are inclined to establish reading behaviour. They seem more sensitive to negative information about books than frequent readers in terms of the intention to read.

We thus should prioritise positive TMBs to provide with users. We name this inspiringness component *pleasantness* after the *apparently pleasant* state of the message (TMB) sender towards the target (books): the joyfulness of the message author towards books. As this naming implies, we emphasise that the *real* attitude of a message sender does not matter, but that how the receiver can judge it does matter. Because our source of exposures is TMBs (online social mentions to books), pleasantness is essentially a textual aspect of inspiringness.[2]

3.1.4 Considerateness

Reducing perceived annoyance of exposure is critical for our system. Even though TMBs meet proximity and pleasantness well, infrequent readers would be overwhelmed if the messages are prompted too much. Shown messages should not be outstanding and in-

[2] While online social mentions can include multimedia data (e.g. photos, videos, and sounds), we can see that the linguistic part constitutes the core of the information they give.

vasive in UI/UX. This idea is well mentioned in digital advertising research as *intrusiveness*: ‘the degree to which advertisements in a media vehicle interrupt the flow of an editorial unit’ (H. Li, Edwards and J.H. Lee, 2002, p. 39). Pop-up, banner, or direct messaging ads that interrupt users’ present information-seeking behaviours result in advertising avoidance, especially when they operate without prior consent (Rejón-Guardia and Martínez-López, 2014). Advertising intrusiveness is also perceived in social media (Luna-Nevarez and Torres, 2015); Bond et al. (2010) showed that some reasons why users feel intrusive to advertisements came from their lack of control over such messages. Not only traditional banner-style ads, OSM also implements *native advertisements* (Wojdynski and Golan, 2016): paid ads that match the visual style and functions of users’ posts in the platforms. This type is recognised as less intrusive than traditional ways (J. Lee, S. Kim and Ham, 2016), but also carries the risk of being false or misleading advertising that deceives customers about the information source and its sponsors (Wojdynski, N.J. Evans and Hoy, 2018). Although it would not be the case of social mentions generated by consumers, from the point of view of *algorithmic transparency* (Diakopoulos and Koliska, 2017), the exposure system ought to clarify how and why TMBs are shown especially if we adopt a native advertising-like method to the system feature.

To make exposure ‘considerate’, we should moderate the appearance of social mentions of books, similar to online advertising. The appearance includes graphical texture and exposing (delivering) experience. For the former aspect, we can thinking of the colour, size, position, and founts of the delivered message objects, for example. The latter consists, at least, of the frequency of exposure and when/where/why to receive. We specify this aspect of considerateness as *behavioural considerateness*.

Textual content of messages is another important aspect of considerateness. In TMBs, we sometimes see some messages that highly recommend certain books. Although they meet pleasantness, too strong recommendations may be perceived as forceful

> 全人類はTITLE[3]読め。よくわかんなくてもいいから読め。なんとなくわかったかな程度で終わってもいいからとりあえず読め。(Every human beings, read TITLE. Read it no matter how you can understand it. Read it immediately even if you may end up understanding the book just vaguely) [a real Japanese tweet example].

This kind of disfavour or slight resistance can be explained by *psychological reactance theory*: how individuals take an action when a threat violates their freedom (J.W. Brehm,

[3] Henceforth, the placeholder TITLE is used to mask book titles mentioned in message examples. Similarly, we use AUTHOR as the placeholder for author names.

1966; S.S. Brehm and J.W. Brehm, 1981). It is one of the important theories in persuasion and attitude change theory, and has been applied to the design of persuasive messages in medical situations (Dillard and Shen, 2005; Miller et al., 2007). Reactance is defined as a ‘motivational state directed toward the reestablishment of threatened or eliminated freedom’ (J.W. Brehm, 1966, p. 15). Psychological reactance has been treated as a personal trait (Burgoon et al., 2002), and can be regarded as a combination of anger emotion and desire to argue or resist (Rains, 2013).

While much work has been carried out to measure reactance in specific contexts, two studies examined that how messages that persuade Japanese college students to read books affect their reactance (Imajo, 2012; Kiyota and Horii, 2017). Both studies showed that forceful phrases raised reactance, although they lead to more adoption of the persuasion than moderate ones. These results imply that the strong-phrased book-recommendation almost inevitably raises psychological reactance. While more forceful messages may leave stronger impression on recipients’ minds (Buller et al., 2000), intermittently sending reactance-provoking messages may well put recipients under stress. Since reactance is unpleasant feeling (Rains, 2013), we recognise such a situation as harassment, and consider it should be avoided. In contrast to behavioural considerateness, we refer to the considerateness in text as *textual considerateness.*

We, therefore, define *considerateness* as the moderateness of the exposure for the user, taking these different aspects into consideration (i.e. behavioural and textual).

3.1.5 Summary of the inspiringness components

So far, we elaborated on the four components of inspiringness. In summary, the inspiring exposure can be defined as ‘the intermittently continuous stimuli of cheerful social mentions (*pleasantness*) about the certain entity (i.e. books) from users’ social networks (*proximity*) that appear in the user’s daily digital environments (*dailiness*) in a moderate manner (*considerateness*)’.

Inspiringness is originated from the functions of physical environments (Chapter 1). Static existence of books offers passive exposure for people. We are also paying attention to people’s mentions about books as a source of unexpected encounters with books, which leads to the idea of making use of social media for the digital surrogate system of physical exposure to books. Such systems with the inspiringness components embedded can still be associated with some situations in physical environments. One example is, ‘during a lunchtime chat with your friends, one of them by chance mentioned a book which s/he has read recently with satisfaction, along the line of the talk’. Both interpretations are possible: (i) ‘you’ in this example participate in the conversation and (ii) ‘you’ do not. The latter, a version of slightly less proximity, can be a situation where 'you' are just walking by and overhearing 'your' friends chatting with 'your' other friends at the canteen, which may match encounters with retweets on Twitter.

3.2 Architecture and necessary modules of the system

Now that we identified the necessary components of inspiringness, we can design a feasible architecture of the digital exposure system. We roughly set its principal feature to providing users with social mentions of books in Twitter, or TMBs (see Section 1.7). In this section, we draw a specific picture of realising each inspiringness component onto the system.

3.2.1 Overall architecture

Throughout the procedure of collecting and delivering TMBs, we must incorporate the all four attributes of inspiringness. We designed the following pipeline, to this end.[4] Note that, for each step in this pipeline, we also listed fine technical modules and italic shows where inspiringness components are embedded:[5]

1. TMB collection
    - *proximity*-based tweet collector
    - TMB identifier
2. inspiring TMB (iTMB) scoring
    - *pleasantness* scorer
    - textual *considerateness* scorer
3. iTMB exposure
    - *inspiringness* coordinator
    - presentation interface on media of *dailiness*

First, we collect TMBs from users' friends (i.e. following accounts), whiile considering proximity. We then measure the pleasantness and considerateness for collected TMB text. TMBs that satisfy proximity, pleasantness, and textual considerateness are called inspiring TMBs (iTMBs), which are provided by this step so far. Finally, we bring such iTMBs to users' daily digital environments with considerateness. Figure 3.1 illustrates this procedure. We will elaborate on each module below.

[4] We omit general technical components, such as databases and servers, in this design, because allocations thereof can vary with pragmatic situations of system operators.

[5] This module composition is not the sole realisation, but a reference implementation. It is possible to merge some modules or to split a module further while the goal of building a digital exposure system with the inspiringness components can be achieved.

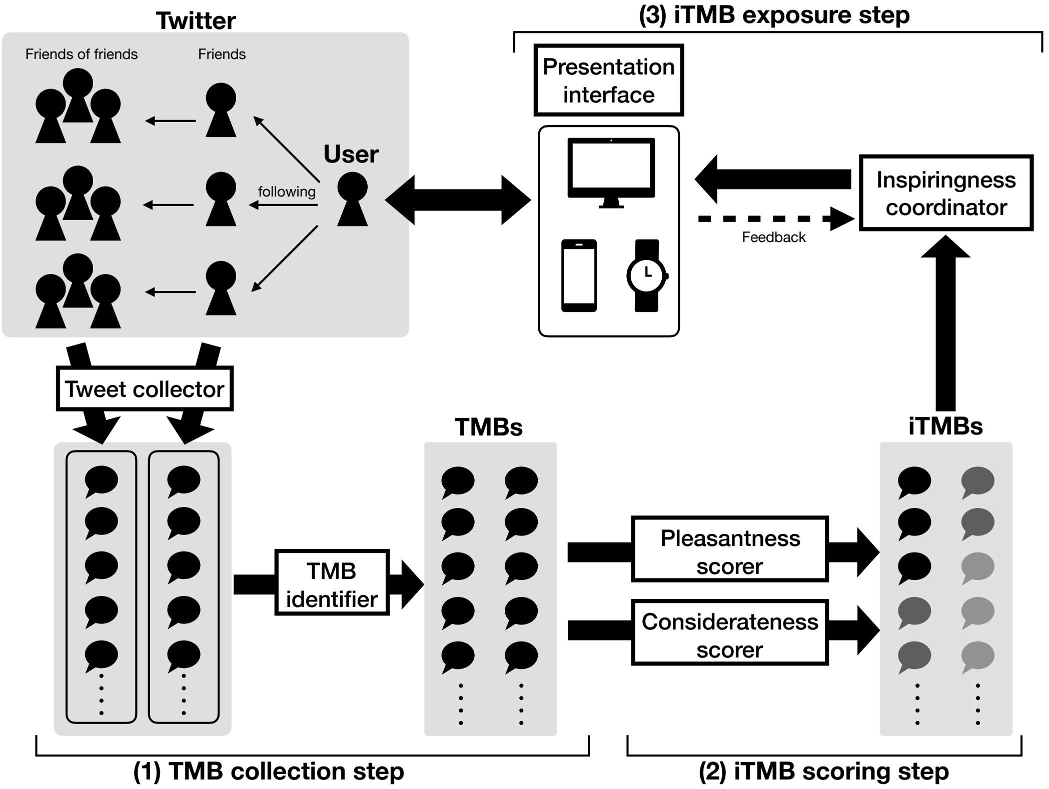


Figure 3.1: Proposed architecture of the digital surrogate system for passive exposures to books

3.2.2 TMB collection step

This step needs to evaluate the proximity of the friends and to identify TMBs from general tweets.

Proximity-based tweet collector

The aim of this module is to meet proximity. We retrieve candidate tweets that contain TMBs from the other accounts who are 'close' to the user. Fundamentally, they are the user's following accounts ('friends' in Twitter's terminology). According to social homophily (Himelboim, McCreery and M. Smith, 2013; McPherson, Smith-Lovin and Cook, 2001), however, topics circulated around infrequent readers might well fewer mentions of books. In the mean time, topics often diffuse across many different user clusters (Dey et al., 2018). We mitigate the issue of scarce existence of TMBs by taking into account 'friends of friends', i.e. up to two-hop accounts in the user's social network. Exposing such TMBs in the end is still valid because the one-hop accounts often re-distribute their friends' tweets (the two hop accounts from the target user) via the retweet function. For the same intention, we include retweets, which might come from more than two hop accounts, into the tweet collection.

This module also scores proximity from users to their source accounts. As mentioned earlier (Section 3.1.2), we can easily rank them with regard to proximity, by assigning interaction metrics between the user and the source accounts, such as the frequency of direct conversations (replying), likes, and retweets. This scores will be used in the iTMB exposure step.

TMB identifier

TMBs have to be identified from the tweet collection because it must contain a large amount of irrelevant tweets to books. This is a non-trivial task due to a variety of the expressions to mention books in the wild (e.g. referring to titles or authors, quoting a passage, and alluding to content). We need to establish a feasible form of this task. Considering the overall system architecture, this filter provides the most crucial source for the digital exposures, especially if we think of the expected scarcity of TMBs in the real world.

3.2.3 iTMB scoring step

After collected TMBs in the first step, the exposure system ranks TMBs according with their inspiringness. While proximity is already met and dailiness is managed in the later step, this second step deals with pleasantness and considerateness. Both are inferred from TMB text parallely, as shown in Figure 3.1.

Pleasantness scorer

This module judges pleasantness from TMBs: how the TMB author is pleased to the mentioned book. This setup is akin to opinion mining (B. Liu and L. Zhang, 2012), which belongs to standard NLP tasks. Since a variety of task formats are proposed for opinion mining, e.g. regarding opinions as discrete classes like positive and negative, or as continuous values ranges from 1.0 to 5.0, we should apply a proper framework to pleasantness scoring. Another issue comes from the fact that popular methods adopt supervised machine learning, which requires a substantial amount of labelled text of the domain (i.e. books) in advance for training the methods.

Textual considerateness scorer

The second module of the iTMB scoring step determines how considerate the text of TMBs is. Inconsiderate language can be defined as phrases to control or stress message receivers and considerateness could be measured by absence or scarcity of such expressions. The most probable representation of inconsiderate text in TMBs could be recommending books as we suggested in Section 3.1.4. The aim of this module is rephrased into detecting the strength of recommendation.

3.2.4 iTMB exposure step

This third step comprises two modules: tuning considerateness variables such as the timing and the amount of iTMB exposures, and defining graphical presentations of iTMBs for digital channels meeting dailiness. These modules for actual exposures to TMBs are another crucial part of the system because users actually face them.

Inspiringness coordinator

By the second step, most inspiringness components are processed for obtained TMBs, i.e. proximity, pleasantness, and textual considerateness. The basic order to serve such iTMBs to users is based on a rank sorted by the value of the above properties; closer, more pleasant, more considerate ones are delivered first. However, effective degrees and balances of inspiringness attributes may change in different personalities. This module offers a capability to adjust inspiringness of TMBs to serve, based on users feedback measured from their usages on the exposure system, such as clicks on iTMBs proposed to users.

This *inspiringness coordinator* also handles non-textual aspects of considerateness, i.e. behavioural considerateness, which affects user experience. That is, this module controls the timing, frequency, and amount of TMB exposure. Users' preferable values will depend on their personality and the type of end digital environments. For instance:

- TMBs may appear more frequently if they are inserted directly into users' Twitter home-timelines than if they are brought as email newsletters;
- for weekday workers, TMB exposure on weekends would be perceived as moderate;
- since psychological reactance varies with individuals, some users may rather accept frequent and information-rich presentations of TMBs.

Presentation interface on media of dailiness

At the final stage of the exposure system, iTMBs are delivered to users' *daily* digital environments. This module provides users with several options for available digital environments, in accordance with their digital lifestyles. Following the present global usage on digital devices and services (Kemp, 2019), the options should include desktop, mobile, and wearable devices, at least. For each device type, we can further select different software application types of dailiness. Emails, messengers, and OSM client applications can be chosen as major digital environments of dailiness (Hashimoto, 2016; Kemp, 2019; Twenge, G.N. Martin and Spitzberg, 2019). For instance, we can send email newsletters containing a list of iTMBs, or occasionally forward an iTMB from a messenger bot. One natural implementation may be a Twitter client application which mingles iTMBs with other tweets.[6]

This module also manages the graphical aspect of *behavioural considerateness* of TMB exposure, e.g. its size, colour, and format. In any destination environment, TMB exposure should follow the original format and style of the platform.

3.3 Overall objectives of the thesis

Now we defined inspiringness and the digital surrogate system for passive exposure to books, we will set the overall objectives of this thesis in a more concrete manner than we made at the beginning (i.e. Section 1.7). We solve the following two objectives:

1. to validate the expected effects of the inspiringness components along the book exposure scenario (Chapter 4)
2. to develop the modules involving NLP techniques (i.e. *core NLP modules*) at the practical level
    - to collect and investigate TMBs (Chapters 5 and 6)
    - to solve the TMB identification task for the TMB identifier module (Chapter 7)
    - to solve the tasks to measure pleasantness and considerateness for the modules of the iTMB scoring step (Chapter 8)

Through achieving these objectives, this thesis can provide the following original contributions:

a. identifying, arranging, and confirming the requirements of the digital surrogate system for physical passive exposures to books and;
b. the task formalisation and development of the NLP back-end modules essential for the system.

In other words, we will show the feasibility of the pipeline-based system-architecture we proposed, given the technical practicability of UI/UX modules. We elaborate on these points below.

[6] The best solution could be implementing our system into official Twitter applications although petitions are beyond our research purpose.

3.3.1 Validation of the inspiringness components

While we have already identified the four components of inspiringness in Section 3.1, we ought to confirm whether they are actually the factors to be considered for inspiring exposure to books. That is, we must show that different states of the components surely influence users' attitude towards reading books. Although we derived each component of inspiringness from existing evidences of relevant fields, some points still remain unclear. For these points, we justify the influences of inspiringness components averaged over different people, including infrequent readers. We will organise what needs to be clarified, later in Chapter 4.

Note that the factors of users' personality traits are beyond the scope of this validation. It is expected that the effectiveness on inspiringness can vary with regard to personality traits, which may raise a finer, advanced research question: how do different states of the inspiringness components affect users' inspiringness according to their different personality traits? Our research instead contributes to offering an integrated experimental environment for future research that tackles such a question, by realising the ways to implement an exposure system.

3.3.2 Development of the core NLP modules

Considering the pipeline architecture of the proposed system, the modules of the first and second steps play a vital role for actualisation of a digital surrogate for passive exposure. We acknowledge that, in the system operation, the quality of UI/UX parts of the system is significant because users directly face them. Whereas the implementation of such UI/UX components are technically feasible, we lack the directly existing methods for preparing TMBs with inspiringness processed, which are the essential source of what the UI applications present to users. Especially, implementation of NLP-related modules, which involves processing of TMB text, faces this difficulty. The reason why solving the demanded tasks is not obvious is because they belong to novel applied tasks—we must translate our tasks into formal NLP applications. Due to the lack of actual TMB data, furthermore, we have even no idea of what difficulties lie on these tasks. The essential issues are, thus:

i. to confirm the existence and characteristics of TMBs via LIS-based procedures, i.e. collecting and sorting data carefully, and then;
ii. to make TMBs accessible as the transition from the current state where we can imagine their existence but cannot easily reach out to them.

This research, therefore, focus on developing the **core NLP modules** which includes the modules that require NLP methodologies, i.e. *TMB identifier*, *pleasantness scorer*, and *considerateness scorer*. We devote most of the rest chapters for this objective. As preparation for the module development, we garner TMBs at Chapter 5 and scrutinise inspiringness of them at Chapter 6. The task for TMB identifier is tackled in Chapter 7. At Chapter 8, we deal with pleasantness and considerateness scoring.

Technically speaking, the outcome also contributes to more general NLP fields by challenging the difficulties of processing tweets in their short informal, high-contextual texts. Besides, the methods to access social mentions of books *in general*[7] with inspiringness scored brings a new perspective to LIS research.

Note that, other than exposure-related modules, we skip *proximity-based tweet collector* in this research too, because its implementation is relatively obvious as we described earlier.

[7] Online mentions of specific kinds of books can be relatively easy to collect by keyword-based information retrieval

4 Validation of the inspiringness components

In the previous chapter, we defined the four components of inspiringness and the system architecture for the exposure system. Before moving towards the system development, we should affirm that those inspiringness components surely constitute inspiringness itself, especially for inducing people, including infrequent readers, to read. To this end, we show that each component affects users' attitude to books positively. While the validity of some components can be proved from related work, that of the others needs to be empirically examined.

First, the necessity of dailiness is obvious in inspiring exposures on digital environments. Dailiness is the requirement that book exposure ought to be delivered to users' daily information behaviour so that they can notice the exposure without any special actions. If we force users to use an unfamiliar digital environment, information delivered therein may well not be consumed because they would forget using the environment. This is why we defined dailiness as the major premise of inspiringness (Section 3.1.1)

Second, the effectiveness of proximity is considerably supported by existing work in attitude change, persuasion, and eWoM (see Section 3.1.2). People known to a user in some ways can attract the user more than complete strangers if other source factors (e.g. expertise, attractiveness, and/or power) are controlled. This applies well to infrequent readers because they should rely more on the peripheral route of ELM (Petty and Cacioppo, 1988), since they are less expertised in reading.

Third, pleasantness is also supported well by eWoM research. As mentioned in Section 3.1.3, it is known that pleasant-looking messages encourage recipients whereas complaints carry negative impressions.

Finally, considerateness has two aspects: behavioural and textual. As we summarised in Section 3.1.4, behavioural considerateness corresponds more-or-less to non-intrusiveness in advertising research, which basically suggests that outstanding presentation should be avoided (Rejón-Guardia and Martínez-López, 2014); textual considerateness lay its foundation on psychological reactance theory, which explains the displeased feeling occurred by forceful messages (S.S. Brehm and J.W. Brehm, 1981).

This psychological reactance theory, however, gives somewhat conflicted findings on explicit recommendation scenarios: strong recommendations (e.g. 'you must consider solar protection'; Buller et al., 2000) or forceful orders (e.g. 'you had better read books'; Imajo, 2012) result in more message acceptance, though they evoke unpleasant feelings on recipients' minds. Only considering a resulted effect on reading desire, textual considerateness might not be significant for inspiring exposure. Moreover, because forceful messages may activate an opposite attitude or action to what the message says (i.e. psychological reactance), unpleasant negative recommendations (e.g. "you shouldn't read this bad book") could rather encourage reading, which conflicts with pleasantness.

Therefore, we need to scrutinise the effect of pleasantness and (textual) considerateness in their combination, towards inspiringness for reading books. We devote this chapter for an empirical investigation on the pleasantness-considerateness interaction. That is, this empirical investigation is committed to showing the effectiveness of pleasantness and considerateness, while that of the other inspiringness components are regarded as evaluated by related work.

4.1 Pleasantness-considerateness interaction

We aim to measure the effect of pleasantness and considerateness in text, especially focusing on our application, i.e. a digital surrogate system for passive exposure to books. We adopt a questionnaire-based method asking how much different TMB-like messages inspire participants to read. The following definitions are recapitulation of a part of inspiringness theory optimised for passive exposure to books using TMBs, within this context:

Inspiringness: the extent to which the receiver of a TMB is inspired to read the mentioned book

Pleasantness: the positiveness of a TMB towards the book

Considerateness: the moderateness of the book recommendation in a TMB

For this research, we define the opposite of considerateness as **forcefulness** and henceforth use mainly this term instead of considerateness because of its compatibility with strength adjectives (cf. 'weak forcefulness' vs. 'strong considerateness'). The present work, thus, aims to understand how recipients are inspired to read a book by, e.g. positively/negatively framed weak/strong recommendations of the book.

Our assumption of the pleasantness-forcefulness interaction are as follows:

- negative recommendations of books (unpleasant TMBs) will decrease inspiringness;

- positive recommendations of books (pleasant TMBs), instead, will increase inspiringness;
- on the other hand, too forceful a positive recommendation may decrease it.

This hypothesis can be indirectly supported from eWoM research and psychological reactance theory: positive eWoM messages are, in general, well associated with the positive attitude of receivers towards the entity (Cheung and Thadani, 2012); according to reactance studies, forceful messages tend to have the opposite effect of what the messages say (S.S. Brehm and J.W. Brehm, 1981). However, the second step in this logic also implies that negative and forceful recommendations of books might increase inspiringness, which conflicts to our inspiringness theory. In the questionnaire, therefore, we prepare TMB-like messages with different levels of *pleasantness* and *forcefulness* so as to measure their effects towards *inspiringness.*

The present work also has some novel viewpoints with regard to related fields. First, the effect of message senders' pleasantness (or the polarity of messages towards the entity) has not yet been clarified in relation to message forcefulness or psychological reactance (Cheung and Thadani, 2012). Second, our study can provide a finer insight about the effect of linguistic characteristics of messages towards psychological reactance, the research of which has not been focusing much on it yet (Miron and J.W. Brehm, 2006; Steindl et al., 2015).

4.2 Method

We designed a questionnaire to ask inspiringness of TMB-like messages in which pleasantness and forcefulness are controlled. We consider the following levels for pleasantness and forcefulness:

Pleasantness: negative, neutral, positive (three levels)

Forcefulness: none, weak, strong, excess (four levels)

The notation of '[*pleasantness-forcefulness*]' (e.g. [positive-weak]) denotes a combination of them. Among the possible $3 \times 4 = 12$ combinations, we exclude weak, strong, and excess in the forcefulness for the neutral pleasantness because of their unnaturalness—a book evaluated as neither good nor bad is hardly ever recommended in any degree. In other words, we consider only [neutral-none] for the *neutral* pleasantness and for *positive* and *negative* pleasantness, all four levels of forcefulness are incorporated. In above-*none* forcefulness (*weak*, *strong*, and *excess*), *positive* messages recommend reading books that message senders appear to evaluate highly, whilst *negative* messages recommend not reading books that they regard as bad. We prepare three samples of TMB-like messages

Table 4.1: The number of TMB-like messages grouped by pleasantness and forcefulness in the questionnaire

Pleasantness	**Forcefulness**			
	None	**Weak**	**Strong**	**Excess**
Positive	3	3	3	3
Neutral	3	–	–	–
Negative	3	3	3	3

for each combination in order to control the effect of other linguistic features than pleasantness and forcefulness in the analysis. TMB-like messages are therefore grouped into 9 combinations of pleasantness and forcefulness, as shown in Table 4.1. These messages are ordered randomly in the questionnaire.

Most of TMB-like messages are borrowed from real examples of our TMB dataset (see Chapter 5). For [negative-weak/strong/excess], we composed such examples by our introspection because they did not appear in the dataset. In order to exclude the effect of book titles or other bibliographic information toward inspiringness, we substitute placeholders for all references to them. This is an example from [positive-none], which masks a book title: “『TITLE』が面白かった (“TITLE” was fun)”. These messages are placed randomly in the questionnaire, to avoid the effect of sequential placements with regard to pleasantness-forcefulness combinations.

For the assessment of whether we obtained fair responses, we inserted two identical messages in the questionnaire. If the responses of a participant were very different between the first appearance and the second appearance, the other responses by that participant would be unreliable. These identical messages (henceforth, reference messages) are put into [neutral-none] examples, i.e. “『○○○○○○』を読んでる (I am reading ‘○○○○○○’)”, since we are mainly interested in the interaction between pleasantness and forcefulness. We carefully placed them keeping a distance of each other in the questionnaire. Other original messages are listed in Appendix A.

4.2.1 Measures

Inspiringness

The primary target response is inspiringness: the extent to which a reader of the TMB is inspired to read the mentioned book. For each TMB-like message, the questionnaire asks participants to report inspiringness of the message as follows: ‘Would you like to try this book?’. We adopted 4 point scale for the answer:

0: ‘I do not think so’,
1: ‘I think so a little’,
2: ‘I think so well’,
3: ‘I think so very well’.

Perceived ‘forcefulness’

We also measured the response of forcefulness of the TMB-like message, in order to evaluate our assignment of message samples to designed forcefulness levels. This is because individual perceptions of forcefulness in messages may vary according to psychological reactance theory (S.S. Brehm and J.W. Brehm, 1981). For the question ‘Is this message forceful (pushy)?’, we use the same 4 point scale as that of inspiringness. The response to this question is referred to as ‘perceived “forcefulness”’[1] in contrast to *designed forcefulness*, i.e. our design of forcefulness levels for TMB-like messages. Note that we did not evaluate pleasantness in this way because we assume the opinion polarity in messages is obvious enough for message readers.

Reading attitude and behaviour

In addition to the above two responses for each message, we also collected participant-wise information about reading behaviour and stance. Amongst participant characteristics, we expect that reading attitude and behaviour may affect inspiringness. Based on existing surveys (see Section 1.3.2), the questions are designed as follows:

(a) Do you like reading books? (5 point scale from 1 = dislike to 5 = like)
(b) Do you think reading books is important? (5 point scale from 1 = disagree to 5 = agree)
(c) How often do you read books?
    - more than one book per week
    - from one to four books in a month
    - one book in a couple of months
    - one book in a year or so
    - I do not read any books at all
(d) How many hours for a day do you spend in average when you read books?
(e) Do you want to change your amount of reading?
    - I want to increase

[1] This ‘forcefulness’ is not exactly the same as our definition of *forcefulness* in inspiringness theory because the question uses the word ‘forceful (pushy)’ as its common usage.

- the current amount is enough
- I want to decrease

Note that we excluded comics and magazines from 'books' in these questions although we did include comics into the range of 'books' throughout the thesis (see Section 1.7.1). We avoid unnecessary conflict in participants' perceptions in the experiment, since under the common sense in Japan, the phrase '読書 (reading books)' seems not to include the act of reading them, usually (Kunimoto et al., 2009). Also, our focal points are pleasantness and considerateness; we do not aim to reveal detailed effects of reading habits in different genres and formats towards inspiringness.

4.2.2 Participants

A total of 35 undergraduates and graduate students in the University of Tokyo participated in exchange for 1,000 JPY book vouchers. To answer the questionnaire written in Japanese, all participants are confirmed as Japanese native speakers. We recruited 25 of the participants from an introductory class of the faculty of education whereas the rest 10 were recruited from a laboratory of data mining based on snowball/chain-referral sampling.

To avoid collecting unnecessary private information, we did not record sex nor age. We also believe that such demographic traits do not influence inspiringness directly even though they may behave like confounding factors in a reflection of the possible skewed distribution of reading preferences in Japan or the University of Tokyo. We can reasonably assume that such demographics of the sample follow that of the University of Tokyo.

4.2.3 Procedure

The participants solved the questions in the questionnaire provided online,[2] in a classroom or a laboratory. The participants were informed the purpose of this questionnaire research by the first page of the online questionnaire (see Appendix Appendix A) as well as a separately attached document for informed consent. We substituted the consent to join the research for answering all questions in the questionnaire, while we also instructed that they can cease to cooperate this research by quitting the questionnaire answering at any time. The book vouchers were passed in exchange for the completion of the questionnaire.

[2] We used Google Form to create the questionnaire. See all questions in Appendix A

Table 4.2: The number of responses (inspiringness and perceived 'forcefulness')

	0	**1**	**2**	**3**
Inspiringness	418	356	140	31
Perceived 'forcefulness'	316	210	210	209

4.3 Result

4.3.1 Descriptive analysis

Responses

First, we checked the reliability of questions we designed for inspiringness and perceived 'forcefulness' by using Cronbach's α (Cronbach, 1951). We obtained 0.899 for inspiringness and 0.923 for perceived 'forcefulness'. These values suggest that the internal consistency of our questions is *acceptable* (Tavakol and Dennick, 2011).

We also checked the deviation between the reference messages in [neutral-none] (see Section 4.2). The difference between the first and second appearance of the questions was small in average (mean): -0.0857 in inspiringness and 0.0571 in forcefulness. Note that both median and mode across participants were 0. We observed that the largest difference between the two reference messages was 2.0 for both inspiringness responses and perceived 'forcefulness' responses, which was produced by two participants. However, their median absolute deviation aggregated over all the other pleasantness-forcefulness combinations were not outstanding in comparison to those of other participants, which can be interpreted as no adversarial responses made. These values above support that all of the participants answered questions fairly.

Next, we observe the distribution of the responses. For each value, counts of question responses over all participants are shown in Table 4.2. Inspiringness has a peak at value 0 ('I do not think so') and decreases as the value increases. Because we masked book titles, most messages seem not to have strong inspiringness. However, this also suggests that linguistic characteristics of book mentions discretely affect inspiringness in a slight, but certain manner. The total responses of perceived 'forcefulness' are balanced among value 1–3 while forcefulness also has a peak at value 0. These values match our design of designed forcefulness in relation to pleasantness levels: three of *none* and two of *weak*/*strong*/*excess*.

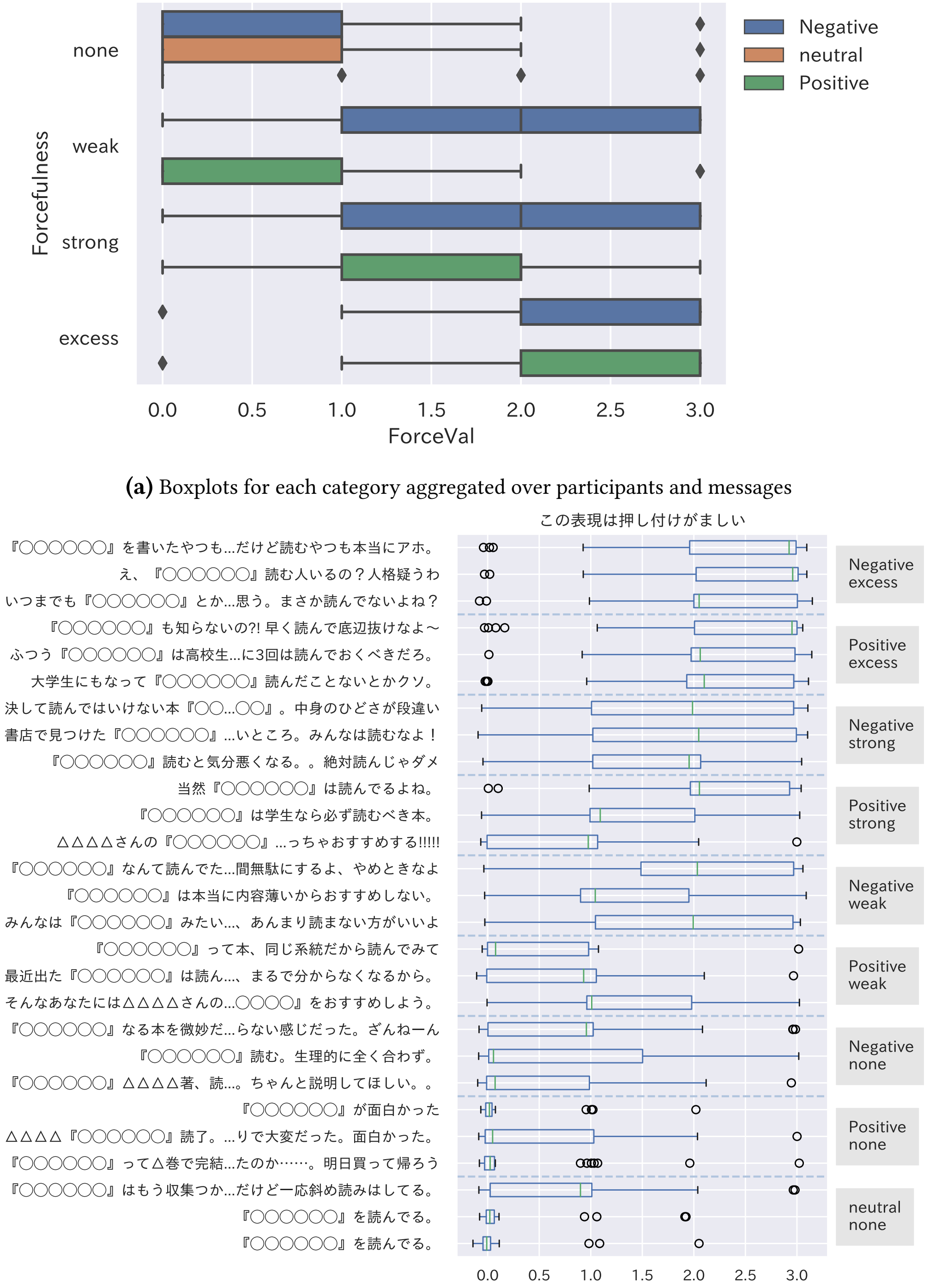
To confirm whether our forcefulness design in TMB-like messages worked as expected or not, we measured the correspondence of the messages grouped by designed forcefulness levels to their perceived 'forcefulness' responses. Figure 4.1 shows boxplots of per-

Table 4.3: Crosstab between forcefulness categories and perceived ‘forcefulness’ responses

Category	Response values			
	0	1	2	3
None	214	64	27	10
Weak	55	72	49	34
Strong	32	56	67	55
Excess	15	18	67	110

ceived ‘forcefulness’ responses for each message grouped by pleasantness and forcefulness levels of TMB-like messages. Figure 4.1a illustrates the boxplots of perceived ‘forcefulness’ responses aggregated over the same pleasantness-forcefulness combinations, whereas Figure 4.3b plots the mean values for each combination. Table 4.3 counts the exact correspondence between designed forcefulness levels and perceived ‘forcefulness’ responses. From these figures and a table, our allocations of messages to each designed forcefulness category match well in the *neutral* and *positive* pleasantness, but responses in the *negative* pleasantness shows a slight discord. Remember we defined forcefulness here as the extent to which the message tries to control the recipients, which is more specific than the general or casual meaning of ‘forcefulness’. This result suggests that a negative assessment (towards books) alone can be interpreted as ‘forceful’. We also measured the degree of ordinal association between designed forcefulness levels and perceived ‘forcefulness’ responses by calculating two rank-order correlation coefficients, i.e. Spearman’s ρ (Spearman, 1904) and Kendall’s τ (Kendall, 1938). Both of the obtained values are substantially distant from 0 (= no ordinal association): 0.601 and 0.521 respectively. For the succeeding analysis, thus, we adopt designed forcefulness as a controlled variable over TMB-like messages. That is, we set the levels of pleasantness and (designed) forcefulness assigned to TMB-like messages as independent variables, whereas we treat inspiringness responses as dependent variables.

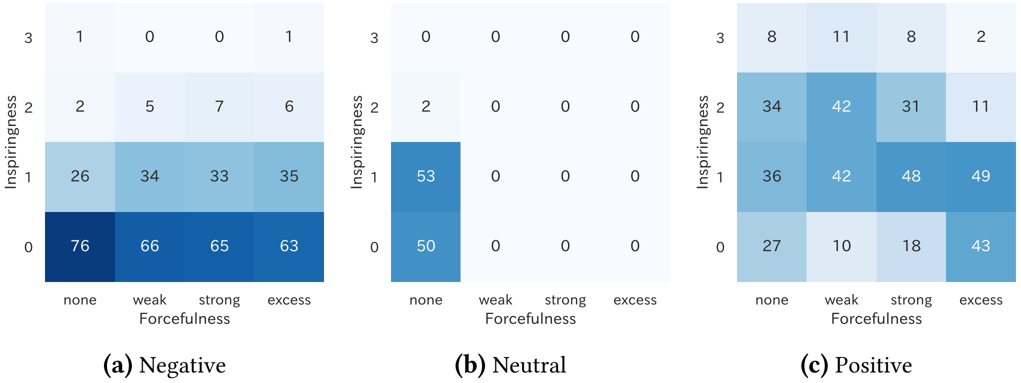
Figure 4.2 shows heatmaps of the crosstabs of overall response counts between inspiringness and designed forcefulness levels split by each pleasantness levels. Figure 4.3 is a line plot of the mean inspiringness responses for each pleasantness-forcefulness category. These figures suggest an interaction between pleasantness and designed forcefulness towards inspiringness. In the *negative* pleasantness, inspiringness is basically low along with all forcefulness levels whereas high forcefulness seems slightly high inspiringness. Inspiringness are relatively higher in the *positive* pleasantness, but the higher forcefulness appears to result in lower inspiringness. These line shapes basically support our assumption



(a) Boxplots for each category aggregated over participants and messages

(b) Boxplots for each question aggregated over participants. For legibility, response values (0–3) are made jittered randomly. TMB-like messages translated into English are available in Appendix A

Figure 4.1: Boxplots of perceived 'forcefulness' responses over all participants grouped by pleasantness and designed forcefulness categories

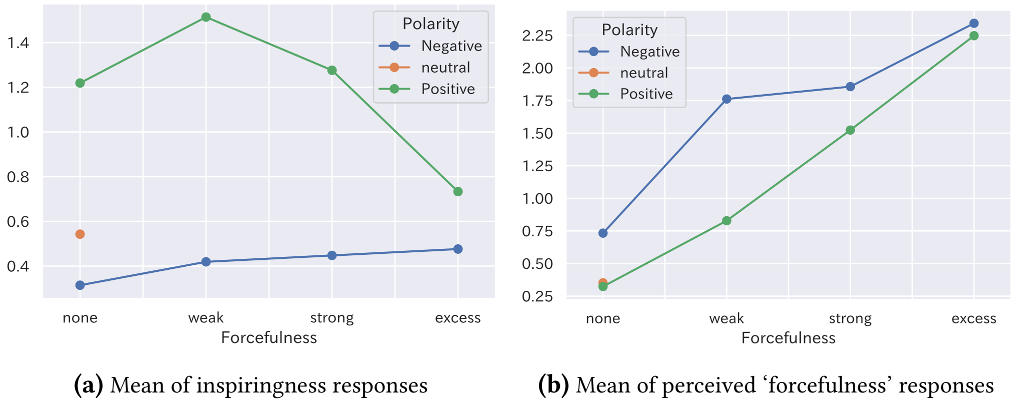


(a) Negative

(b) Neutral

(c) Positive

Figure 4.2: Crosstab heatmaps of inspiringness responses counted over all participants and messages in relation to designed forcefulness levels grouped by pleasantness levels



(a) Mean of inspiringness responses

(b) Mean of perceived 'forcefulness' responses

Figure 4.3: Mean of the responses for inspiringness and perceived 'forcefulness' grouped by pleasantness

for pleasantness-considerateness interaction, but we still ought to take into consideration random effects of participants and TMB-like messages.

Reading attitude and behaviour

Next, we report the reading attitude and behaviour of the participants to assess to what extent we can generalise the result. Total counts for each question are listed in Table 4.4. This table also provides the breakdown of these counts in terms of the difference in faculties (humanity students vs. science students).

From the answers of the question (a), 'Do you like reading books?', more than 70% of the participants like reading books. This seems normal because several surveys showed that more than 60% of samples in Japan basically like reading (Japan MEXT, 2015; Japan

Table 4.4: Responses for the questions asking reading attitude and behaviour. The values in parentheses denote column-wise percentage for each question. The numbers assigned for each answer correspond to the values used for coding.

Question	Humanity	Science	All
(a) Do you like reading books?			
1. (Dislike)	2 (8)	0 (0)	2 (6)
2.	1 (4)	1 (10)	2 (6)
3.	4 (16)	1 (10)	5 (14)
4.	10 (40)	3 (30)	13 (37)
5. (Like)	8 (32)	5 (50)	13 (37)
(b) Do you think reading books is important?			
1. (Disagree)	0 (0)	0 (0)	0 (0)
2.	0 (0)	0 (0)	0 (0)
3.	5 (20)	1 (10)	6 (17)
4.	6 (24)	3 (30)	9 (26)
5. (Agree)	14 (56)	6 (60)	20 (57)
(c) How often do you read books?			
0. I do not read any books at all	0 (0)	0 (0)	0 (0)
1. One book in a year or so	4 (16)	1 (10)	5 (14)
2. One book in a couple of months	6 (24)	4 (40)	10 (28)
3. From one to four books in a month	12 (48)	4 (40)	16 (45)
4. More than one book per week	3 (12)	1 (10)	4 (11)
(d) How many hours for a day do you spend in average when you read books?			
$[0.0, 0.5)$	8 (32)	3 (30)	11 (31)
$[0.5, 1.0)$	4 (16)	2 (20)	6 (17)
$[1.0, 1.5)$	8 (32)	2 (20)	10 (29)
$[1.5, 2.0)$	0 (0)	3 (30)	0 (0)
$[2.0, 2.5)$	4 (16)	0 (0)	7 (20)
$[2.5, 4.5)$ $(= [2.5, 3.0) \cap \cdots \cap [4.0, 4.5))$	0 (0)	0 (0)	0 (0)
$[4.5, 5.0)$	1 (4)	0 (0)	1 (29)
(e) Do you want to change your amount of reading?			
1. want to increase	23 (92)	7 (70)	30 (85)
0. stay in the same	2 (8)	3 (30)	5 (14)
-1. want to decrease	0 (0)	0 (0)	0 (0)

Publishing Industry Foundation for Culture, 2009; National Institution For Youth Education, 2013).

For the question (b), 'Do you think reading books is important?', nobody in the sample thought that reading was unimportant. Although we do not have relevant survey data to this question, the distribution may be deviated from the average perception of overall Japanese population, since education could be correlated to the positive perception.

The responses to the question (c) ('How often do you read books?') shape a peak around the answer 2 and 3. The definition of *non-readers* in (Mainichi Shimbun, 2013) are those who read less than one book per month, which corresponds to the answer from 0 to 2. Our sample consisted of 57% readers and 43% non-readers. This proportion (i.e. around a half of the population) basically matches to the result of other surveys in Japan (Japan MEXT, 2015; Mainichi Shimbun, 2019; see also Section 1.3.2).

For the question (d) ('How many hours for a day do you spend in average when you read books?'), Table 4.4 gives histograms in 30-minutes bins. Our sample can be divided by half at 1.0-hour point. We can see two peaks in the ranges $(0.0, 0.5]$ and $(1.0, 1.5]$. Several statistics agrees on frequent reading time in average which spans around 30 minutes (National Institution For Youth Education, 2013; a survey of National Federation of University Co-operative Associations of Japan[3]).

In the result of question (e), 'Do you want to change your amount of reading?', most of the participants (85%) 'want to increase', while nobody 'want(s) to decrease'. This is deviated from the Japanese average, which is 60.4% in 2018 (Japanese Agency for Cultural Affairs, 2019).

In summary, our sample can be characterised as basically interested in reading, but their actual reading behaviours are close to the Japanese average. While we should be aware of the difference in the attitude variables (Questions a, b, and e), from the behavioural aspects (Questions c and d), the outcome of this research should possess a certain generalisability at least towards a subset of Japanese population of similar demographic properties (e.g. college students). Our sample also includes 42% of infrequent readers when we regard those who read less than one book per month as them, which is a reasonable size for assessing the effects of inspiringness components (i.e. pleasantness and textual considerateness) on them.

Figure 4.4 is a heatmap of Kendall's τ values among all questions of reading behaviours, telling correlations between each response. The question (a) is positively correlated with (b)-(d), but the question (e) has almost no correlations to other questions except the question (b).

[3] 26 February 2018 (https://www.univcoop.or.jp/press/life/report53.html — accessed on 26 September 2019)

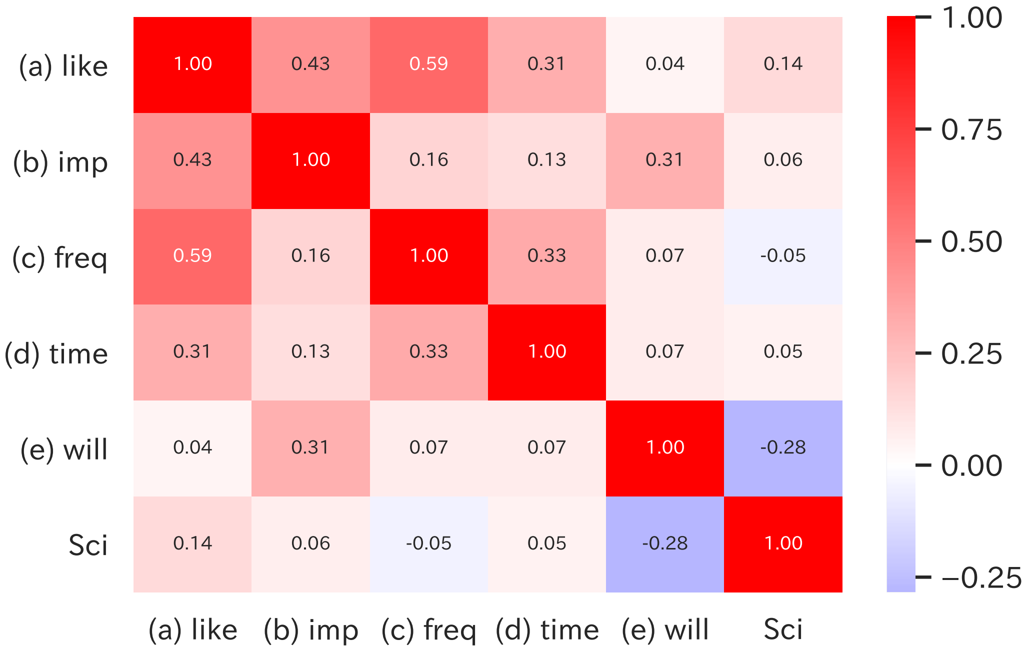


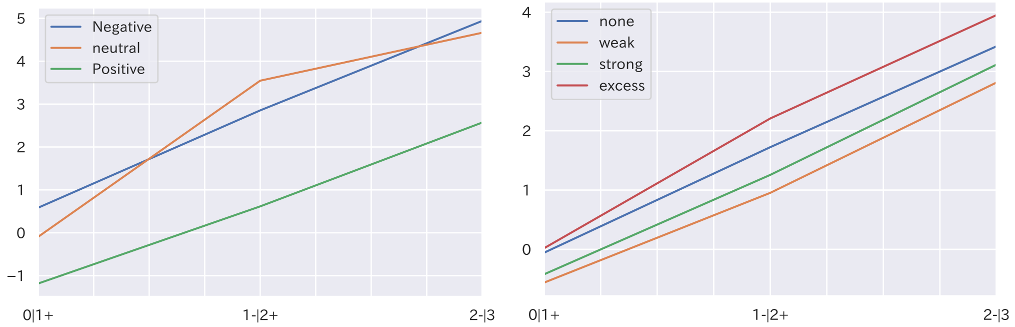
Figure 4.4: Rank-order correlation coefficients (Kendall's τ) among reading behaviour responses. Sci stands for science students (= 1; 0 stands for humanity students).

Finally, we check whether we should take into account the difference between the students' faculties (humanity or science) in our sample. From Table 4.4, there is no large difference in the response distributions as we expected. Also, Figure 4.4 tells that students' faculties (the row 'Sci'; dummy coding of science students = 1 and humanity students = 0) has subtle or no correlations only, except a weak negative association with the question (e). Most of them may come from the relatively smaller sample size (= 10) of the science students. We thus mix all students and analyse the sample as one, rather than we split the data or consider the faculty as an independent variable for inspiringness.

4.3.2 Regression analysis

Modelling

Given the ordinal responses in inspiringness and individual variations among participants and questions, we first consider to apply a mixed effect model using ordinal regression, e.g. cumulative link mixed model (Tutz and Hennevogl, 1996). It is based on the assumption of proportional odds (or parallel lines; Ari and Yildiz, 2014), and this does not hold as shown



(a) Pleasantness

(b) Forcefulness

Figure 4.5: Line plots of logits for cumulative odds of inspiringness responses

in Figure 4.5, which draws log cumulative odds of inspiringness responses with regard to pleasantness are drawn.

Therefore, we also adopt a multinomial logistic regression model as this model is a generalised version of ordinal logistic regression.[4] We compare these models each other to assess the effect of pleasantness and forcefulness.

As a linear predictor (η), we formulate our model into the following equation:

$$\eta = \begin{bmatrix} \mathbf{x}_{\text{pos}} & \mathbf{x}_{\text{neg}} & \mathbf{X}_{\text{pos}\times\text{force}} & \mathbf{X}_{\text{neg}\times\text{force}} \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{Z}_{\text{participant}} & \mathbf{Z}_{\text{message}} \end{bmatrix} \mathbf{u} + \mathbf{e} \tag{4.1}$$

The symbols denote the following meanings:

- $\mathbf{x}$s are design vectors of the independent variables
  - ‘pos’ and ‘neg’ stand for *positive* and *negative* in pleasantness
  - ‘force’ stands for designed forcefulness
- $\mathbf{X}$s are design matrices for the interaction among independent variables
  - $\mathbf{X}_{\text{pos}\times\text{force}}$ — the interactions between positive pleasantness and all forcefulness levels
  - $\mathbf{X}_{\text{neg}\times\text{force}}$ — the interactions between negative pleasantness and all forcefulness levels
- $\boldsymbol{\beta}$ is a vector of fixed effects
- $\mathbf{Z}$s are design matrices for random effects
- $\mathbf{u}$ is a vector of random effects
- $\mathbf{e}$ is a vector of random errors

[4] Another option is to use partial proportional odds models which mitigate the assumption of proportional odds (Sasidharan and Menéndez, 2014). However, they are not capable of observing interaction between pleasantness and forcefulness because they isolate the independent variables that violate proportional odds assumption aside from other independent variables.

Table 4.5: Statistical models for the regression analysis of inspiringness. R stands for the responses of 'reading attitude and behaviour' questions.

	Ordinal model	Categorical model
Use η	Ordinal	Categorical
Use $\eta^{(\mathrm{R})}$	Ordinal w/ R	Categorical w/ R

Note that we omit the notation of intercepts from this equation here, but models hold them.

In this predictor, we define [neutral-none] as the baseline. Our primary interest, the interaction between pleasantness and forcefulness, is included as $\mathbf{X}$. By coding *positive* and *negative* of pleasantness separately ($\mathbf{x}_{\mathrm{pos/neg}}$), we removed non-existing interactions between *neutral* pleasantness and above-*none* forcefulness. The coefficients for $\mathbf{x}_{\mathrm{pos/neg}}$ thus correspond to [positive/negative-none], and above-*none* levels of forcefulness are considered in two $\mathbf{X}$s. Using the dummy (or treatment) coding, we will obtain simple effects.

We additionally examine the effect of participants' reading attitude and behaviour towards inspiringness. This version of models uses another linear predictor $\eta^{(\mathrm{R})}$, i.e. Equation (4.2), which includes the responses of reading attitude and behaviour as independent variables $\mathbf{x}_{\mathrm{Q}(\cdot)}$ where 'Q(·)' stands for the question (·) in which · takes a, b, …, e. We did not consider interactions among them for brevity.

$$\eta^{(\mathrm{R})} = \begin{bmatrix} \mathbf{x}_{\mathrm{pos}} & \cdots & \mathbf{X}_{\mathrm{neg}\times\mathrm{force}} & \mathbf{x}_{\mathrm{Q(a)}} & \cdots & \mathbf{x}_{\mathrm{Q(e)}} \end{bmatrix} \boldsymbol{\beta} + \cdots \tag{4.2}$$

In summary, we consider four combinations of models for the regression analysis as shown in Table 4.5. We refer to these models as the names in this table.

The ordinal models are defined as:

$$\Pr(y \leq k \mid k = 0, 1, 2) = \frac{\exp(\eta)}{1 + \exp(\eta)} \tag{4.3}$$

This model estimates the parameters for the effects along with thresholds (or cutpoints) as intercepts.

In categorical models, Equation (4.4), different linear predictors η_k are independently used because the model separately fits log odds between the baseline level and the other levels in the dependent variable. We thus have three binary logistic regressions in our case because inspiringness responses have four levels. For $k = 1, 2, 3$, η_k is defined as substituting $\boldsymbol{\beta}_k$ for $\boldsymbol{\beta}$ in Equations (4.1) and (4.2). We set the baseline level of inspiringness to 0.

$$\Pr(y = k \mid k = 1, 2, 3) = \frac{\exp(\eta_k)}{1 + \sum_{j=1}^{3} \exp(\eta_j)} \tag{4.4}$$

Parameter estimation

We estimate parameters using Bayesian framework mainly because of its modelling flexibility, the better estimation of parameters, and intuitive understanding of the result (Eager and J. Roy, 2017; J. Lee, E. Walker et al., 2017). For computation, we used the statistical software R (R Core Team, 2019, ver. 3.6.0) and brms (Bürkner, 2017, ver. 2.9.0).

We set prior distributions to non or weakly informative distribution, to be less optimistic.[5] The non informative prior, or the flat distribution, was applied to our ordinal models for the fixed effect parameters. For the categorical models, we used Student's *t* priors with 5 degrees of freedom to the fixed effect parameters. We adopted this weak informative prior for faster and reliable convergence since the categorical models have three times more parameters than the ordinal models. Also note that larger degrees of freedom are known as appropriate for logistic regression (Ghosh, Y. Li and Mitra, 2018). In both models, random effect parameters were estimated by using Student's *t* priors with 3 degrees of freedom.

The models are fitted by the NUTS sampler (Hoffman and Gelman, 2014) using 4 chains each with 2000 iterations, the first 1000 of which was used for warm-up. All models were converged well because potential scale reduction factors (Gelman and Rubin, 1992) were 1.00 for all parameters.

Interpretation

In logistic regressions, we can interpret the plus or minus sign of estimated coefficients directly as the polarity of the effect to inspiringness (increasing it or decreasing it, respectively).[6] Besides, the absolute value of the coefficient means the strength toward inspiringness responses.

The estimated parameter coefficients and their 95% credible intervals are illustrated in Figures 4.6 to 4.9. Note that a 95% credible interval is the 2.5 and 97.5 percentiles of the posterior distribution, interpreted as the range within which the estimated coefficient

[5] We referred to a guide provided by Stan's web page to choose prior distributions: https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations

[6] Strictly speaking, because of using log odds, the plus/minus sign of the coefficient means an increasing/decreasing effect to log odds between certain level(s) of inspiringness and others.

for the parameter stays within 95% probability limits. If a credible interval does not include 0, the sign of the coefficient is 'certain'. We may also express 'probably' positive or negative for an effect if 0 is relatively close to either bound of the credible interval of the effect. If the length of a credible interval is relatively short, the estimation of the coefficient is 'confident'. We use the term 'strong' or 'weak' for describing the relative value of estimates. Using this terminology, we can summarise the result as follows, focusing on our hypothesis about pleasantness and forcefulness:

Ordinal

(1) the effect of *positive* ([positive-none]) is certainly positive and strong

(2) the interaction effect of [positive-excess] is certainly negative and strong

(3) the effect of *negative* is probably negative and weak

(4) the interaction effect of [positive-weak] is probably positive and weak

Ordinal w/ R

(5) the response value 3 of the question (a) is certainly positive, but unconfident

(6) the response value 3 of the question (c) is probably negative and unconfident

(7) the question (d) has confidently no effect

(8) the effect of reading attitude and behaviour is not proportional to their ordinal response values

Categorical

(9) the effect of *positive* is certainly positive and strong in the comparison 0-vs-1 and 0-vs-2

(10) the interaction effect of [positive-weak] is probably positive, but weak, among all comparisons

(11) the effect of *negative* is certainly negative and weak in the response value comparison 0-vs-1

(12) the interaction effect of [positive-excess] is confidently negative and weak in the comparison 0-vs-2

Categorical w/ R

(13) no reading attitude and behaviour have certainty in their effects

Overall

(14) in both ordinal and categorical models, the estimates of the coefficients for pleasantness and forcefulness remains regardless of the incorporation of reading behaviours

Table 4.6: Model comparison using LOO. Numbers after ± are the standard error.

Models	$\Delta\text{ELPD}_{\text{LOO}}$	LOOIC	$\Delta\text{ELPD}_{\text{WAIC}}$	WAIC
Categorical w/ R	—	1571.80 ± 47.29	—	1561.83 ± 46.44
Categorical	−1.30 ± 2.05	1574.40 ± 45.45	−1.90 ± 2.06	1565.64 ± 44.62
Ordinal	−17.13 ± 13.29	1606.06 ± 42.96	−21.91 ± 13.14	1605.65 ± 42.95
Ordinal w/ R	−17.93 ± 13.47	1607.66 ± 43.70	−22.70 ± 13.31	1607.24 ± 43.69

Model comparison We compared these four models by using leave-one-out (LOO) cross-validation (Vehtari, Gelman and Gabry, 2017). Table 4.6 provides the difference of expected log pointwise predictive density (ΔELPD), leave-one-out information criterion (LOOIC), and widely applicable information criterion (WAIC). The higher ELPD means better fit, while the two information criteria are interpreted as the other way around. These values suggest that the categorical models are better than the ordinal models because the differences in ELPD, LOOIC, and WAIC are almost same amounts of their standard errors. However, the difference of whether incorporating reading behaviours or not appears small in contrast of their standard errors.

4.4 Discussion

According to the result of model comparison, we shall put more importance on the categorical models. In contrast, the effect of reading behaviours would be negligible not only because of the result of the model comparison, but also because of the small confidence and certainty in their parameter estimates (e.g. Items (5), (6), (13) and (14)).

With regard to our hypothesis, we can summarise the result as follows:

- *Positive* pleasantness certainly has a strong positive effect to inspiringness (∵ Item (9); also weakly supported by Item (1))
- *Positive* pleasantness and *excess* forcefulness certainly has a weak interaction to decrease inspiringness (∵ Item (12); also weakly supported by Item (2))
- *Positive* pleasantness and *weak* forcefulness probably has a weak interaction to further increase inspiringness (∵ Item (10); also weakly supported by Item (4))
- *Negative* pleasantness certainly has a weak negative effect to inspiringness (∵ Item (11); also weakly supported by Item (3))

These support our hypothesis well and thus ensure inspiringness theory we proposed.

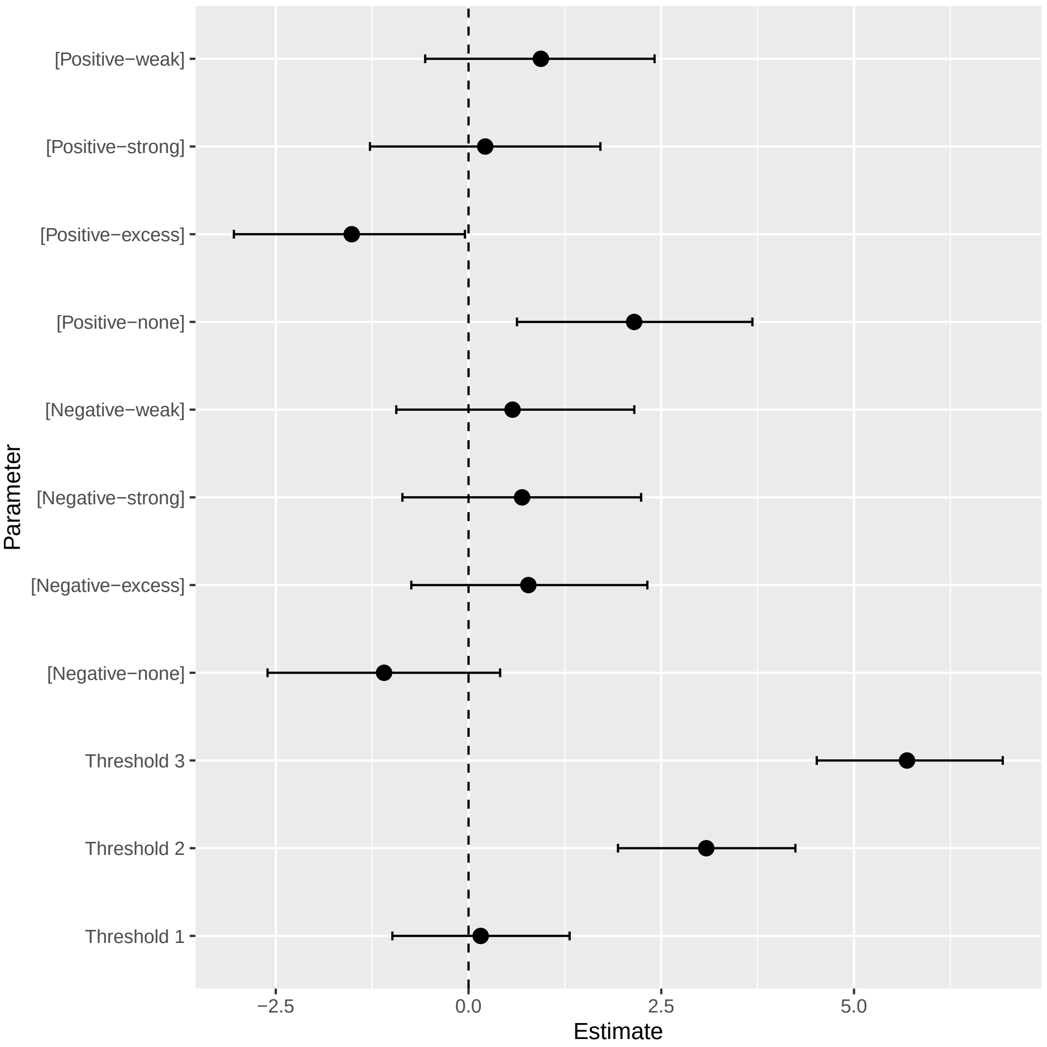


Figure 4.6: Estimated parameter values and their credible intervals of the ordinal model (without reading behaviour responses)

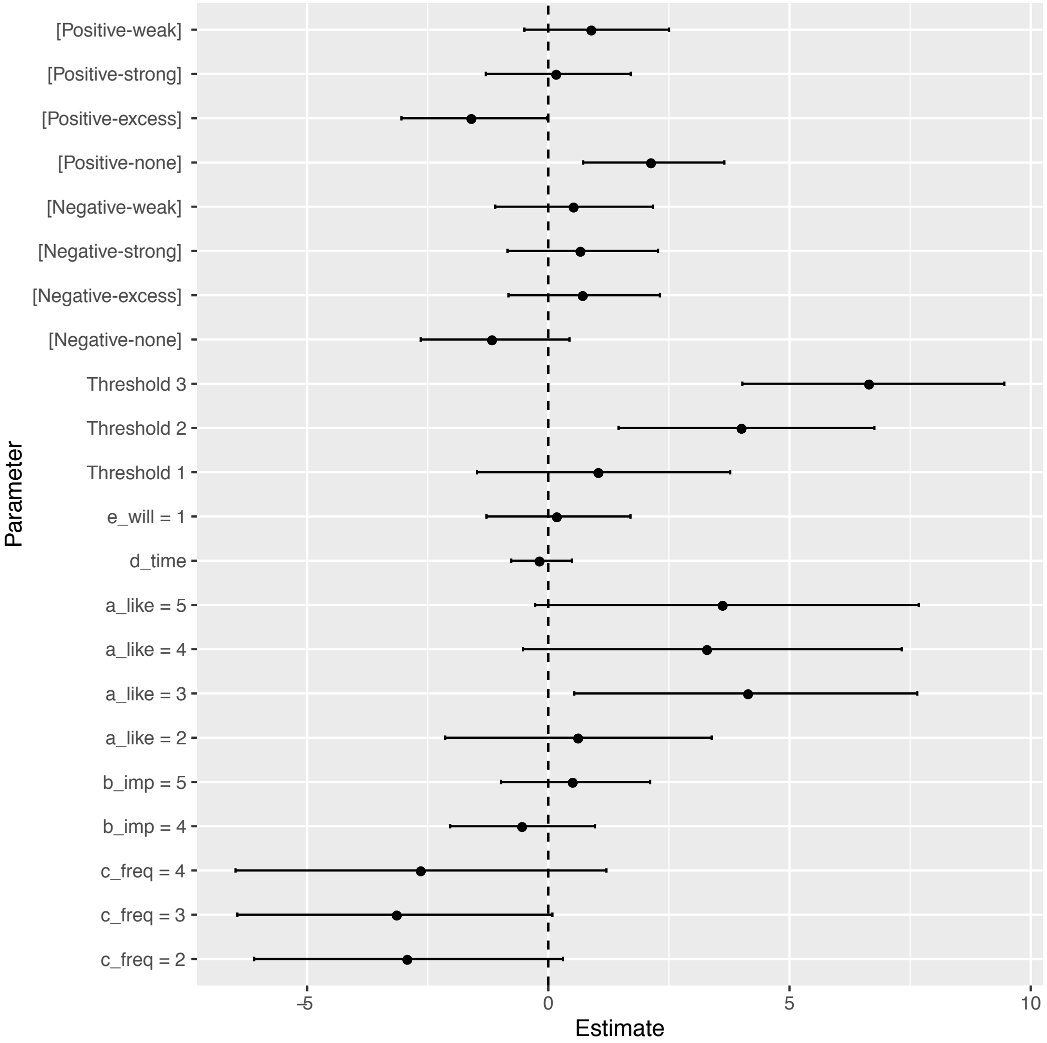


Figure 4.7: Estimated parameter values and their credible intervals of the ordinal model with reading behaviour responses

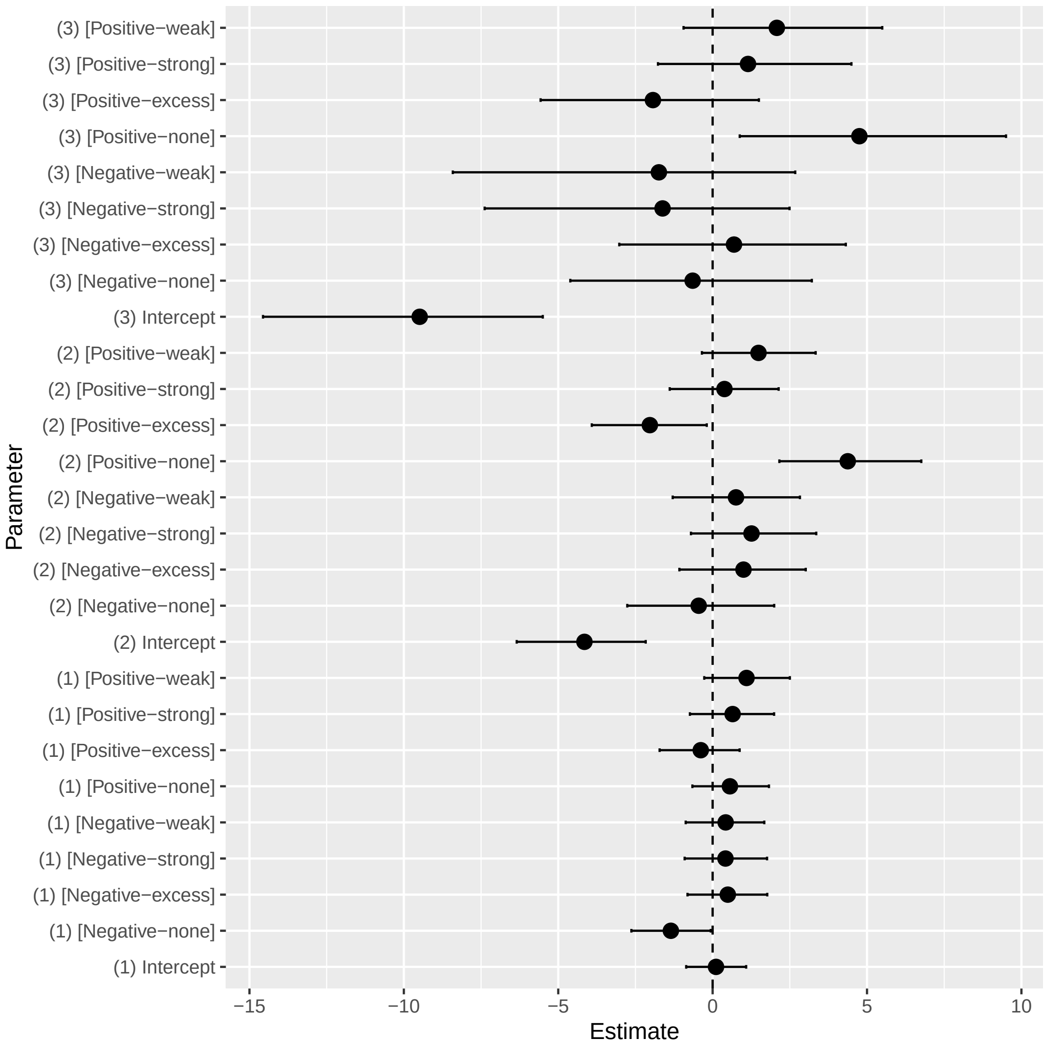


Figure 4.8: Estimated parameter values and their credible intervals of the categorical model (without reading behaviour responses)

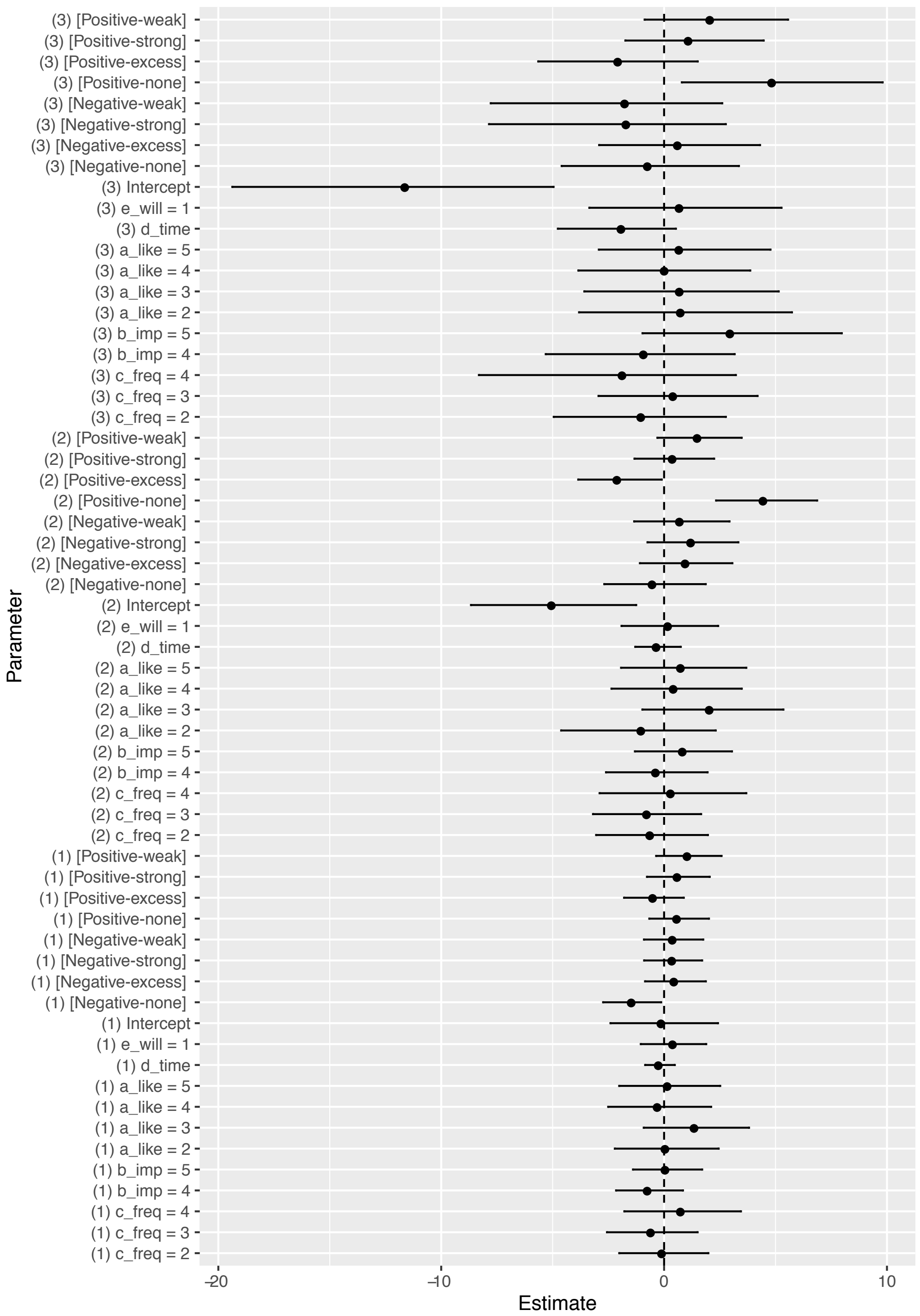


Figure 4.9: Estimated parameter values and their credible intervals of the categorical model with reading behaviour responses

4.4.1 Implication

The results can contribute to the related fields, i.e. psychological reactance and eWoM research. In contrast to existing studies, our experiment scales four levels of forcefulness unlike the typical binary-level setup (i.e. forceful or not; e.g. Miller et al., 2007). It showed that counter-attitude messages shape a convex curve as forcefulness increases. While previous work has reported a linear decreasing under the binary setup, our outcome revealed a finer characteristic of reactance by integrating 'recommendation' or 'suggestion' into the line of forcing or controlling.

We supplementarily measured forcefulness responses (Section 4.2.1) and they seem to be also associated with negative pleasantness (see Figure 4.1). This implies that 'forcefulness' in the daily communication context involves negativity of opinions. Conversely, stating negative comments can be perceived as 'forceful'. While some eWoM research showed the strong effect of eWoM with the negative polarity (e.g. Liebrecht, Hustinx and van Mulken, 2019), our finding identified 'forcefulness' (not exactly the same as our definition of *forcefulness*) as a component of linguistic negativity.

4.4.2 Future directions

Improving the generalisability in demographic properties of the sample can describe the effect of inspiringness and considerateness finer. At least, our results may well apply to the similar demographic groups to our sample, such as other university students in Japan. Can we extend the applicability to further generalised populations?

As we reviewed in Section 4.3.1, the actual reading habit of our sample was reasonably close to Japanese average, though its reading attitude seemed somewhat deviated. We expected that such reading-behaviour variables could be the most influential traits, amongst personalities and demographic properties, towards one's sensitivity to inspiringness. However, it turned out that all reading-behaviour variables, even including the attitude ones (i.e. Question a, b, and e), had very small effects in comparison to pleasantness and considerateness. One expectation is that other characteristics on the sample side could be less influential.

To investigate this point, we need a scalable environment for experiments. Digital surrogate systems for passive exposure to books can serve the role.

Part III

TMB corpora

5 TMB corpus creation

As preparation for core-NLP-module implementation, we collect tweets that mention books (TMBs). Analysing collected TMBs can bring insights for how to identify TMBs (in TMB identification) or how to measure inspiringness of TMBs (in iTMB scoring). While the TMB identifier finds TMBs from general tweets, iTMB scorer measures textual aspects of inspiringness from TMBs. Because their starting points of processing differ, we should prepare suitable datasets for them separately. Using different methods, we compile two datasets: a TMB identification dataset and a TMB inspiringness dataset. We also designed an annotating guideline to label TMBs since human inspection is required to judge TMBs and their inspiringness.

The rest of this chapter is organised as follows. First, we define the range of TMBs we target in this research in Section 5.1. Second, the methods to collect two datasets are explained in Section 5.2. Then, Section 5.3 shows the annotation guideline. Finally, we report the quality of annotation in Section 5.4.

5.1 Target TMBs

The phrase 'tweets that mention books', in a literal sense, can mean a wide variety of mention styles. In this research, we limit TMBs to (Japanese) tweets that mention *specific* books. This is because they will become better exposure to books than general book mentions, especially for infrequent readers, by helping them to determine what books to read. We regard the mentioned book as *specific* if it follows the characteristics below:

Identifiability: We exclude tweets that mention books which cannot be identified due to the lack of information (e.g. only saying 'I read *that* book').

Individuality: We do not target tweets that just state the act of reading in general, or an overview of books from a publisher or an author (e.g. "I'm happy that bookstores these days carry more wrestling-related books!"). Since physical bookshelves offer access to each book, we should consider specific utterances as their online surrogates, as in 'I have just finished reading *Moby-Dick*'. Although abstract or metaphysical discussions about books or readings may be regarded as reading cultures or behaviours, we believe that they will not correspond to an encounter of an individual book.

We define 'books' as all published or planned-to-publish books, including e-books. Tweets that mention the followings are, therefore, excluded:

- magazines,
- web-only articles, and
- private (not publicly distributed) books

Note that the definition of 'books' follows UNESCO (1965) as we declared in Section 1.7.1.

Finally, we handle tweets posted by individuals because we are interested in the surrogate for passive exposure to books provided through human communications. Given that there is the substantial number of automatically generated tweets in Twitter (A. Wang, 2010), we should filter them out in the TMB collection step (cf. Section 3.2.2).

In summary, our target TMBs are *manually generated Japanese tweets that mention specific books.*

5.2 Collection methods

We prepare two datasets: a TMB identification dataset and a TMB inspiringness dataset. Whereas the TMB identification dataset should contain both TMBs and non-TMBs, the TMB inspiringness dataset can focus on TMBs. We apply different methods to collect TMBs for these datasets.

5.2.1 TMB identification dataset

This dataset will be used mainly for obtaining insights about TMB identification and for evaluating our TMB identifier (Chapter 7). In TMB identification, we must distinguish TMBs from general tweets. We use book titles as the key information.

Specific books can be referred to by several bibliographic elements, such as titles, authors, and/or publishers. Amongst them, *titles* are intrinsically the core information to identify books. Titles are also the most identifiable and memorable for humans in comparison to, e.g. International Standard Book Numbers (ISBN), even though ISBNs are designed to uniquely identify books. For example,

> *The Girl on the Train* is a surprisingly good book. A real page-turner with very interesting character at the center. [a real English tweet example]

A prior preliminary analysis has also shown that almost all tweets use titles to refer to specific books.[1] Since TMBs using book titles are included in tweets that contain title strings (TCTSs), where *title strings* denote textual strings that exactly match existing book titles, we can start collecting TMBs from TCTSs.

We prepared about 10,000 TCTSs for our experiments. This is the largest boundary of data sizes in previous studies, and it is still appropriate for human annotation. Note that most of similar tasks, which will be introduced in Section 7.2, handled around 1,000–10,000 annotated tweets to classify. The procedure to collect 10,000 TCTSs is as follows:

1. create an exhaustive list of title strings,
2. collect a large amount of TCTSs based on the list, and
3. choose a subset of 10 thousand tweets, in such a way as to contain as many but diverse TMBs as possible.

In Step 1, the list of comprehensive book titles was compiled from a Japanese bibliographic database.[2] It contains 1,421,556 titles of books written in Japanese and published by 2016.

In Step 2, we used this list to collect 74,330 TCTSs in total, during the period from 30th April to 5th May 2015.[3] We employed the Twitter *stream* API[4], which randomly samples the Twitter public stream, instead of the *search* API[5], because querying titles to the search API suffers from rate limits due to the huge number of titles.

In Step 3, **10,791** tweets that contain a title that appeared less than three times were selected from 74,330 TCTSs. Based on a preliminary investigation, we found that a large amount of non-TMBs existed and that the frequency of mentioned titles was distributed in a manner resembling Zipf's law. That is, a small number of titles occur quite often while a large number of titles appear very rarely. Tweets containing less frequently mentioned title-strings can cover a wide variety of TMBs, and are likely to contain relatively fewer non-TMBs.

After annotating these tweets for TMB or not, the way of which will be explained in Section 5.3.1, we obtained **450** TMBs, **10,341** non-TMBs. In this data set, there are 441 distinct users in TMBs, 9,042 in non-TMBs. Table 5.1 summarises these counts.

[1] From the Twitter public stream, we collected 2,258 tweets that contain the word '読了 (finished reading)' in Japanese between 12–22, September 2016 and annotated a 10% sample (226 tweets). We obtained 211 (93%) mentions of book titles but only 122 (54%) author names and less than 100 other bibliographic fields.

[2] We used Webcat Plus which unifies the wide range of bibliographic databases, such as the national library (National Diet Library, Japan), university libraries, and commercial book catalogues in Japan.

[3] Since some TMBs mention books to be published in the near future at the period of making the tweets, we used the book title list that contains books published by 2016.

[4] https://stream.twitter.com/1.1/statuses/sample.json

[5] https://api.twitter.com/1.1/search/tweets.json

Table 5.1: TMB identification dataset

	Num. of tweets	Num. of accounts
TMB	450	441
non-TMB	10,341	9,042
Total	10,791	9,483

5.2.2 TMB inspiringness dataset

This dataset will be used for analysing TMBs (in Chapter 6) and for evaluating our iTMB scoring modules. As shown in Table 5.1, it turned out that even TCTSs does not contain a large portion of TMBs. We should collect more TMBs in order to investigate their characteristics. We adopt hashtag search for this demand, which is popularly used in Twitter studies (Bosco, Patti and Bolioli, 2013; Gonzales, 2014; Graells-Garrido, Baeza-Yates and Lalmas, 2019; A. Kim et al., 2017; Mohammad, Kiritchenko and J. Martin, 2013). Using book or reading-related hashtags may gather TMBs with higher precision.

Based on a prior preliminary analysis, i.e. searching Twitter for book-related hashtags, we selected the following four hashtags, which seems to be contained in relatively many TMBs: #読書 (reading), #読了 (finished reading), #書評 (book review), and #本 (books).

Considering a huge proportion of bot tweets, in addition, we included tweets posted by official Twitter applications only, since they are not designed for bot-like automation. This simple rule-based procedure was derived from the result of TMB identification Section 7.5.2. Although this may also remove TMBs produced via online social reading services (see Section 2.1), we prioritised hand-crafted TMBs than them because of their fixed format and the ease of their collection.

From 16 June to 5 August 2018, 9,997 unique tweets were obtained. After annotating these tweets for TMB or not, the way of which will be explained in Section 5.3.1, we obtained 8,198 TMBs consisting of 2,426 #読書 (reading), 4,626 #読了 (finished reading), 411 #書評 (book review), and 1,517 #本 (books). Note that these numbers do not add up to 8,198 because some tweets contain multiple tokens of these hashtags. TMBs in this dataset were posted from 2,818 unique users as shown in Table 5.2. Over half of the users (0.59%) just posted 1 TMBs, and 90% of users generated less than 7 TMBs in this dataset. The maximum count of TMBs posted by one user was 103. These TMBs mentioned 5,457 book titles, 4,307 (78.9%) of which were mentioned only once. The most frequently mentioned book title appeared 37 times.

Table 5.2: TMB inspiringness dataset

	Num. of tweets	Num. of accounts
TMB	8,198	2,818
non-TMB	1,799	1,097
Total	9,997	3,925

5.3 Design of annotation guideline

Our annotation scheme consists of two layers: TMB-or-not and inspiringness annotations. First, annotators label whether the tweet is a TMB or not. Then, for TMBs only, textual aspects of inspiringness are labelled.

5.3.1 TMB-or-not annotation

As the first layer of this annotation scheme, we distinguish TMBs from non-TMBs. The definition of TMBs follows what we specified for *the target TMBs* in Section 5.1.

We asked annotators to search for book information if it is necessary to judge whether the mention is TMB or not. This is because what is mentioned in a tweet is often hard to interpret. For example, substantial entities, e.g. films and songs, are entitled the same as book titles due not only to coincidence but also to transmedia franchises. We regard tweets that only mention non-book versions of certain books as non-TMBs. In the TMB identification dataset, which comprises TCTSs, another kind of non-TMBs appearing frequently comes from the fact that book titles often consist of ordinary expressions (e.g. *Kidnapped*, *See Me*, etc.).

5.3.2 Inspiringness annotation

For TMBs found through the TMB-or-not annotation, annotators are asked to label inspiringness of TMBs. In this stage, textual aspects of inspiringness are targeted because the data will be served for iTMB scoring. The main components to annotate here are *pleasantness* and (textual) *considerateness*. We will formalise these concepts to make the annotation feasible. Additionally, the *purpose* of tweeting about books is also annotated for a better understanding of the TMB activity, since we have almost no idea with the reason why people mention books in Twitter.

If a TMB mentions multiple books, each one of the books is the target of these annotations. We define a pair of a TMB and one of books mentioned in the TMB as a (TMB) *record*, and let annotators label records rather than TMBs.

Pleasantness

Pleasantness is ‘the joyfulness of the message author towards the mentioned target (i.e. books) perceivable from the message’ (Section 3.1). We asked annotators to label the tweet author’s opinion towards the books mentioned in a tweet, based on the tweet text. Opinions can be defined in several ways, such as the polarity (e.g. positive vs. negative) and the intensity (e.g. 1–5 point scale) (B. Liu and L. Zhang, 2012). Basically following the ‘polarity’ scheme as we did in Chapter 4, we experimentally consider the balance between positive and negative opinions in the TMB. The opinion polarity, or a discrete valence system, can stabilise the annotation quality, while some mentions like book reviews may describe both good and bad parts of the mentioned book. The following six values, therefore, are defined:

Positive: the author of the TMB seems to like the book, probably because it is interesting, fun, and/or valuable. E.g. “I’m crazy about TITLE. I love this one best of all volumes in the series. I cannot describe my feeling other than love.” [originally in Japanese]

Negative: the author of the TMB seems to dislike the book, probably because it is uninteresting, boring, and/or not valuable (i.e. opposite of positive). E.g. ‘I have finished reading TITLE written by AUTHOR. I don’t like it at all. [originally in Japanese]’

Neutral: unable to decide positive or negative from the TMB alone since no such expressions appear. E.g. ‘I have finished reading TITLE.’ [an artificial example]

Positive > Negative: the TMB has both positive and negative opinions, and the positive part is superior. E.g. ‘Though the theme of this book was mediocre, its characters and story were quite interesting.’ [an artificial example]

Positive < Negative: the TMB has both positive and negative opinions, and the negative part is superior. E.g. ‘I like some of the characters in this novel. But the story was getting worse and worse. I quit reading it before the last chapter.’ [an artificial example]

Positive = Negative: the TMB has both positive and negative opinions, both of which are balanced. E.g. ‘TITLE. Fun story, but a bit lengthy. Rated 3/5.’ [an artificial example]

Textual considerateness

Considerateness is defined as ‘the moderateness of the exposure for the user, or how less forceful the exposure is to the user’ (Section 3.1). Especially, in the TMB inspiringness annotation, we target *textual considerateness*: how forceful expressions are used in TMBs. Considering what the situation looks like, where TMB authors use the phrases with a little or more pressure on TMB readers, the purpose of *recommending* books should be the one. In a forceful situation, a TMB may order recipients to read a book. We therefore focus on recommending phrases used in TMBs in textual considerateness annotation.

If annotators find TMBs with recommending phrases (or simply, recommending TMBs), we asked them to extract the phrases and to label its strength. Furthermore, we asked annotators to extract its target audience if any is explicitly declared, and to label the scale of the audience. This is because the target audience may affect forcefulness or considerateness of a recommendation.

The recommending phrase: extract the phrase that recommends the book. If another phrase that strengthens the recommending phrase is found near the recommending phrase, include all of them as one phrase (e.g. 'I *totally recommend* TITLE *without any hesitation!*' [originally in Japanese]).

The strength of the recommending phrase: choose a level of the strength of the recommending phrase; we set three levels taking into account psychological reactance.

Weak: the recommendation is made in a moderate way, such as 'recommend', 'want you to read', and 'good to read'.

Strong: the recommendation is strong or almost forcing the reader to read the book, e.g. 'had better to read', 'must read', and 'should read'.

Excess: the forcefulness of recommendation is excessive, leading to unpleasant phrases such as:

- insulting those who have not yet read the book (e.g. 'absurd of you not to read it' [an artificial example])
- putting strong peer pressure to read the book (e.g. 'of course you wanna-be musicians have already read the book, right?' [an artificial example])

The target audience for the recommendation: extract the expression describing the audience of the recommendation in recommending TMB if any (e.g. '*Young people* should read TITLE by AUTHOR' [originally in Japanese]). For brevity, we also refer to this as *the recommendation audience.*

The scale of the recommendation audience: choose the scale or range of the target audience for the recommendation from the following categorisation (*italic* phrases corresponds to the recommendation audience.):

Individuals: the book is recommended to a specific person or a couple of specific people (e.g. '*@user* I recommend my favourite, TITLE!' [originally in Japanese])

Specific group: the book is recommended to a group of specified people (e.g. 'TITLE is a must-read book for *those who are going to launch new venture.*' [originally in Japanese])

Everybody: the book is recommended (a) to everyone (including vaguely specified large profiles such as 'all humanity' and 'human race'), or (b) to no one in particular, i.e. no target is specified. The case (b) is covered because it seems to recommend the book at least to all followers of the author of the TMB (e.g. 'TITLE is still interesting. *Everyone* let's read this story which begins from CHAPTER.' [originally in Japanese])

Purpose

This is a supplemental annotation target aside from inspiringness, which can contribute to the understanding of iTMBs. When we formalised textual considerateness, we pay attention to the *purpose* of mentioning books in TMBs. Other than recommendation, there should be more purposes for TMBs.

We identified six purposes for TMBs, based on an open coding in a preliminary analysis. All examples below are originally in Japanese.

To share a review (review): sharing a review of the mentioned book (e.g. 'I have read TITLE by AUTHOR. It's really fun to read.').

To report an action (report): reporting the action of reading or buying the mentioned book (e.g. 'I have just finished reading TITLE.').

To recommend (recom): recommending or suggesting the book for someone (e.g. '`@user` I absolutely recommend to you TITLE by AUTHOR!!').

To advertise (ad): advertising the book, e.g. as authors or publishers, for marketing purpose (e.g. '[Now on sale] TITLE. A dark Gothics fantasy!').

To express expectations (expect): showing expectations of a future action or situation about the book (e.g. 'I would love to get TITLE and TITLE! Definitely!').

To cite or refer to (refer): referring to or citing the book or its content for other purposes (e.g. 'A flower was blossoming by the green road near from Hibiya park. According to "TITLE", it came from China at Edo period and was planted on gardens and shrines. It smells like banana, delicious!').

We asked annotators to choose the most applicable label. Although several categories might apply to a TMB (like *report* and *review*), This single label system can make the annotation more accurate than multiple labelling. Since *review* and *report* are expected to be frequent, multiple labelling would make annotators bored, leading a lower annotation quality.

5.4 Labelling quality

Both datasets were manually annotated by five people in total. The result of TMB-or-not annotation for both datasets were reported earlier. We will elaborate on the outcome of inspiringness annotation in Chapter 6. This section describes the quality of our annotation through inter-coder agreements.

5.4.1 Quality of TMB-or-not annotation

We assessed the guideline for TMB-or-not annotation by measuring inter-coder agreement on a sampled subset of the TMB identification dataset. Two people (including the author) annotated the same 5% set of the TCTSs (i.e. 540 tweets). We observed 0.775 in Cohen's κ (J. Cohen, 1960). Given the high level of inter-coder agreement, which satisfies the standard for the quality of computational linguistic corpora proposed in Artstein and Poesio (2008), the remaining tweets in the TMB identification dataset were annotated by one person only, i.e. the author.

5.4.2 Quality of inspiringness annotation

Similarly, the TMB inspiringness annotation was evaluated on a sample subset of the TMB inspiringness dataset. In annotation, the TMB inspiringness dataset was split into four parts, each of which was labelled by a different external annotator (i.e. five distinct annotators in total). To measure the quality of annotation, the author also annotated 200 TMBs (around 2%) of the TMB inspiringness dataset. The sampled subset consisted of four sets of 50 TMBs, each of which was sampled separately from the corresponding four splits. Inter-coder agreement was calculated in J. Cohen (1960)'s κ, and we obtained the following values:

- 0.774 in pleasantness,
- 0.700 in the strength of recommendation phrases,
- 0.824 in the scale of recommendation audience, and
- 0.801 in the purposes.

From the standard interpretation, these values can be regarded as 'substantial' agreements (Landis and Koch, 1977). We thus regard the annotations on both datasets as coherent.

6 Descriptive analysis of TMB corpora

We collected TMBs via two different method, i.e. book title-based and hashtag-based. In order to realise the core NLP modules, we must investigate the characteristics of TMBs. Remember that we have almost no idea with how people use Twitter to share their reading behaviours with their online friends. Especially, inspiringness has not yet been investigated over TMBs. In this chapter, we focus on examining inspiringness of TMBs across two datasets: the TMB identification dataset and the TMB inspiringness dataset. Note that the difference between TMBs and non-TMBs will be reviewed in Chapter 7, where the task of TMB identification is tackled.

6.1 Objective

We study the linguistic characteristics of TMBs across two datasets: the TMB identification dataset and the TMB inspiringness dataset. Starting from the fundamental properties such as character counts and frequent words, we aim to reveal the characteristics of inspiringness of TMBs. For TMBs, we annotated pleasantness and textual considerateness, as well as the purpose of mentioning books. Describing differences amongst distinct levels of such attributes will bring insights that contribute especially to the implementation of iTMB scoring modules. The analyses is made with both quantitative (label counts, tweet lengths, and keywords) and qualitative (bottom-up categorisation) methods, and showed that TMBs can be a good source of online exposure to books.

The rest of this chapter is organised as follows. In Section 6.2, existing findings from related studies to the attributes we investigate. Then, we explain the detail of the datasets we use in Section 6.3. In Section 6.4, we give results of the quantitative analysis on TMBs including label counts for pleasantness, considerateness, and purposes, which are not reported in Chapter 5. Next, Section 6.5 provides a fine sight on textual considerateness. Section 6.6 concludes these analysis.

6.2 Known findings related to TMB attributes

6.2.1 Opinion polarity

In relation to pleasantness, we refer to the studies of electronic Word of Mouth (eWoM) and the research about opinion mining on Twitter.

The distribution of the opinion polarity of eWoM depends on topics and platforms. Chevalier and Mayzlin (2006) studied online book reviews as eWoM and found that reviews were dominantly positive over different book review sites. Pang, L. Lee and Vaithyanathan (2002) published a movie review dataset created from IMDB[1] in which positive reviews were 1.7 times more frequent than negative reviews. Van de Kauter, Breesch and Hoste (2015) created a sentiment analysis corpora made of Belgian financial newspaper articles where the ratio of positive : neutral : negative almost equals to 3:2:2.

Opinion mining, or automatically detecting opinions towards entities, is also popular in Twitter. For instance, Semantic Evaluation (SemEval) workshops have held a series of shared tasks on Twitter sentiment analysis (Nakov, Ritter et al., 2016; Nakov, Rosenthal et al., 2013). Note that these tasks define 'sentiment' as the opinion polarity towards certain entities, while 'sentiment' may be defined as differently from 'opinions' in some scenarios (B. Liu and L. Zhang, 2012). In the datasets prepared for these tasks, in which various topics were mentioned, the distribution of the sentiment polarity varies from 2:3:1 to 2:2:1 in positive : neutral : negative. Guzman, Alkadhi and Seyff (2017) collected 1,000 tweets that mention software and annotated the sentiment with a five point scale from *very negative* to *very positive* including *neutral* as the mid point. Neutral tweets consisted of 85% of the data, which was believed to be caused by the high proportion of bot tweets. For other examples, positive messages among tweets reviewing resort spots outweighs negative ones (Philander and Zhong, 2016), while eight times more negative tweets than positive tweets were found over the mentions about antibiotics in livestock (Steede et al., 2018). From these findings, we conclude that the polarity distribution in tweets or eWoM is dependent highly on topics.

Automated sentiment analysis and opinion mining of tweets are still a challenging problem, because of their informal, short, and noisy characteristics. The top score of SemEval 2016 Task4 (Nakov, Ritter et al., 2016), an aspect-based opinion-polarity classification task of tweets, was 0.633 $F1$-score (between positive and negative labels). The Google Cloud Natural Language API[2] is one of only a few tools that support Japanese sentiment analysis and opinion mining with good performance. The API also provides an entity-level

[1] https://www.imdb.com/

[2] https://cloud.google.com/natural-language/

sentiment analysis method which can be used to detect the opinion of mentioned books, but not yet available for Japanese inputs. Since we thus cannot rely on automatic methods, we annotated our tweets manually (see Section 5.3).

6.2.2 Recommending messages

How often do messages that recommend specific entities appear in social media? As we will mention in Section 6.2.3, in Twitter, recommending tweets constitute 2–20% of tweets that mention certain entities, depending on topics (Vosoughi and D. Roy, 2016). They are known as positive stimuli for receivers to accept the mentioned entities. J. Huang et al. (2012) revealed that eWoM can improve not only the prior expectations of recommended items, but also the posterior evaluations of recommended items.

From the perspective of linguistic features of recommending messages, many studies in the clinical field confirmed that strong phrases are more likely to change receivers' attitudes and behaviours (Akl et al., 2012; Buller et al., 2000). Conversely, recommendations that are too strong may cause psychological reactance (S.S. Brehm and J.W. Brehm, 1981) for receivers; receivers may feel repulsion against the direction given by such messages. Psychological experiments to prove this phenomenon often use so-called *controlling language* (e.g. ‘should’, ‘ought’, ‘must’, and ‘need’) as *strong* or *forceful* expressions in contrast to *weak* phrases like suggestions and recommendations (Dillard and Shen, 2005; Miller et al., 2007; Quick and Considine, 2008).

Despite the richness in research that evaluates the effect of recommending messages, there is less work that analyses their descriptive characteristics. Packard and Berger (2016) found that *novice* consumers used explicit endorsements (e.g. ‘I recommend it’) more often than implicit endorsements (e.g. ‘I liked it’ and ‘I enjoyed it’), which was the other way around in *knowledgeable* consumers. Another research (Labrador et al., 2014) made a rhetorical structure analysis of online advertisement of digital cameras, video-cameras, television sets, e-book readers, and digital frames. It reported that online advertisements are written in an informal style characterised by second person pronouns (‘you’), imperatives (‘surround yourself with …’), contractions (“you're”), and puns (e.g. camera: ‘a lot of memory for lots of memories’). No studies, however, focus on the variety and strength of recommending phrases. To the best of our knowledge, our research is the first attempt to describe what kinds of book recommendations Twitter users face.

The target audience, which we take into account, is a necessary component of recommendation. Imajo (2012) found that the awareness of being included (or not) in the target audience of recommending messages affects psychological reactance of recipients. In general, we can assume that recommending messages in direct conversations should target the recipients explicitly or implicitly. If messages are publicly oriented, the range of the target audience may become vague. For example, as Kitamura, Sasaki and Kawai (2016) found, 36.8% of users who post tweets reviewing media contents more than once in a week (110 people) were not conscious of the target audience. Other than this research, there is barely any research that investigates the range and the variety of the target audience in public recommending messages like tweets.

6.2.3 Purposes of messages

In the field of linguistics, message purposes have been handled by *dialogue* (or *speech*) *acts* in which the intention of utterances is categorised. Several domain-independent taxonomies have been proposed in linguistic philosophy research such as Austin (1962) and Searle (1975). Often based on these, task-oriented taxonomies have been developed in computer science fields, e.g. instant chat messages (Ivanovic, 2005; S.N. Kim, Cavedon and Baldwin, 2012) and web forums (S.N. Kim, L. Wang and Baldwin, 2010). Such work also aims to predict dialogue acts of sentences in targeted corpora. Some studies label dialogue acts in tweets. Vosoughi and D. Roy (2016) proposed five categories for tweets following the Searle (1975)'s taxonomy as *Tweet Acts*: Assertion, Recommendation, Expression, Question, Request, and Miscellaneous. They reported the difference in the frequency of categories among types of tweets. Tweets that mentioned specific entities consisted of 52% Expression and 34% Assertion, whereas tweets that mentioned events contained 47% Assertion and 36% Expression. Recommendation constituted only 3% of both types of tweets (with entities and with events). In contrast, tweets that mentioned long-standing topics (e.g. cooking and travelling) consisted more of Recommendation (23%; the second most frequent, next to Assertion). R. Zhang et al. (2013) also adapted Searle (1975)'s taxonomy for their experimental data: Statement, Question, Suggestion, Comment, and Miscellaneous. They also reported a similar difference found in Vosoughi and D. Roy (2016).

Although no research excluding ours handles TMBs, some papers focus on tweets related to specific topics. Oraby et al. (2017) investigated customer service conversations in Twitter, and proposed a finer-grained tagset based on Ivanovic (2005) and S.N. Kim, Cavedon and Baldwin (2012). It has two levels, the higher one of which consists of Greeting, Statement, Request, Question, Answer, and Social Act. The second level has 24 labels in total. The top five frequent tags are statement_info, request_info, statement_complaint, statement_expr(ess)_negative, and statement_suggestion.

Compared with the general dialogue act categories, the purposes we defined in Section 5.3.2 are more specific to the TMB domain. For example, we define *recom* and *ad* separately, while they would correspond to 'Recommendation' in Tweet Acts (Vosoughi and D. Roy, 2016). To distinguish recommendation made by individuals (eWoM) from intentional marketing messages matters in this research because conversation-based passive exposure to books are concerned. Marketing messages in social media are often diffused as embedded ads as well, which may well give users an impression different from eWoM.[3]

Apart from dialogue act frameworks, a Japanese study (Kitamura, Sasaki and Kawai, 2016) describes the purposes behind Japanese tweets. According to the result of a questionnaire survey on 730 Japanese Twitter users a substantial amount of tweets are posted without conscious reasons or target audience. Over a half of 21 tweet categories defined by Kitamura, Sasaki and Kawai, 15% of users generate tweets in that manner, e.g. tweets about weather, greetings, and expressing free time. Among these 21 tweet categories, tweets that review media contents (e.g. films, TV shows, and books) are related to TMBs. The study reports that 299 out of 730 users (40.96%) produce these tweets more than once in a week, and that users who often find it fun to tweet this category (or with the *consummatory* reason) constitute 39.1% of the 299 users. The survey further conducted a quantitative textual analysis of the reasons to tweet provided in free forms of the questionnaire. The result showed that the phrase '共有したい (want to share)' appeared most frequently over 12 out of 21 tweet categories. Furthermore, the phrase 'お薦めしたい (want to recommend)' was characteristically frequent in tweets that review media contents. This suggests that reviewing and recommendation would also be found in the purposes of TMBs.

6.3 Data

Among the TMB identification dataset (Section 5.2.1), we extracted 332 TMBs posted via official Twitter applications, to match the condition with TMBs in the TMB inspiringness dataset (Section 5.2.2). Henceforth, we refer to these 332 TMBs as the dataset *S* (i.e. small), and all TMBs in the TMB inspiringness dataset as the dataset *L* (i.e. large). We split TMBs in both datasets into TMB records, i.e. pairs of a TMB and one of the mentioned books in the TMB (see Section 5.3.2). The dataset *S* has **371** records, while *L* contains **9,127** records.

6.4 Quantitative analysis

We examine our datasets from the following perspectives: purposes, opinions, and recommending TMBs. For each perspective, we first provide the label counts to describe their distributions. Second, we look into the textual characteristics among labels. We measure

[3] For instance, ads in social media contributes to dissatisfaction of users (ACSI, 2018).

Table 6.1: The representative values of the tweet character lengths among several Japanese tweet datasets

Dataset	**Mean**	**Median**	**Mode**
Dataset *S*	61.7	56	44
Dataset *L*	87.6	98	110
Neubig and Duh (2013)	45	—	—
(Twitter blog[4])	—	—	15

tweet lengths as estimates of the textual format/style, and characteristic words (keywords) for quantitative analysis of contents.

6.4.1 Overall TMBs

Before diving into each perspective in detail, we describe the textual characteristics of overall TMBs.

Tweet Length We measured the length of TMBs as an simple approximation of the amount of information content. To calculate the length, we removed URLs, @-reply tokens, and hashtags from tweet texts, as well as title strings. The median length of titles was 8 in *S* and 7 in *L*. We counted characters in the text as the length unit for tweets. This procedure was also applied to succeeding analyses.

The representative values of the whole dataset are shown in Table 6.1 These values are relatively longer than general Japanese tweets. For example, the *mean* number of characters per Japanese tweet was reported as 45 in Neubig and Duh (2013), whereas ours are 62–88. Also, according to the Twitter official blog[4], '[m]ost Japanese Tweets are 15 characters', or the *mode* value was only 15.

Comparing our datasets with each other, the tweet length of *L* is almost twice as long as that of *S* (in, e.g. median and mode). Since longer tweets would require more intention and effort, this difference indicates that *S* may contains more instant and casual tweets, and that *L* may consists of more information-rich, designed tweets. We will take into account this nature of our datasets for succeeding analyses.

Keywords We counted content words in the dataset *L* in order to describe the overall topical characteristics. This research adopt nouns, verbs, and adjectives as content words and removed Japanese stop-words listed in Kokubu, Yamazaki and Nosaka (2013). URIs,

[4] http://bit.ly/2fQ2b7W, published at 27 September 2017.

Table 6.2: The top 20 frequent content words that appear in dataset L

1–10	**Count**	11–20	**Count**
読む (read)	3536	感じる (feel)	386
本 (book)	1598	読書 (book reading)	341
読了 (finished reading)	931	シリーズ (series)	327
面白い (interesting)	840	生きる (live)	325
作品 (work)	773	心 (mind)	324
物語 (story)	592	書評 (book review)	322
主人公 (hero)	448	買う (buy)	308
小説 (novel)	409	著者 (author)	289
世界 (world)	398	描く (draw)	284
最後 (last)	391	お話 (tale)	279

Hashtags, @-reply tokens, and book titles were also filtered out. Japanese sentences were tokenised and lemmatised by using a Part-of-Speech and morphological analyser `MeCab`[5] with a neologism-enhanced dictionary (Sato, 2015).

Table 6.2 lists top 20 frequent words in dataset L. We can see that book/reading-related words appear many times, such as '読む (read)', '本 (book)', '読了 (finished reading)', '作品 (work)', '物語 (story)', and '小説 (novel)'. Other words in this list seem to come from review or summary text, e.g. '面白い (interesting)', '世界 (world)', and '主人公 (hero)'.

6.4.2 Purpose

Although our primary focus is on inspiringness (pleasantness and considerateness), we start particular analyses from the purpose because it can be regaeded as a fundamental nature of TMBs.

Label counts

We start by observing the distribution of purposes of Twitter users for mentioning books. In general, most Japanese tweets report what users have done or felt, whereas substantial users post tweets that review media contents more than once in a week (see Section 6.2.3). Do such reviews constitute the most of TMBs? Or, can we see users' reports about reading behaviours more often?

Table 6.3 shows label counts and the proportion of purposes.[6] We can see that *review* tweets rank at first among both datasets, which means that reviewing is the most frequent

[5] http://taku910.github.io/mecab/

[6] We excluded one record (tweet) from L that was not covered by the six categories of TMB purpose. In the record, the tweet author *asked* audiences for their reviews of a book. Although we are able to create a new *question* label, no other similar tweets were found in either dataset.

Table 6.3: The number and the proportion of TMBs for each class of the purposes to mention books.

(a) Dataset *S*

Purpose	Count	Percentage
Review	114	30.73
Refer	98	26.42
Report	78	21.02
Expect	30	8.09
Recom	27	7.28
Ad	24	6.47
All	371	100.00

(b) Dataset *L*

Purpose	Count	Percentage
Review	5191	56.88
Report	1707	18.70
Ad	965	10.57
Refer	605	6.63
Recom	426	4.67
Expect	232	2.54
All	9796	100.00

purpose of users in Twitter for mentioning specific books in Twitter. The distributions of purposes between *S* and *L* are, however, largely different. The dataset *S* can be divided into two parts in which belonging purposes are almost evenly distributed:

- *review*, *refer*, and *report* (around 30%, 26%, 21%; ≈ 80% in total)
- *expect*, *recom*, and *ad* (around 8%, 7%, 6%; ≈ 20% in total)

In contrast, *L* is characterised by the more than a half proportion of *review* and by a 2^{-x}-like decrease in the succeeding purposes. The proportion of *refer* and *expect* is much fewer than that of *S* (= 8% vs. 2.5% in *L*), whereas the almost twice as much amount of *ad* is also characteristic to *L*.

The casualness of *S* is shown straightforwardly in much more amount of the *expect* purpose. Furthermore, *Refer* tweets, in which books appear as a ‘secondary’ reference, show more diverse styles in *S* than *L*, as follows:

Dataset *S*

- 『TITLE』の物語の発端の 2015 年 5 月だ! … (This month, May 2015, is the beginning of the story of ‘TITLE’! …) [TITLE as a reference for the date]
- AUTHOR の『TITLE』にはっきり書いてありますけれど、日本の地理的条件は、… (As AUTHOR’s TITLE clearly noted, Japan’s geographic condition is …) [TITLE as a reference supporting an own argument]

Dataset *L*

AUTHOR『TITLE1』#読了 やっぱりこの人は読み進めさせる力が半端ないなあと改めて感じた作品です。『TITLE2』の衝撃を期待していただけに、ラストが少し物足りなかった感もありますが、… (AUTHOR ‘TITLE1’ #FinishedReading I again realised this author is super capable to appeal to readers. Although the climax is not quite satisfactory comparing with the impact of her previous ‘TITLE2’, …) [referring to TITLE2 as a related work to TITLE1; a common style of *refer* in *L*]

The public nature of L matches to the predominance of *review* tweets where detailed comments and explanations of books are provided. The larger amount of *ad* also stems from the fact that advertising tweets are more likely to consist of hashtags than ordinary human-generated tweets in order to earn a wide range of consumer[7].

Note that both of our datasets contain around 14% of *recom* and *ad* tweets altogether. both of the purposes correspond to the act of 'Recommendation' in Tweet Acts (Vosoughi and D. Roy, 2016). As mentioned in Section 6.2.3, Vosoughi and D. Roy (2016) reported that tweets that mention 'long-standing topics' have relatively more 'Recommendation' messages. This suggests that reading books as a topic belongs to 'long-standing topics' rather than tweets that mention 'some entities' in general.

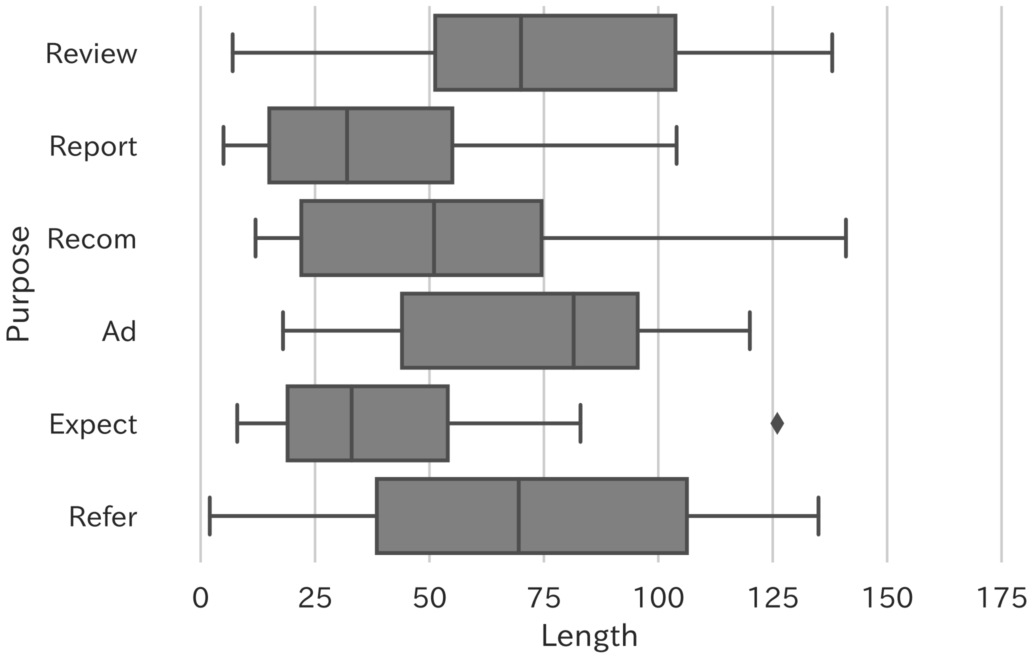
Textual analysis

Next, we describe the characteristics of the textual content of the tweets, per purpose category.

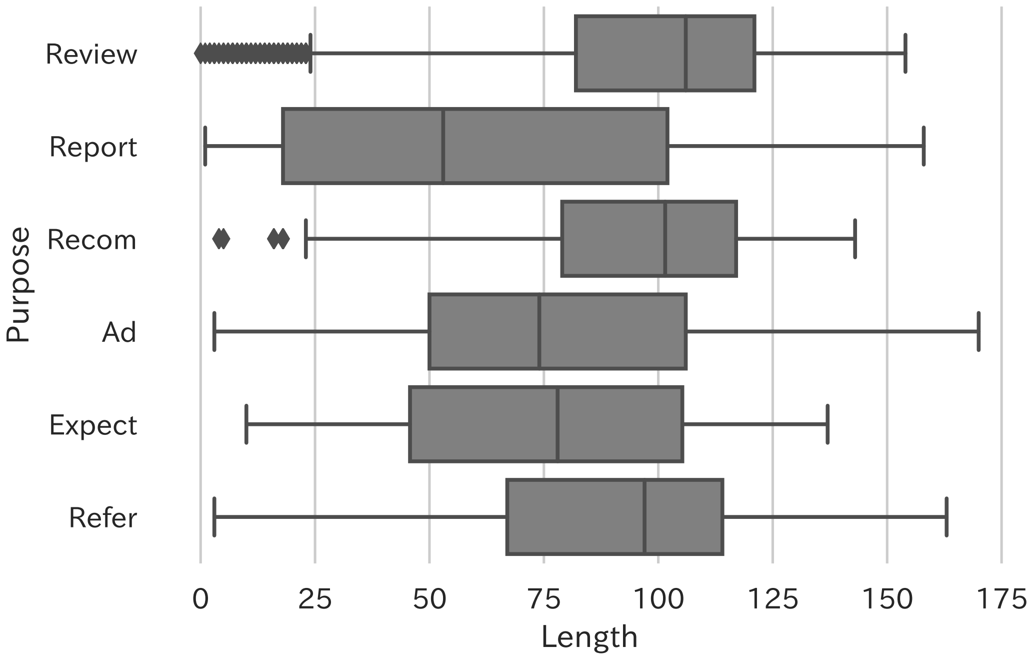
Tweet Length The fact that TMBs are longer than general tweets (Section 6.4.1) might stem from the large volume of *review* tweets, as these could be longer to describe books. To examine this, we drew box plots of the length of TMBs grouped by purposes, as shown in Figure 6.1. These figures show that frequent purposes such as *review* and *refer* are longer than the others. Besides, in all purposes, tweets in L are much longer than those in S. *Report* and *expect* tweets, both of which are short in S, have notably more characters in L. This is largely because *report* and *expect* tweets in L tend to contain extra comments or reviews along with the phrases to report reading progress (*report*) or express expectation for related works (*expect*). That is, *report* and *expect* in L are similar to *review*.

Keywords We further investigated the textual content by finding characteristic words for each purpose. TMBs of different purposes should contain different words that characterise themselves. However, simple (content) word counting independently for each purpose may hinder such characteristics due to the high frequency of common book/reading-related words like shown in Table 6.2. In order to extract keywords, thus, we adopted the term frequency—inverse document frequency (TFIDF) weighting. This can emphasise the words that appear intensively in specific documents. We concatenated all tweets in a purpose group so that each group correspond to a document in TFIDF calculation, which can highlight purpose-wise keywords. Then, TFIDF values were normalised per document (purpose group). Furthermore, we used the words appearing among 3–5 purpose groups

[7] https://sproutsocial.com/insights/hashtag-marketing-tactics/



(a) Dataset *S*



(b) Dataset *L*

Figure 6.1: Boxplots of the length of tweets grouped by the labels of purposes.

only, to cut off too common and specific words as noise. We report the result of this analysis only for *L* because this weighting produced a noisy result for *S* probably due to the data size.

Table 6.4 is the result of TFIDF calculation described above, in which top 25 weighted words are listed for each purpose. We can see that very frequent words like ‘読む (read)’ and ‘本 (book)’ were successfully rejected by this operation. ‘読了 (finished reading)’ is shared across *review*, *report*, *expect*, and *refer*. While this word is obviously the most characteristic to *report*, TMBs of the other four purposes often report the progress of reading as well.

What words characterise purposes? Here we focus on words that occur exclusively in each purpose. *Review* tweets consist of the words related to emotion, description, and evaluation (e.g. ranks **2**, **5**, **11**, **13**, **18**, **20**, and **25**). *Report* has time-related words (e.g. ranks **2**, **3**, **5**, and **11**) and behavioural words (e.g. ranks 1, 10, and 22). In *recom* tweets, other than explicit recommending words (ranks *1*, *2*, and *25*), we can see value-oriented words (ranks 10, 17, and 19). TMBs of *ad* purpose contain metadata about books, such as book series names (rank **6** and **10**) and publisher names (rank **25**). Supplemental words to specify bibliographic information are also found (ranks 4, 20, and 21). In *expect*, functional or modification words are notably ranked in high positions at, e.g. **7**, **9**, **10**, **11**, **12**, and **14**. We can notice prospecting words like ranks 1, 3, 15, and 16 as well. Finally, *refer* TMBs share 12 words with *review* TMBs among top 25 words. This might be because substantial *review* tweets mention other works as references, and corresponding TMB records were labelled *refer* like the example in Section 6.4.2. Appearance of the words referring to relevant works and authors (ranks **12** and **24**) support this phenomenon.

6.4.3 Pleasantness: The opinion polarity

Label counts

In Table 6.5, label counts of the opinion polarity are shown for both datasets. While we consider the composition of positive and negative opinions for each tweet (record), tweets that contain both parts of expressions (*positive* >, =, < *negative*) were very few: 3.78% in *S* and 4.60% in *L*. This is due to the short character limit and the instant nature of posting in Twitter. Besides, since Twitter is not specialised to book reviews, most users may not make *critical* reviews. We can notice that only a few records hold the purely *negative* opinion (2.66% in *S*; 1.38% in *L*).

We reasonably conclude that most TMBs are purely *positive* or no opinions (*neutral*). This partly agree with Chevalier and Mayzlin (2006) which reported that negative book

Table 6.4: Top 25 TFIDF-weighted words for each purpose in dataset *L*. PER/LOC/ORG/PRD denote named entities of person, location, organisation, and product respectively. SEP means sentence ending particles used for sentence-level emphasis in spoken Japanese. MIS denotes mis-tokenised words. Ranks with boldface or underline corresponds to the mentions in body text.

	Review	TFIDF
1	読了 (finished reading)	0.4375
2	感じ (something like)	0.1472
3	人生 (human life)	0.1431
4	笑 (LOL)	0.1175
5	分かる (understand)	0.1120
6	登場人物 (character)	0.1085
7	いう (say)	0.1078
8	文章 (text)	0.0981
9	そんな (such)	0.0975
10	過去 (past)	0.0944
11	美しい (beautiful)	0.0871
12	ブログ (blog)	0.0850
13	強い (strong)	0.0843
14	家族 (family)	0.0781
15	出来る (able to do)	0.0781
16	やる (do)	0.0781
17	男 (man)	0.0760
18	切ない (painful)	0.0733
19	少女 (girl)	0.0728
20	見える (can see)	0.0726
21	特に (especially)	0.0720
22	時代 (era)	0.0719
23	短編 (short story)	0.0712
24	舞台 (stage)	0.0698
25	凄い (awesome)	0.0698

	Report	TFIDF
1	読了 (finished reading)	0.5540
2	昨日 (yesterday)	0.1682
3	7 月 (July)	0.1566
4	少女 (girl)	0.1444
5	6 月 (June)	0.1243
6	短編 (short story)	0.1131
7	男 (man)	0.1015
8	10	0.0974
9	やる (do)	0.0928
10	終える (finish)	0.0899
11	2018	0.0899
12	笑 (LOL)	0.0870
13	治 (PER)	0.0784
14	無い (absent)	0.0772
15	訳 (reason)	0.0754
16	人生 (human life)	0.0754
17	記録 (record)	0.0725
18	家族 (family)	0.0725
19	二人 (couple)	0.0725
20	呼ぶ (call)	0.0672
21	過去 (past)	0.0638
22	見つける (find)	0.0638
23	見える (can see)	0.0638
24	猫 (cat)	0.0638
25	齋藤 (PER)	0.0627

	Recom	TFIDF
1	おススメ (recommend)	0.3099
2	是非 (really)	0.2748
3	読了 (finished reading)	0.2314
4	選 (selection)	0.1256
5	人生 (human life)	0.1230
6	皆さま (everyone)	0.1173
7	そんな (such)	0.1157
8	美しい (beautiful)	0.1085
9	感じ (something like)	0.1085
10	役立つ (useful)	0.1085
11	大人 (adult)	0.1013
12	ビジネス (business)	0.1013
13	つく (have)	0.0940
14	霊魂 (soul)	0.0880
15	笑 (LOL)	0.0868
16	無い (absent)	0.0838
17	良書 (good book)	0.0754
18	考え方 (idea)	0.0754
19	好きな人 (those who love)	0.0754
20	0	0.0754
21	頃 (about)	0.0723
22	短編 (short story)	0.0723
23	味わう (taste)	0.0723
24	個人的 (personal)	0.0723
25	お勧め (recommend)	0.0723

	Ad	TFIDF
1	霊魂 (soul)	0.2222
2	シミルボン (ORG)	0.2055
3	九州 (LOC)	0.1851
4	歌集 (poem collection)	0.1605
5	新 (new)	0.1375
6	新潮文庫 (PRD)	0.1233
7	ほか (etc.)	0.1187
8	死後 (after death)	0.1173
9	販売 (sale)	0.1111
10	角川文庫 (PRD)	0.1111
11	わく (rise)	0.1111
12	絶賛 (praise)	0.1111
13	レビュー (review)	0.1111
14	ランキング (ranking)	0.1096
15	社長 (president)	0.1058
16	記念 (memorial)	0.1049
17	幽い (PRD)	0.1049
18	チェック (check)	0.1005
19	7	0.1005
20	編著 (edited)	0.0987
21	号 (issue)	0.0987
22	社会 (society)	0.0959
23	リーダー (leader)	0.0926
24	日本 (Japan)	0.0913
25	新潮社 (ORG)	0.0868

	Expect	TFIDF
1	次 (next)	0.3803
2	読了 (finised reading)	0.3541
3	楽しみ (looking forward)	0.3148
4	笑 (LOL)	0.2098
5	図鑑 (illustrated book)	0.1671
6	サイズ (size)	0.1241
7	ぞ (SEP)	0.1215
8	い本 (MIS)	0.1049
9	やる (do)	0.0918
10	とりあえず (anyway)	0.0918
11	さて (now/well,)	0.0918
12	ゆっくり (slowly)	0.0886
13	知念実希人 (PER)	0.0787
14	出来る (able to do)	0.0787
15	わくわく (exciting)	0.0759
16	次作 (sequel)	0.0709
17	kindle 版 (kindle ver.)	0.0709
18	終える (finish)	0.0656
19	楽しむ (enjoy)	0.0656
20	忘れる (forget)	0.0656
21	実感 (realisation)	0.0656
22	ラスト (last scene)	0.0656
23	題名 (title)	0.0607
24	病棟 (ward)	0.0607
25	リアル (reality)	0.0607

	Refer	TFIDF
1	読了 (finished reading)	0.2939
2	登場人物 (character)	0.1286
3	木 (tree)	0.1225
4	笑 (LOL)	0.1163
5	男 (man)	0.1163
6	感じ (something like)	0.1041
7	美しい (beautiful)	0.0980
8	開催 (held)	0.0922
9	犯人 (criminal)	0.0922
10	短編 (short story)	0.0918
11	文章 (text)	0.0857
12	前作 (preceding work)	0.0857
13	レポート (report)	0.0827
14	語る (talk)	0.0796
15	時代 (era)	0.0796
16	分かる (understand)	0.0735
17	シミルボン (ORG)	0.0735
18	やる (do)	0.0735
19	かなり (quite)	0.0735
20	いう (say)	0.0735
21	読書会 (reading circle)	0.0709
22	どんでん返し (unexpected ending)	0.0709
23	印象 (impression)	0.0674
24	作者 (author)	0.0674
25	7	0.0674

Table 6.5: The number of TMBs with the proportion for each class in the opinion polarity towards mentioned books.

(a) Dataset *S*

Opinion	Count	Percentage
Positive	195	52.56
Neutral	150	40.43
Negative	12	2.66
Positive > Negative	9	2.43
Positive < Negative	3	0.81
Positive = Negative	2	0.54

(b) Dataset *L*

Opinion	Count	Percentage
Positive	5487	60.12
Neutral	3094	33.90
Positive > Negative	258	2.83
Negative	126	1.38
Positive = Negative	85	0.93
Positive < Negative	77	0.84

reviews in online websites were around 7–15% only, and that positive reviews occupied 70–80%, among book-review websites. Our TMB datasets have the different distribution of *neutral* opinions: while Chevalier and Mayzlin (2006) found 6–10% of neutral reviews, *neutral* records are nearly the same or a half amount of *positive* records in *S* or *L* respectively. Twitter is filled with simple reports (see Section 6.2.3), the majority of which is expected to be neutral.

We further investigated the relationship between opinions and purposes. Table 6.6 provides crosstabs of label counts between the opinion polarity and the tweet purpose annotations. *Review* contains most of opinion-bearing records. That is, large numbers of records in each opinion class except *neutral* are also labelled *review*. Positive tweets (*positive* and *P* > *N*) constitute 70–80% of *review* tweets among *S* and *L*, which agrees with Chevalier and Mayzlin (2006). *Neutral* tweets, in contrast, are found mainly in *report* and *refer* purposes. *Neutral* and *review* TMBs, which constitute 10% of *L*, often describe only the content summaries of mentioned books without any opinion-bearing expressions.

Textual analysis

Tweet Length In Figure 6.2, box plots of the number of characters in TMB text are grouped by the opinion polarity. A common characteristic among both datasets is that all opinion-bearing categories are longer than *neutral* TMBs. Moreover, tweets containing both positive and negative opinions are longer than positive or negative only records. This is reasonable because expressing both opinions may need more characters. Also remember that the *review* purpose held most of opinion-bearing TMBs, and that the *report* purpose mainly consisted of *neutral* TMBs; their lengths are correlated.

Keywords Table 6.7 lists top 25 TFIDF-weighted words of each class in the opinion polarity. The setup and procedure for TFIDF calculation are the same as in Section 6.4.2.

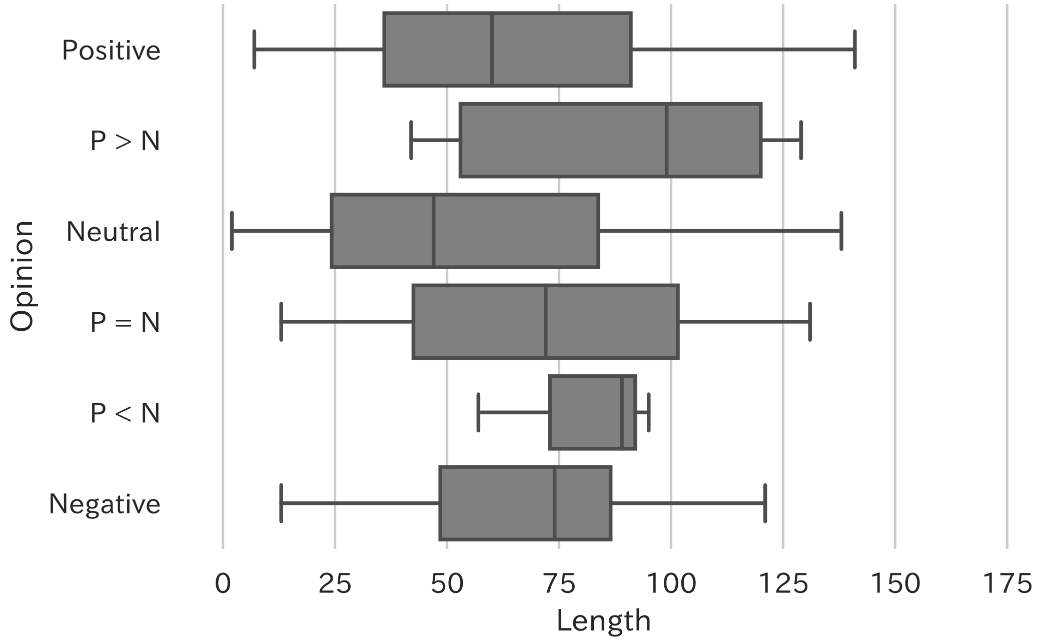
Table 6.6: Results of the opinion polarity and purpose annotation. Values in parentheses are the proportion (%).

(a) Dataset *S*

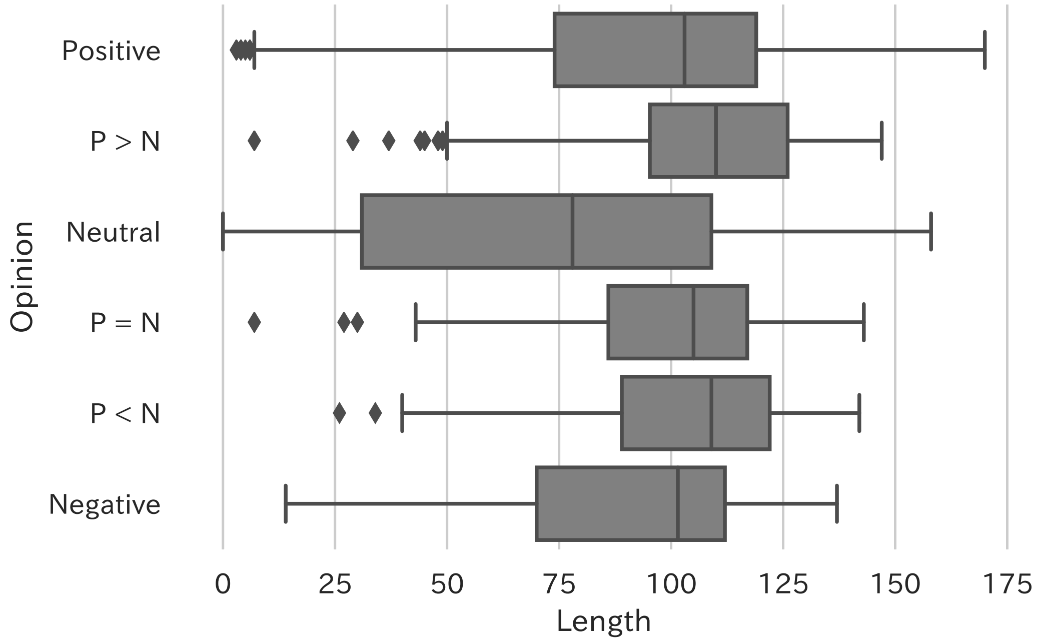
Purpose	Opinion						All
	Positive	**P > N**	**Neutral**	**P = N**	**P < N**	**Negative**	
Review	85 (22.91)	8 (2.16)	10 (2.70)	1 (0.27)	2 (0.54)	8 (2.16)	114 (30.73)
Report	15 (4.04)	0 (0.00)	61 (16.44)	1 (0.27)	0 (0.00)	1 (0.27)	78 (21.02)
Recom	27 (7.28)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	27 (7.28)
Ad	23 (6.20)	0 (0.00)	1 (0.27)	0 (0.00)	0 (0.00)	0 (0.00)	24 (6.47)
Expect	27 (7.28)	0 (0.00)	2 (0.54)	0 (0.00)	1 (0.27)	0 (0.00)	30 (8.09)
Refer	18 (4.85)	1 (0.27)	76 (20.49)	0 (0.00)	0 (0.00)	3 (0.81)	98 (26.42)
All	195 (52.56)	9 (2.43)	150 (40.43)	2 (0.54)	3 (0.81)	12 (3.23)	371 (100.00)

(b) Dataset *L*

Purpose	Opinion						All
	Positive	**P > N**	**Neutral**	**P = N**	**P < N**	**Negative**	
Review	3680 (40.32)	247 (2.71)	1002 (10.98)	73 (0.80)	75 (0.82)	114 (1.25)	5191 (56.88)
Report	220 (2.41)	3 (0.03)	1474 (16.15)	4 (0.04)	2 (0.02)	4 (0.04)	1707 (18.70)
Recom	401 (4.39)	6 (0.07)	17 (0.19)	2 (0.02)	0 (0.00)	0 (0.00)	426 (4.67)
Ad	883 (9.68)	0 (0.00)	82 (0.90)	0 (0.00)	0 (0.00)	0 (0.00)	965 (10.57)
Expect	160 (1.75)	1 (0.01)	69 (0.76)	2 (0.02)	0 (0.00)	0 (0.00)	232 (2.54)
Refer	143 (1.57)	1 (0.01)	449 (4.92)	4 (0.04)	0 (0.00)	8 (0.09)	605 (6.63)
All	5487 (60.12)	258 (2.83)	3093 (33.89)	85 (0.93)	77 (0.84)	126 (1.38)	9126 (100.00)



(a) Dataset *S*



(b) Dataset *L*

Figure 6.2: Boxplots of the length of tweets grouped by the labels of the opinion polarity.

Table 6.7: Top 25 TFIDF-weighted words for each opinion polarity in dataset *L*.

	Positive	TFIDF
1	冊 (*n* piece of books)	0.2845
2	書評 (book review)	0.1747
3	目 (*n*-th)	0.1684
4	見る (see)	0.1381
5	せる (make someone do)	0.1343
6	生きる (live)	0.1293
7	考える (think)	0.1238
8	著者 (author)	0.1188
9	による (by)	0.1179
10	ー (MIS)	0.1133
11	時 (when)	0.1122
12	くれる (gratefully do)	0.1122
13	など (etc.)	0.1029
14	そして (and)	0.0990
15	紹介 (introduce)	0.0917
16	すごい (awesome)	0.0908
17	reviews	0.0907
18	お (prefix)	0.0875
19	2	0.0858
20	all	0.0784
21	怖い (scary)	0.0781
22	美しい (beautiful)	0.0743
23	本当に (really)	0.0743
24	幸せ (happy)	0.0739
25	作家 (author)	0.0732

	Neutral	TFIDF
1	書評 (book review)	0.2274
2	ブログ (blog)	0.1931
3	シミルボン (ORG)	0.1395
4	2018	0.1335
5	更新 (update)	0.1287
6	昨日 (yesterday)	0.1210
7	6	0.1127
8	彼女 (she)	0.1091
9	大 (big)	0.1091
10	彼 (he)	0.1019
11	家族 (family)	0.0983
12	愛 (love)	0.0965
13	変わる (change)	0.0965
14	幸せ (happiness)	0.0948
15	4	0.0948
16	会 (community)	0.0876
17	恋 (love)	0.0876
18	あなた (you)	0.0858
19	国 (nation)	0.0793
20	猫 (cat)	0.0751
21	場所 (place)	0.0733
22	どこ (where)	0.0715
23	それでも (but)	0.0709
24	大人 (adult)	0.0697
25	円 (yen)	0.0688

	Negative	TFIDF
1	目 (-th)	0.1453
2	しれる (may)	0.1272
3	映画 (film)	0.1262
4	見る (see)	0.1090
5	所 (place)	0.1090
6	読者 (readers)	0.0908
7	オチ (end)	0.0908
8	2	0.0908
9	真相 (truth)	0.0842
10	カバー (cover)	0.0842
11	どんでん返し (unexpected ending)	0.0842
12	づらい (difficult to do)	0.0842
13	迫る (approach)	0.0737
14	授業 (class)	0.0737
15	巡る (go around)	0.0737
16	太陽 (the sun)	0.0737
17	司馬遼太郎 (PER)	0.0737
18	ダメ (no good)	0.0737
19	ほう (MIS)	0.0737
20	きっと (hopefully)	0.0737
21	p	0.0737
22	騙す (deceive)	0.0727
23	著者 (author)	0.0727
24	考える (think)	0.0727
25	死 (death)	0.0727

	P > N	TFIDF
1	冊 (*counter suffix of books*)	0.1628
2	目 (-th)	0.1542
3	変わる (change)	0.1290
4	楽しめる (enjoyable)	0.1191
5	途中 (in the half way)	0.1114
6	見る (see)	0.1114
7	せる (make someone do)	0.1028
8	者 (person)	0.0893
9	設定 (setting)	0.0857
10	著者 (author)	0.0857
11	物足りない (unsatisfactory)	0.0857
12	ものの (but)	0.0857
13	それでも (but)	0.0810
14	一気 (at a stretch)	0.0794
15	づらい (difficult to do)	0.0794
16	章 (chapter)	0.0771
17	登場 (appearance)	0.0771
18	引き込む (attract)	0.0771
19	再読 (read again)	0.0771
20	なかなか (considerably)	0.0771
21	そして (and)	0.0771
22	すごい (awesome)	0.0771
23	題名 (title)	0.0695
24	興味深い (interesting)	0.0695
25	師 (master)	0.0695

	P = N	TFIDF
1	家族 (family)	0.1291
2	設定 (setting)	0.1115
3	考える (think)	0.1115
4	怖い (scary)	0.1115
5	おもしろい (interesting)	0.1115
6	真剣 (serius)	0.0904
7	物凄い (quite)	0.0904
8	師匠 (master)	0.0904
9	疲れる (get tired)	0.0892
10	時 (when)	0.0892
11	所 (place)	0.0892
12	他 (other)	0.0892
13	ファンタジー (fantasy)	0.0892
14	わ (MIS)	0.0892
15	なんか (kinda)	0.0892
16	辛い (painful)	0.0775
17	版 (version)	0.0775
18	決して (never)	0.0775
19	手紙 (letter)	0.0775
20	彼女 (she)	0.0775
21	弱い (weak)	0.0775
22	一つ (one)	0.0775
23	モヤモヤ (unsatisfactory)	0.0775
24	による (by)	0.0775
25	雰囲気 (atmosphere)	0.0669

	P < N	TFIDF
1	見る (see)	0.1448
2	うーん (hmm)	0.1305
3	合う (match)	0.1207
4	など (etc.)	0.1207
5	恩田陸 (PER)	0.1118
6	店 (store)	0.0979
7	ドラえもん (Doraemon)	0.0979
8	すっきり (clear)	0.0979
9	見える (seeable)	0.0966
10	特に (especially)	0.0966
11	全然 (not at all)	0.0966
12	人物 (person)	0.0966
13	一 (one)	0.0966
14	よく (well)	0.0966
15	若い (young)	0.0839
16	少年 (boy)	0.0839
17	出来事 (event)	0.0839
18	とき (when)	0.0839
19	っていう (…that)	0.0839
20	しんどい (painful)	0.0839
21	いまいち (not very)	0.0839
22	雰囲気 (atmosphere)	0.0724
23	要素 (element)	0.0724
24	聞く (listen)	0.0724
25	納得 (become convinced)	0.0724

Table 6.8: Annotation results of the strength of recommending phrases and the scale of audience. The number inside parentheses shows the count of the recommendation audience written explicitly.

(a) Dataset *S*

Scale of audience	Strength of recommending phrases			All
	Weak	Strong	Excess	
Individuals	10 (10)	2 (2)	0 (0)	12 (12)
Specific group	2 (2)	4 (3)	0 (0)	6 (5)
Everybody	7 (0)	6 (1)	0 (0)	13 (1)
All	19 (12)	12 (6)	0 (0)	31 (18)

(b) Dataset *L*

Scale of audience	Strength of recommending phrases			All
	Weak	Strong	Excess	
Individuals	3 (3)	0 (0)	0 (0)	3 (3)
Specific group	167 (165)	22 (22)	6 (6)	195 (193)
Everybody	221 (35)	37 (2)	6 (0)	264 (37)
All	391 (203)	59 (24)	12 (6)	462 (233)

In comparison to the case of purposes, these lists show a noisier result. We can still see, however, a couple of polarity-showing words for *positive* (e.g. ranks **16**, **21**, **22**, **24**) and *negative* (e.g. ranks **12**, **18**, **22**). TMBs having both opinions rather show more polarity-showing words of both positive and negative. They are ranked at **4**, **18**, and **22** in $P > N$, at **4**, **5**, **7**, **9**, **16**, **21**, **23** in $P = N$, and at **2**, **8**, **14**, **20**, **21** in $P < N$. These opinion groups are more characterised by contrastive words (ranks 3, 12, 13, and **21** in $P > N$) and negation words (ranks **15** and **20** in $P > N$; ranks 18 in $P = N$; ranks 11 in $P < N$).

6.4.4 Textual considerateness: Recommending TMBs

Label counts

We report the count of labels for recommending TMBs. First of all, records that contain recommending phrases, or recommending TMBs, constituted a small portion of both datasets: 31 (8.4%) in *S* and 462 (5.1%) in *L*. Among them, records that specify the recommendation audience were 18 in *S* and 233 in *L*, which resulted in almost the half of recommending TMB records in both datasets.

In Table 6.8, crosstabs are shown between the strength of recommending phrases and the scale of the recommendation target audience. As the strength increases, in all audience scales, the number of records drops; the *excess* case was absent in *S*. Most recommending TMBs use considerate (*weak*) phrases, and thus excessive (*excess*) recommendations appear rarely. For the audience scale, *S* had more records that targeted *individuals* than *L* did. This would stem from the fact that *S* contains more replying tweets (14.8% in *S*; 2.85% in *L*). Few users may put hashtags when replying to others due to the public nature of hashtags (Bruns and Burgess, 2011; Scott, 2015; van den Berg, 2014), which could cause few replying

Table 6.9: Annotation results of recommending TMBs and purposes.

(a) Dataset *S*

Purpose	Target scale and strength						All
	Individuals		Certain group		Everybody		
	Weak	Strong	Weak	Strong	Weak	Strong	
Recom	10	2	2	3	4	4	**25**
Ad	0	0	0	0	3	2	**5**
Refer	0	0	0	1	0	0	**1**
All	**10**	**2**	**2**	**4**	**7**	**6**	**31**

(b) Dataset *L*

Purpose	Target scale and strength							All
	Individuals	Certain group			Everybody			
	Weak	Weak	Strong	Excess	Weak	Strong	Excess	
Review	0	8	1	0	10	3	0	**22**
Report	0	0	0	0	3	0	0	**3**
Recom	3	156	21	6	200	33	6	**425**
Ad	0	3	0	0	5	1	0	**9**
Expect	0	0	0	0	1	0	0	**1**
Refer	0	0	0	0	2	0	0	**2**
All	**3**	**167**	**21**	**6**	**221**	**37**	**6**	**462**

tweets in *L*. Recommendations that targets *specific group* and *everybody* were roughly balanced in both datasets, but the cases of *everybody* were larger in *L*. Since hashtags-bearing tweets are automatically tied to public streams of topics, users may also consider potential audience via hashtag search.

Table 6.9 shows what purposes are found among recommending TMBs. *Recom* was expectedly assigned to almost all of recommending TMBs. A few of them were, however, annotated as *review* tweets that mainly state a review of the book with a supplemental phrase of recommendation. This may happen because our annotation scheme only assignes a single, best applicable purpose to each TMB record, and because many people make tweets that review media contents not only for sharing but also for recommending the contents (Kitamura, Sasaki and Kawai, 2016). A couple of records were labelled other than *recom* or *review* because these records consist of reporting or referring phrases and slight recommendation phrases. Surprisingly, few *ad* tweets appeared in recommending TMBs. This may be because *ad* tweets just state the start of selling the books or the sales status of the book rather than explicitly saying recommending phrases, following Packard and Berger (2016).

Table 6.10: Annotation results of recommending TMBs and the opinion polarity

(a) Dataset *S*

Purpose	Target scale and strength						All
	Individuals		Specific group		Everybody		
	Weak	Strong	Weak	Strong	Weak	Strong	
Positive	10	2	2	4	7	6	31

(b) Dataset *L*

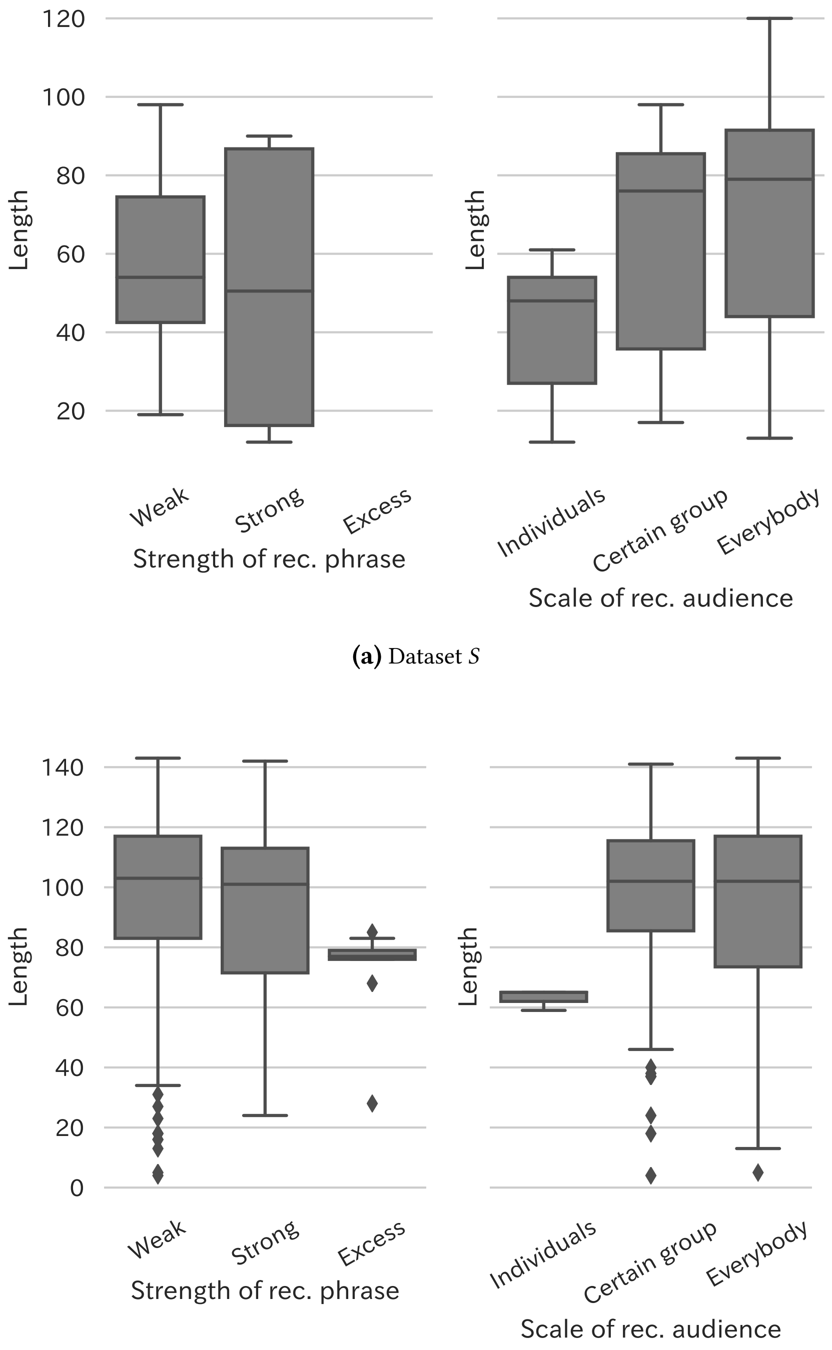
Purpose	Target scale and strength							All
	Individuals	Specific group			Everybody			
	Weak	Weak	Strong	Excess	Weak	Strong	Excess	
Neutral	0	10	2	0	10	2	0	**24**
Positive	3	152	20	6	208	35	6	**430**
P = N	0	1	0	0	1	0	0	**2**
P > N	0	4	0	0	2	0	0	**6**
All	**3**	**167**	**22**	**6**	**221**	**37**	**6**	**462**

We also checked the opinion polarity in recommending TMBs as shown in Table 6.10. Almost all of recommending TMBs had positive opinions and no purely *negative* records were found in both datasets. A few neutral records appeared in *L* (24 in *neutral* and 2 in *P* = *N*). The authors of *neutral* tweets recommend a book to some audiences without mentioning their own impressions. The balanced instances (*P* = *N*) introduce both positive and negative sides of the mentioned book, and recommend it under the condition 'if you may find it interesting'.

Textual analysis

Tweet Length We first measured the characteristic lengths to recommending TMBs: recommending phrases and audience specifications. Our two datasets have the same median value of 7 for recommending phrases. The median lengths of the recommendation audience were 9 in *L* and 8 in *S*.

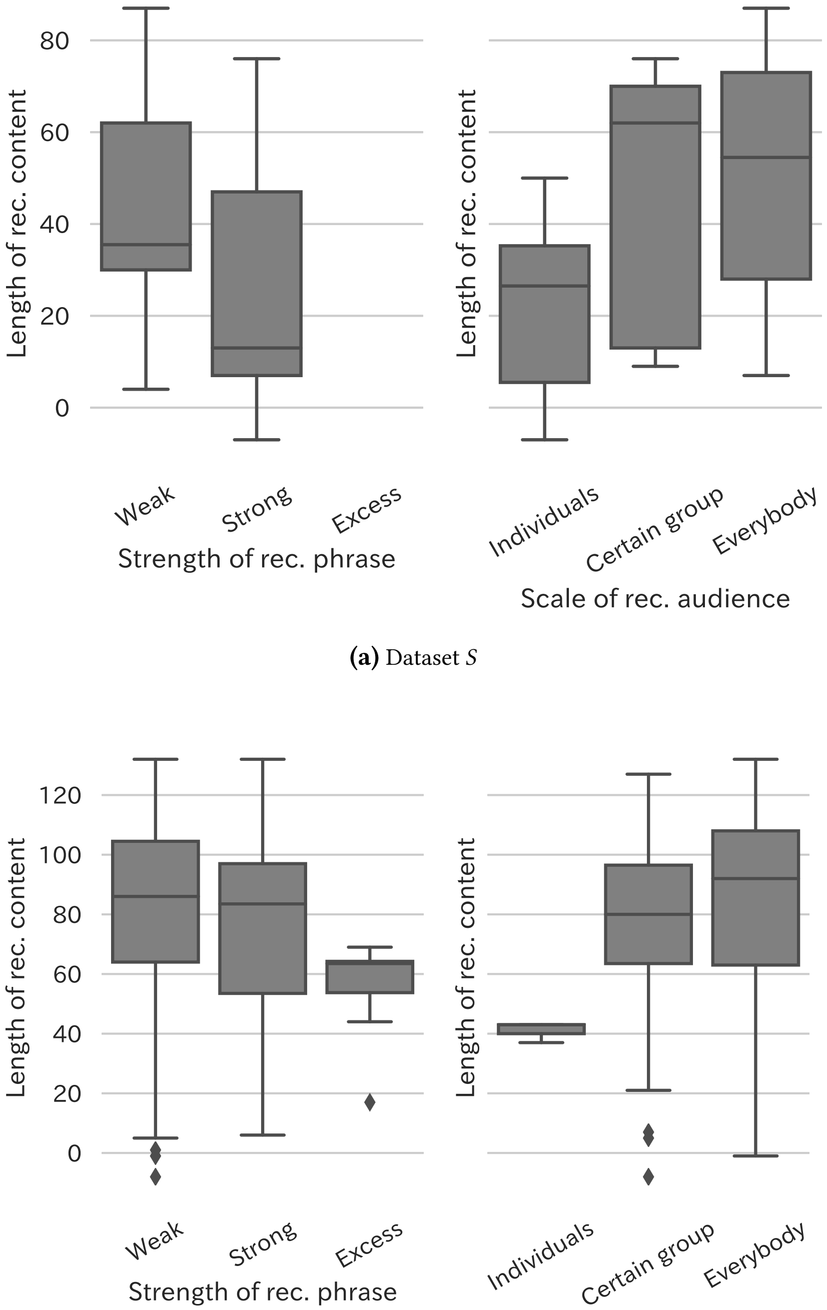
Next, Figure 6.3 shows the length of recommending TMBs grouped separately by the strength of recommending phrases or the scale of the recommendation audience. We noticed that, focusing on the median, the length is shorter when the strength of recommendation is greater. For the recommendation audience, the opposite relation is shown. It seems that (i) considerate recommendations use more characters to hedge their wording whereas stronger recommendations say simpler words with fewer reasons or explanations, and that (ii) more explanations are needed for larger audiences.



(a) Dataset *S*

(b) Dataset *L*

Figure 6.3: Box plots of the length of recommending TMBs grouped by the strength of recommendation phrases and the recommendation audience.



(a) Dataset *S*

(b) Dataset *L*

Figure 6.4: Box plots of the length of recommendation content grouped by the strength of recommendation phrases and the recommendation audience.

In order to examine these points, we further calculated the length of the *content* of recommending TMB records. We estimated the length of recommendation content by subtracting the length of recommendation phrases and audiences from the whole length. For example, given a tweet "*boys*, why not reading *TITLE*?? It's pretty helpful!!" [originally in Japanese], the content length is the number of characters not in italics. Figure 6.4 are box plots of the length of recommendation content grouped by the strength of recommendation phrases and the audience scale. Some outliers are plotted at lower than 0 because a couple of records embed recommendation phrases or targets within hashtags like '#ビジネス小説好きに**オススメ** (#**Recommended**ToBussinessNovelLovers)'[8], while we removed hashtags from the calculation of recommendation-content lengths. The points (i) and (ii) above are supported; these content lengths are in a reverse proportion to the strength and in a direct proportion to the audience scale.

Keywords Although we applied the same procedure of TFIDF weighting for recommending TMBs, the result did not show clear patterns. Instead of quantitative methods, we qualitatively investigate what phrases and targets were mentioned in recommending TMBs in the next Section 6.5.

6.5 Qualitative analysis of textual considerateness

We categorised expressions appearing in recommendation phrases and recommendation audiences with regard to the textual contents. In this analysis, we investigate *L* only, because the size of *S* is small, and almost all of them are covered by the examples in *L*.

6.5.1 Recommendation phrases

We grouped the categories of recommendation phrases for each strength.

Weak-level strength

For 394 *weak* phrases, 84.26% are covered by the following three categories, examples of which are also included in the annotation guideline (Section 5.3.2):

Recommendation (210, 53.71%): 'お薦め (recommend)', '推薦図書です/オススメ本 (a recommended book)', "推しである (that's my recommendation)", …

Hope (69, 17.65%): '読んでもらいたいです (I hope you read this book)', '読んでほしい一作です (This is a piece of work that I hope you read)', …

[8] The hyphen in this hashtag is put just for typography. Twitter hashtags do not allow hyphens.

Suggestion (67, 17.01%): ‘ぜひご一読を！ (Please definitely have a read-through of this book!)’, ‘読んでみてください！ (Please try!)’, ‘ぜひ (certainly)’, …

The majority is simple and straight forward phrases of just ‘recommendation’, while weaker expressions like ‘hope’ and ‘suggestion’ appear around 17% for each. The rest 15.74% of expressions can be categorised as follows:

Warranty (25, 6.39%): “読みやすいかも！ (it would be easy to read!)”, “楽しめる小説だと思う。(I think it’s an enjoyable novel.)”, …

Request (14, 3.58%): ‘ぜひ、お読みください (Please certainly read)’, ‘ぜひ知ってください (Please really know)’, …

Target audience only (6, 1.53%): ‘働く人みんなへ (for all of workers)’, ‘TOPIC について知りたい方は以下の三部作を (to those who want to know about TOPIC, the following trilogy)’, …

The instances of ‘warranty’ are somewhat implicit since they just state how the tweet author think others can enjoy the book, but they still imply the will of recommendation. In ‘request’, the polite language ‘ください’ delivers softer impression even though the syntax of these phrases take imperative forms.[9] A few examples use the ‘target audience only’.

Among these categories, we further found a few tweets (= 10) that use specialised phrases refer to the content of the mentioned books:

- (suggest) – ‘この歪んだ教室を覗いてみませんか？ (why not witness this warped classroom?)’
- (hope) – “このトリックを知ってもらいたい！ (I hope you know this trick!)”

They sound like advertising slogans, although they were apparently generated by consumers.

Strong-level strength

Strong expressions consist of the following three categories:

Order (25, 42.37%): ‘これは読んで！！ (do read this!!)’, ‘必読 (must read)’, …

Strengthened (22, 37.29%): “この本を読まないとね！ (you have to read this book, don’t you?)”, ‘本当におすすめです！！ (recommend this book honestly!!)’, ‘読んでほしい！ 絶対！ (wish you to read! Definitely!)’, …

Obligation (12, 20.34%): ‘絶対読むべき (should read absolutely)’, ‘読んだ方がいい。(had better to read [the book].)’, …

[9] This characteristic would also apply to other languages that often use honorifics in conversations, such as Korean and Thai.

All patterns above are covered by or naturally extended from the examples in the annotation guidelines. We grouped emphasised versions of *weak* expressions into 'strengthened'. Even some extreme cases were found in this group, including the phrases like advertising slogan:

- '哲学書デビューはこいつで決まり！ (Best for your philosophy book debut!)'
- 'この時代の必須アイテム (A must-have item of this era)'
- 'どうか夏の暑い暑い日中に、エアコンもかけずに読み切ってほしい一冊。 (wish you to read through this one at a stretch without using your air conditioner in a super hot daytime of Summer.)'
- '是非ともオススメ。一行たりとも見逃さずに読んで頂きたいです。 (Recommend this at any cost. I sincerely want you to read it without missing any single line.)'
- '全力で AUTHOR さんのオススメ小説を紹介します！！ (I will introduce my recommendation of books by AUTHOR to the best of my ability!!)'
- 'いいから読んで！！ とだけ言いたい (All I wanna say is "just read it without argument!!")'

Although the long phrases above might seem forceful in a certain context, they just over-emphasise positive recommendations and have no aggressive meaning like *excess* instances, which are introduced next.

Excess-level strength

Finally, we introduce recommending phrases labelled *excess*. We should note that 11 out of 12 samples mentioned the same book and their accounts were deleted or suspended when we made the analysis, which suspects that they were generated by stealth marketing.[10]

- 'ヤバイよ！ 早く読まなくちゃ (omg! must read it in a hurry)'
- "読まんで…いいと？ (are you thinking …it's acceptable not to read the book?)"
- "知らないのはヤバイよ、 (you'll get in trouble if you don't read the book)"
- '買いに行けるね！ 注文出来ますよ！ (you can go to buy it! you can order it!)'
- '迷ってないで早く読まないとだよ (must read it without hesitation)'
- '早く読まないと!! (must read it as soon as possible!!)'

They can be labelled threat appeal or peer pressure, implying that not reading the book is inappropriate.

Only one example that was annotated as the *excess* scale other than the above was: '鼻先に突きつけたくなった。 (I wanted to thrust this book just in front of their noses.)'. This is regarded as an aggressive expression to make the particular people read the book.

[10] We tried filtering bot tweets by application names used to post tweets. These tweets were posted from Twitter for iPhone, which may mean they were posted by the human labour of innumerable workers.

Summary of recommendation phrases

In recommending TMBs, a diverse variety of phrases were found in the *weak* and *strong*-level strength and able to be categorised by grammatical or semantic perspectives. Whereas most recommendation expressions followed examples that our annotation guideline provided, a few phrases were creatively specialised to the book content, imitating advertising slogans. Expressions of the *excess* strength may need much more examples to describe their characteristics, but we confirmed that they certainly exist in Twitter.

6.5.2 Recommendation audiences

Now we report the categorisation of the recommendation audience specified in recommending TMBs that targeted *specific groups* or *everybody*. *Individuals* are skipped since the target users were exactly denoted by the @-reply format among the examples.

Specific group

The perspectives of declaring *specific group* (195 TMBs) were categorised into the following three types:

Interest and personality (99, 50.77%): '時代物が好きな人 (people who like history books)', 'ロボアニメ好き (robot Anime lovers)', '読書が嫌いな人 (those who do not like books)', '後味悪い話が好きな方 (those who like bad endings)', '上手な文章を書きたい人 (those who want to write good)', '問題解決をしたい人 (those who desire to solve problems)', …

Circumstances and experience (50, 25.64%): '受験生の人とか (somebody like studying for entrance exams)', '夏休みに入る前・入った人 (those who are in or right before the summer vacation)', '心が疲れてる人 (mentally exhausted people)', 'まだの方/未読の方/読んでない方 (those who have not yet read this book)', '読書慣れしてない人 (those who are not accustomed to reading)', …

Demographic profile (37, 18.97%): 'ちびっ子 (kids)', '10〜20 代の人 (10–20s)', '高校生以上 (high school students or older)', '男性 (male)', …

The most frequent pattern was the *interest and personality*. Some tweets target very specific audience as follows:

- (*interest and personality*) – 'ちゃんと弱い人を書いている作家さんが好きな人 (those who prefer to authors that describe weak people precisely)'
- (*circumstances and experience*) – '生きづらい人、思考が堂々巡りして行き詰まっている人など (somebody feeling hard about life, or are at deadlock due to a thinking loop)'

A few examples (7 in total) use a combination of these perspectives, for instance:

- ‘若い世代やあまり本を読まない人 (young generations or infrequent readers)’
- ‘工事現場や重機が好きな男の子 (boys who like construction sites and heavy construction equipment)’

As in these expressions, they implicitly associate some personal traits with certain demographics or situations.

In the *specific group* scale, two samples did not have phrases to specify any audience because they implicitly targeted students to whom the school homework to write book reviews are assigned: e.g. ‘今日は #読書感想文 にオススメの #本 を紹介します (Today, I introduce a #book recommended for #BookReviewSchoolComposition)…’.

Everybody

Among 264 TMBs that recommend books to *everybody*, 37 (14.02%) of them use explicit phrases. The majority of such expressions, which appear in 22 (8.33%) of the 264 TMBs, were simple mentions meaning ‘all’ or ‘everybody’ (e.g. ‘みなさん’, ‘皆’, ‘みんな’, ‘誰にでも’, and ‘皆さま’). Except for these, two types of target-audience specification-styles are found:

A set and its compliment (6, 2.27%): “山に登る方にも登らない方にも (not only those who climb mountains but also those who don’t)”, ‘数学嫌いも数学好きも (maths haters and maths lovers)’, “理系に興味のある人もない人も (not only those who have interest in science but also those who don’t)”, ‘大人も子供も (grown-ups and kids)’, …

A broad range (6, 2.27%): ‘幅広い年代の方 (people of various ages)’, ‘性別・年齢問わず、いろんな人 (various people regardless of sex and ages)’, ‘多くの人/たくさんの人 (many people)’, …

We also found two other mentions: one using the second person ‘あなた (you)’ and; one that gradually expands the range of audiences to everybody, i.e. ‘今、ここで悩んでる人、生き方がわからなくなった人、全ての人 (those who are worrying right now and here, those who have lost how to live along, and all people)’.

Summary of recommendation audiences

The authors of recommending TMBs typically consider the match between book contents and preferences/situations of audiences. Some examples specify target audiences in a highly detailed way.

6.6 Discussion

6.6.1 Summary

We described the characteristics of how current Twitter users are exposed to tweets that mention books (TMBs), focusing on Japanese environment. Our focal points of the analysis were textual aspects of inspiringness, i.e. pleasantness and considerateness. Pleasantness was formulated into opinion polarity, whereas textual considerateness was measured from recommending TMBs. We also investigated the purpose of mentioning books. These attributes were derived from the TMB-corpus creation (Chapter 5), where two TMB corpora collected by different methods (i.e. the TMB identification dataset and the TMB inspiringness dataset) were labelled following the annotation guideline. We analysed the subsets of all TMBs in these datasets (i.e. the dataset S from TMB identification dataset and the dataset L from TMB inspiringness dataset) by using both quantitative and qualitative methods.

While most of the general tweets are known as simple reports of events or feelings, TMBs consisted of much more tweets that review or comment on books. The dataset L, hashtag-based, also contained more tweets that advertise books. TMBs had more characters than general tweets because of the large proportion of TMBs that review, advertise, and refer to books. Using TFIDF-based weighting, we observed characteristic words for each purpose we defined.

The opinion polarity toward books was basically positively framed. While neutral opinions were still substantial, negatively framed opinions were very few. This distribution, especially inside the subset of reviewing TMBs, is similar to that of book-review websites where positive opinions are dominant; TMBs in general can be characterised by the relatively large amount of neutral tweets. TMBs that contain both positive and negative opinions rarely appeared, probably due to the short character limit and the instant nature of posting. TMBs containing opinions were longer than neutral tweets. TFIDF weighted words characterised tweets that have both positive and negative opinions by switching words like ‘but’.

Recommending TMBs constitute a small portion (5–8%) of our datasets. A half of them explicitly specify the audience for the recommendation. The number of recommending TMBs decreased as the strength of recommendation increased. We found that recommending TMBs with stronger phrases have fewer textual contents, which suggested that tweet authors who are eager to recommend books tend to rely mostly on recommending phrases and to provide less reasons and explanations. We made a qualitative analysis to group expressions used in recommending phrases and recommendation audiences. This showed diverse variations of the recommendation style in terms of grammar and semantic.

Table 6.11: Top 10 hashtags with frequencies in dataset *L*

1	**#(book title)**	1651
2	**#読書好きと繋がりたい (#WannaConnectWithBookLovers)**	853
3	**#読書記録 (#ReadingRecord)**	505
4	**#book**	216
5	**#読書好きな人と繋がりたい (#WannaConnectWithReadingLovers)**	195
6	**#小説 (#Novel)**	178
7	**#日経新聞 (#NewsPaperNikkei)**	186
8	**#本好きな人と繋がりたい (#WannaConnectWithBookLikers)**	170
9	**#漫画 (#Manga)**	154
10	**#読書垢 (#ReadingAccount)**	136

TMBs can be evaluated as a good source of passive exposure to books from the following points:

- since reviews on books are the most common way of mentioning books, those who are exposed to TMBs can be inspired by knowing some detail of books;
- the majority of TMBs are positive, which is known to give positive impressions on mentioned entities (i.e. books);
- TMB receivers are less likely to be discouraged to read books by encountering negative opinions since they appear very rarely;
- few occurrences of explicitly recommending TMBs provides smaller possibilities of evoking psychological reactance.

However, we should note that the amount of TMBs are small in the whole tweet population. *S* constitutes only around 3% of a set of tweets that contain book-title strings, while the size of *L* is relatively small in contrast to the period of collection. This suggest that we need to propagate TMBs in order to make them to be an online surrogate for traditional exposure to books, such as bookshops in town.

6.6.2 Limitations

We focused on Japanese TMBs and acknowledge that some styles of mentioning books and label distributions may differ in different languages. Japanese texts can contain more information than alphabet-based texts like English due to its number of characters (Neubig and Duh, 2013), which has allowed Japanese Twitter users to make self-contained lengthy tweets like reviewing books more easily from the start of Twitter's service operation. Since the character limit in Twitter was expanded[4], however, such tweets with more information may emerge also in alphabet-based languages. This implies that our findings on Japanese tweets can be now generalised to other languages, such as label distributions, tendency of tweet lengths, and the category of recommendation phrases and audiences.

The dataset L was collected by hashtag-based search, which may produce a skewed distribution of tweets. In fact, the distributions of L we examined were different from those of S. Basically, tweets with hashtags constitute a very small proportion (e.g. less than 15% in Guzman, Alkadhi and Seyff (2017)'s dataset), probably due to the necessity of an additional action to put a # mark. Table 6.11 lists the top 10 hashtags in L, excluding the book-related keywords to collect L, i.e. #読書 (book reading), #読了 (finished reading), #書評 (book review), and #本 (books). The top hashtag '#(book title)' denotes the hashtags of specific book titles (like '#HuckleberryFinn'). We can see some hashtags aiming to gain followers, e.g. #読書好きと繋がりたい (#WannaConnectWithBookLovers; rank **2**, **5**, and **8** are the same concept in different phrasings). This also expects a reciprocal communication, i.e. mutual following-followed relationships (Anger and Kittl, 2011). The tweet authors in L are more likely to have desire to expand their follower networks. The dataset S, in contrast, contained only 36 hashtags in 371 records. We barely found 2 #読書 (reading) and 1 #書評 (book review) among the book-related hashtags we used for L.

Another issue is the existence of spam accounts. Even though we excluded the tweets posted from non-official Twitter client applications, we noticed that about 5.1% of tweets in L became unavailable because their original accounts had been deleted or suspended by the point of 2 December 2018. We still included them in our analysis, since they were not so many and we could not find hardly any difference in tweet text from the similar posts by accounts alive.

6.6.3 Outlooks

From a methodological point of view, this research can be referred to as a specific case study of tweets which requires careful design of certain linguistic concepts to be annotated. The present study is, as far as we know, the first attempt to examine the strength or forcefulness over the phrases of recommending certain entities in casual online mentions. Our framework can also be applied to other entities such as films, videos, medicines, and tourism.

While we believe that the most of our findings and approaches are still applicable to other languages, it is worth examining the difference in TMBs between languages. For example, conventions in making book titles and circumstances of online book services can affect the style of TMBs. Similarly, chronological analysis of TMBs would provide interesting insights in relation to the periodical change of Twitter usage. As we analysed TMBs regardless of the difference in user accounts, a user-wise analysis may show other insights as individuals have different language style. We found expressions of recommendation targets were linguistically diverse. Besides, identifying the target audience of recommendation could be another interesting application because the target audience may have effects on the feeling of message receivers (Imajo, 2012).

While this research showed that TMBs can be regarded as a good source for online exposure to books, it remains unclear how often TMBs are received by Twitter users. This type of research might be challenging because Twitter reorganises tweets of users' home timelines based on users' activities. Just counting followers of TMB authors may not correspond to the actual experience of encountering TMBs. For a better estimation, we should also take into account TMB receivers' behaviours, such as when and how often they use Twitter. Some filter-bubble research, where ideologically biased information consumption is investigated, uses user profile-based estimation on a large scale data (Eady et al., 2019) or following/followed network analysis (Himelboim, M. Smith and Shneiderman, 2013). We plan to design a suitable method to examine actual exposure to TMBs.

Finally, we should note that the insights we obtained in this chapter will be utilised for designing the tasks and methods of iTMB scoring in Chapter 8.

Part IV

Core NLP modules of the system

7 TMB identification

Note that this chapter is the reorganisation of the following paper:

> Yada, S., Kageura, K., and Paris, C., 2019 (online first). Identification of tweets that mention books. *International Journal on Digital Libraries*. DOI: 10.1007/s00799-019-00273-4

As the second module in the TMB collection step, the TMB identifier retrieve TMBs from general tweets. In this chapter, we define a TMB identification task and solve it by machine learning with task-oriented features.

The main contributions of this chapter are the following.

- We tackled the task of identifying book titles among informal texts, which is held to be one of the most difficult named entity recognition tasks, i.e. identifying named entities in a wide variety of forms from informal texts.
- We focused on user-generated posts about books, and tried removing not only unrelated noise posts but also machine-generated posts.
- Our proposed method achieved comparable performance (0.76 F1-score[1]) to the highest-scored methods in related tasks.

The rest of this chapter is organised as follows. We first give a formal definition of our task in Section 7.1. Then, Section 7.2 summarises existing studies similar to our identification task. We elaborate on our methods in Section 7.3. The procedures and results of experiments are described in Section 7.4. In Section 7.5, we provides the results of analyses and diagnosis on the performance of our method. Finally, Section 7.6 concludes the chapter.

7.1 Task definition

TMBs can take various forms, which makes the task of identifying TMBs hard. The phrase 'tweets that mention books', in a literal sense, can mean a wide variety of mention styles.

[1] F1-score (F) is defined by $F = 2 \cdot P \cdot R/(P + R)$, or the harmonic mean of precision (P; the number of relevant documents divided by the number of retrieved documents) and recall (R; the number of relevant documents in the retrieved documents divided by the total number of relevant documents).

In order to make the task of TMB identification feasible, we target a specific type of TMB as the first step, which is TMBs using book titles.

As we already defined in Section 5.1, our target TMBs are *manually generated Japanese tweets that mention specific books.* The most popular way to specify books in social communication is to refer to *book titles* amongst bibliographic information fields, as we explained in Section 5.2.1. Here is a TMB example using book titles, found in English Twitter:

> *When Breath Becomes Air*[2] is the most profound, life changing book I have ever come across. It will stick with me through my whole life

It is, thus, reasonable to start from TMBs with book titles. Henceforth, otherwise noted, 'TMBs' denotes tweets that mention books referring to their titles.

This allows us to start with simple pattern matching for extracting tweets that contain the same expressions as book titles, or *title strings.* Collecting such tweets that contain title strings (TCTSs) is supported by up-to-date, comprehensive lists of book titles, availability of which are usually guaranteed through, for example, book catalogues or bibliographic databases compiled by the social effort from national libraries and/or bookstores. In other words, we can start from TCTSs rather than general tweets to find TMBs.

Because TCTSs still contain non-TMBs, however, we have to remove these to obtain the desired TMBs. It is still essential to distinguish TMBs and non-TMBs amongst TCTSs. Amongst TCTSs, we identified two types of non-TMBs: *bot* and *noise* tweets.

Bot tweets are TCTSs that may or may not refer to books but are posted automatically, or more specifically, those which are not created by humans' manual actions. Recall that we excluded them from the range of TMBs we target. Most of these tweets are posted by automated accounts, or bots. Some are apparently intended for carrying promotional information of books, like discounts of books at miscellaneous online bookstores, while others lure users to websites filled with advertisements. One example is:

> <Get Max 500 yen Discount!> A Book Title / Its Author [Post Payment available] [Free Shipping if 2500+ yen] [05P06May15] URL [originally in Japanese]

Almost all of these tweets are automatically generated following some schedules and/or other conditions like page updates in external websites. These cases are covered by bot (or spam) account detection research we introduced in Section 7.2.

[2] In this thesis, italic strings within tweet examples denote actual book titles.

However, in *bot tweets*, we also include cases where human accounts let automation services post tweets in place of themselves. By such automation, they intend, for instance, to circulate their interest in books or to promote their own work. The following is an example:

> This is the scheduled tweet about my favourites!!
> Anime/Game/Manga/Fan Fiction/ An anime title /. . . (24 instances). . . / (originally in Japanese)

Although some studies try to define such accounts as *cyborgs* (Chu et al., 2012) and struggle to put the threshold between bots and cyborgs, we do not make this distinction, because tweet-level automation is of concern here.

Note that we set one exception to the definition of bot tweets: social reading services (SRSs). SRSs (e.g. GoodReads) allow users to automatically post tweets corresponding to their actions like marking a book as read[3]. We regard these integrated tweets as TMBs because they reflect users' online manual actions related to reading.

Noise tweets are manual TCTSs that do not refer to books. One of major noise types comes from the fact that many book titles consist of ordinary expressions, e.g. *Kidnapped*, *A Night in Paris*, *From Anna*, and *See Me*. A real English tweet example is:

> To *the girl on the train* who is currently drawing her eyebrows on. No.

Another type is TCTSs that mention other media contents, such as films, TV shows, songs, and video games, whose titles are identical to existing book titles. For instance, if a tweet said "I can't stop watching *man in the high castle. . .*", it would hardly ever mean the original 1962 novel written by Philip K. Dick, but would mean its TV show version. As we will see, the number of noise tweets is much larger than that of TMBs amongst TCTSs.

The TMB identification task is now formulated as a multi-class classification task over TMBs, bot, and noise tweets. We thus define the TMB identification task as follows: *to classify given TCTSs into TMBs, bot, and noise tweets*, among which TMBs are our primary focus.

7.2 Related identification tasks

Identification of tweets that mention certain entities

Some studies carried out the task of identifying tweets that mention a particular type of entities among noisy tweets. For example, the CLEF RepLab workshop had a shared task of

[3] See, for example,
https://www.goodreads.com/help/show/68-how-do-i-add-my-reading-updates-to-twitter

tweet filtering on English and Spanish tweets collected by keyword search (Amigó, Carrillo de Albornoz et al., 2013; Amigó, Corujo et al., 2012). Target entities included automotive industries, banking companies, universities, and music artists. The data sets contained three times more relevant tweets than irrelevant (noise) tweets, which is the opposite situation to ours. Participants of the task often utilised external knowledge bases such as Wikipedia[4] and Freebase[5] information to extend features available in tweets themselves.

Erdmann et al. (2013) tackled the issue of extracting TV program titles from tweets. Since the work was aware of noise tweets generated by searching for TV show titles, it trained a Support Vector Machine (SVM) (Cortes and Vapnik, 1995) for ambiguous (or noisy) TV show titles while unambiguous titles were filtered by a rule-based method checking if titles were registered as Wikipedia articles. For the SVM classifier, rule-based features were used such as occurrence of TV-related words, and of character names of mentioned TV programs.

Prasetyo et al. (2012) applied text classification to collect tweets that contain useful information related to engineering software systems from tweets that contain software hashtags. They represented tweets as bag-of-words vectors and expand vocabulary by looking at web pages to which the tweets refer. These features were fed to an SVM classifier which achieved 0.71 F1-score.

Guzman, Alkadhi and Seyff (2017) conducted a multi-label classification task of tweets that mention software in which the tweets were assigned to one or more labels about the relevance to the three different types of stakeholder groups: technical (software developers in companies), non-technical (non-developers in companies), and general public (end-users). Using Bag-of-Words features weighted by term frequency–inverse document frequency (TFIDF (Salton and C.-S. Yang, 1973)), the work compared the performance with different classification algorithms, such as naive Bayes (McCallum and Nigam, 1998), SVM, and random forests (Breiman, 2001). These classifiers performed differently in precision and recall but similarly in F1-score ($F \approx 0.75$), although the label of technical stake holders had lower quality (the maximum F was 0.52). They also carried out bot account detection on the same data set, which we will mention later.

Aramaki, Maskawa and Morita (2011) predicted the outbreak of influenza. Like our task, they first collected tweets possibly related to influenza by using a set of keywords such as ‘flu’, and then classified them into tweets posted by influenza patients and those that were not. They used Bag-of-Words features in tweets with 6-words window size, i.e. including only up to $6 \times 2 = 12$ words from left and right sides of influenza-related terms

[4] https://www.wikipedia.org/

[5] https://developers.google.com/freebase/

in the tweet. According to the result, SVM performed best ($F = 0.756$) among different classifiers, e.g. logistic regression (or maximum entropy modelling; MaxEnt), naive Bayes, nearest neighbours, random forests, AdaBoost (Freund and Schapire, 1997), and Bagging (Breiman, 1996). To collect tweets that mention the user's health, Tuarob, Tucker et al. (2014) also proposed an ensemble of different classifiers (e.g. naive Bayes, random forests, and SVM) trained by different features such as n-grams, dictionaries, topic models, and sentiments. Their performance distributed between 0.51 and 0.77 in F1-score.

Bot detection

Bot detection is a popular research field on Twitter (Alothali et al., 2018; Chu et al., 2012; A. Wang, 2010). Studies in this field aim to detect *bot accounts*, or automated Twitter users. Varol et al. (2017) estimated that between 9% and 15% of Twitter accounts are bots. Chu et al. (2012) further defined *cyborgs*: bot-assisted humans and human-assisted bots. Twitter users can register their accounts to an automation services to post, e.g. the updates of their websites, which is one example of bot-assisted humans. A customer support bot run by an automated program may be occasionally taken over by a real human staff in a complex conversation with customers; this is one type of human-assisted bot.

Bot accounts often post *spam* tweets (Chu et al., 2012). Therefore, we can find that spam account detection research uses similar frameworks to bot account detection (Alothali et al., 2018; Verma, Divya and Sofat, 2014; T. Wu et al., 2018). Unlike bots, spam has several definitions for different aims and motivations. One defines spam as 'spreading malicious, phishing, or unsolicited commercial content in tweets' Chu et al., 2012. Some papers do not provide a clear definition (e.g. (Grier et al., 2010; McCord and Chuah, 2011; A. Wang, 2010; T. Wu et al., 2018)). T. Wu et al. (2018) only implies that spam is one form of attacks from criminal accounts (which are called spammers in the work). We did not adopt the term 'spam' because our definition of bot tweets includes ones from innocent bots and from humans who just intend to connect to more people (cf. Section 7.1).

According to several surveys (Alothali et al., 2018; Verma, Divya and Sofat, 2014; T. Wu et al., 2018), features used in bot or spam detection research can be summarised as follows: user/account information, social graphs, tweeting behaviour, and textual contents. User/account information includes ages of accounts, screen names, and numbers of following users, users followed by, past tweets, liked tweets, Retweets, etc. Social graph features incorporate following/followed networks of users. Tweeting behaviour typically exploits times and intervals of users' recent tweets, e.g. the ratio of original tweets and Retweets. Finally, textual contents found in each tweet and users' profiles are examined in terms of numbers of hashtags, URLs, mentions, etc. In addition to features, the surveys report that most studies applied similar classification algorithms as the research of Section 7.2 did. The performance ranges from 85% to 98% in F1-scores or other metrics (Alothali et al., 2018).

Botometer[6] and DeBot (Chavoshi, Hamooni and Mueen, 2016) are instances of state-of-the-art bot detectors that publicly provide their application programming interfaces (APIs). Botometer uses more than 1,000 features and makes full use of users' past tweets. DeBot utilises the correlated behaviour of multiple bot accounts.

While almost all of bot or spam detection work formulates the task into spam/bot *account* classification, our study filters bot *tweets*. We thus aim to implement a novel method to filter bot tweets using the features available from just one tweet. Although Guzman, Alkadhi and Seyff (2017) tried a similar setting of bot detection, it still formulates the task into classifying whether a given tweet was posted by a bot account or not.

Recognition of a single named entity

If we focus on book titles in TCTSs, the TMB identification task can be considered as a named entity recognition (NER) task targeting a single type of named entities (NEs): book titles. A number of studies focus on extracting a single NE. One of the most popular named entities is biochemical-substance names (Crichton et al., 2017; Gridach, 2017; Kou, W.W. Cohen and Murphy, 2005).

Book titles can appear in contexts completely unrelated to books because book titles take a wide variety of linguistic forms from noun phrases to adjectival and verbal phrases; even clauses and sentences are permitted. Furthermore, they often consist of ordinary expressions such as *Kidnapped* or *A Night in Paris*. These are the reasons why book titles are supposed to be formatted with styles in formal texts; they should be made italic in English and parenthesised in Japanese. In fact, such patterns are exploited in a few studies that addressed book title recognition (Brin, 1999; Downey, Broadhead and Etzioni, 2007) from large semi-formal web corpora.

We cannot assume orthographically formatted texts in social media. Derczynski et al. (2014) summarised the difficulties of NER in Twitter: short messages, noisy content, social context, user-generated, and multilingual. They also confirmed that general NER methods, designed for more-or-less formal written documents (e.g. newspaper articles (Sang and De Meulder, 2003)), perform worse than the Twitter-specific NER methods, such as T-NER (Ritter et al., 2011). T-NER, a pipeline of NE segmenter and NE classifier, is still a strong baseline of NER in tweets (Habib and Van Keulen, 2016). However, its score for film and

[6] This was formerly called BotOrNot (C.A. Davis et al., 2016; Varol et al., 2017) (https://botometer.iuni.iu.edu/).

TV show titles, which have similar characteristics to book titles, was around 10 points lower than the overall score (F1-score 0.565 in average vs. 0.66 in total). The best method (Limsopatham and Collier, 2016) in a tweet NER shared task (Strauss et al., 2016), the evaluation data set of which was an augmented version of Ritter et al. (2011)'s collection, also failed to recognise film and TV show titles (0.11 and 0.06 in F1-score respectively, in contrast to 0.52 in total). Thus, especially in tweet-like noisy texts, book titles appear to be far more difficult to recognise than common traditional NEs such as names of people, locations, and organisations (Sang and De Meulder, 2003; Sekine and Nobata, 2004).

7.2.1 Text classification algorithms

For text classification tasks, such machine learning (ML) algorithms as naive Bayes (McCallum and Nigam, 1998), maximum entropy modelling (MaxEnt; Nigam, 1999), and Support Vector Machine (SVM; Cortes and Vapnik, 1995) have been widely used. These methods have also been applied to the classification of tweets, for example, identifying spam tweets (McCord and Chuah, 2011; A. Wang, 2010) or sentiment and opinion analyses (Go, Bhayani and L. Huang, 2009; Pak and Paroubek, 2010), showing different results depending on the applications and parameter settings.

Recent studies apply deep neural network (NN) models to text classification (Lai et al., 2015; Ororbia Ii, Giles and Reitter, 2015). Most of these models require larger corpora than the non-neural algorithms to achieve good performance. For specific NE recognition tasks or specialised information retrieval tasks, it is often difficult to collect a sufficient amount of annotated data to learn (cf. Crichton et al., 2017) When the target data is limited to a particular text type, size issues become even more critical.[7]

7.3 Methods

To tackle the TMB identification task defined above, we propose a two-step binary-classification pipeline. We first provide the motivation for adopting this pipeline, and then elaborate on what features we adopt and propose. We also specify the classification algorithms we used.

7.3.1 Classification strategy

Through the process of our manual annotation of TCTSs, we found that bot tweets have specific features in format and in metadata compared to TMBs and noise tweets. As mentioned above, bot tweets are automatically generated by schedules and conditions. In

[7] Replacing the algorithms we adopted in this chapter to simple NN models (e.g. multilayer perceptron, and convolutional or recurrent-NN layers) caused no significant improvement.

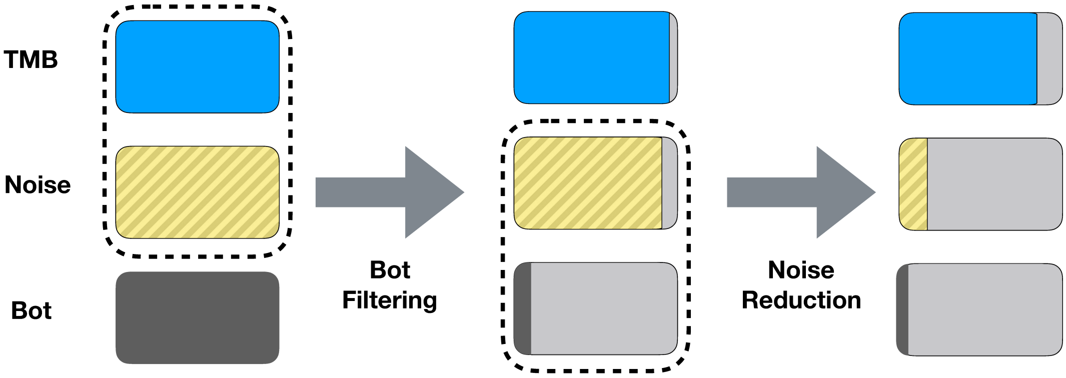


Figure 7.1: The architecture of our proposed method, or the two-step pipeline

bot/spam account detection tasks, in fact, metadata features are popular (P. Kaur, Singhal and J. Kaur, 2016; T. Wu et al., 2018).

In contrast, TMBs and noise tweets are both human generated. Their difference is not in their behavioural aspect but in their textual expressions. Remember the aforementioned, hypothetical example of noise tweets: "I can't stop watching *man in the high castle…*". We can recognise that this tweet is not about a book because it uses the word 'watching'. If this word was 'reading', we would consider the tweet to be a TMB.

Based on these observations, we split this task into two binary classification tasks, thus proposing the following **two-step pipeline** to solve the TMB identification task (Figure 7.1):

1. *bot filtering* (bot tweets vs. TMBs and noise tweets), and
2. *noise reduction* (noise tweets vs. TMBs).

7.3.2 Feature design

Information extracted from tweets can be classified into body text and metadata. Metadata is divided into two types: metadata of tweets themselves and metadata of users who posted the tweets. The former includes the time of posting and the number of retweets and likes, whereas the latter includes profile text and the number of following and followed accounts. Each step of the two-step pipeline aims to classify tweets with different characteristics. We define separate features for these steps. The proposed features, which we elaborate on below, are summarised in Table 7.1.

Table 7.1: Features

Bot Filtering		
Metadata	`app`:	application name tokens used to post tweets
	`f/f`:	the ratio of followings/followers
Tweet body texts	`hashtag`:	the number of hashtags
	`url`:	tokens of URL host names
Noise Reduction		
Metadata	`app`:	application name tokens used to post tweets
Tweet body texts	`url`:	tokens of URL host names
	Word tokens:	tokens of words in tweets
	`distinct`:	distinguish tokens before/after titles
	`bib`:	abstract bibliographic information tokens
	`titleness`:	add PoS tags of tokens within titles

Table 7.2: Top 3 frequently used Twitter client applications to post tweets in our data set.

Bot	Twitter for iPhone	38.3%
	Twitter for Android	19.5%
	Twitter Web Client	16.9%
Non-bot	dlvr.it	15.6%
	IFTTT	11.7%
	tweeterfeed	4.5%

Bot filtering

Bot tweets are defined as automated tweets. A study reported that 87% of tweets by bot accounts were posted via automated tools, whereas 70% of human tweets were generated from Twitter's official website or mobile devices (Chu et al., 2012). Our data set (Section 7.4.1) also shows different distributions of Twitter client applications used for tweeting between bot tweets and non-bot (TMB and noise) tweets (Table 7.2).

In spite of this strong discriminative power, the difference of applications is rarely used in bot/spam detection research (Alothali et al., 2018; Verma, Divya and Sofat, 2014). Even the state-of-the-art bot detectors do not utilise this information. Chu et al. (2012) used it as features by calculating the ratio of automated applications used in recent tweets and that of manual clients for each user based on handcrafted lists of a few applications. Guzman, Alkadhi and Seyff (2017) took a simpler approach to check whether the string 'bot' appears in the user name or the application name, but this approach did not work very well. Considering that these existing methods do not scale when the size of the data increases, we propose a way to make the classifier learn the difference of applications between bot and non-bot tweets directly from the data: the one-hot vector representations of application names (`app`, on which we elaborate below).

However, some bots pretend to be human by automating clients for humans. For example, tweets via 'Twitter for Web' can be generated by using web browser automation tools like Selenium[8]. Therefore, we additionally adopt widely used features in bot/spam account detection studies. While account-level rich features such as user information and tweeting behaviour have been popular (Verma, Divya and Sofat, 2014; T. Wu et al., 2018), we used simpler features available in tweet text. This is mainly because we should avoid relying too much on account-level features. This point comes from the fact that bot tweets can be posted from human accounts, as we introduced in Section 7.1. Besides, retrieving additional tweets, which is required for behavioural features, costs time due to the rate limit of the Twitter API. We adopted two features from the body text and one feature from metadata. The former two are the number of hashtags (`hashtag`) and the tokens of host names in URLs (`url`). Many studies agree that bot/spam tweets contain more hashtags (McCord and Chuah, 2011; Verma, Divya and Sofat, 2014). It is also reported that similar URLs, typically linked to malicious or unsolicited websites, appear across tweets from different bot/spam accounts (Chu et al., 2012; Grier et al., 2010). These features perform constantly well for different data sets (T. Wu et al., 2018). Finally, among typical account information features, we selected the ratio of the number of following accounts to that of follower users (`f/f`). Bots are unlikely to be followed but tend to follow many other accounts (Alothali et al., 2018; McCord and Chuah, 2011). Unlike the absolute numbers of following or follower accounts, which are frequently used, the ratio stays constant even if data sets grow. This characteristic can help the model's generalisability to unseen data in which some users may have a very large number of following or follower accounts.

These features in bot filtering are generated as follows.

Metadata

Application name (`app`): Every distinct application name is represented as a one-hot vector of $\mathbb{R}^{|V| \times 1}$ with all 0s and one 1 at the index of that application name in the sorted application names.

Following/followers ratio (`f/f`): We use the numbers of following and follower accounts of each author of tweets and calculate the ratio of following/followers. If a follower count is zero, we assign -1 to this value.

Tweet Body Text

[8] https://www.seleniumhq.org/

Hashtag count (`hashtag`): The number of hashtags within the tweet body text is counted.

URL host name tokens (`url`): This generates a Bag-of-Words matrix with normalised inverse document frequency (IDF) of host names extracted from URLs that appear in tweets (e.g. ‘`twitter.com`’ in ‘https://twitter.com/en/privacy’). This aims to capture common malicious or unsolicited URLs used in bot tweets. While Bag-of-Words features are often weighted by TFIDF, we adopt IDF weighting only. Due to the character limit of tweets, it is expected that the same token will not occur so frequently in a tweet. In this situation, IDF tends to perform better than TFIDF as reported in some studies (Tuarob and Mitrpanont, 2017; Yada and Kageura, 2015).

Noise reduction

From the overview of the related tasks in Section 7.2, textual features in tweet body text supported by domain knowledge (e.g. Wikipedia and dictionaries) play a key role. As we mentioned in Section 7.1, noise can be classified into the following two major types:

(a) book title strings are referred to as common meanings because titles consist of ordinary expressions (e.g. *1984*, *Dune*, *See Me*, and *The Right Stuff*) and;

(b) book title strings are mentioned as other works or things with identical names such as film, music, and personal or geographic names (e.g. *Gone Girl*, *Forrest Gump*, and *Michael Jackson*).

In both cases, contextual words around title strings give critical information to differentiate TMBs from noise. Consider the difference in these mock examples:

(a) ‘I have bought “*see me*” at a bookshop’ (TMB) vs. ‘plz *see me*!’ (Noise)

(b) ‘I freaked out after reading *gone girl*’ (TMB) vs. ‘Must-see movie, *Gone Girl*’ (Noise)

Thus, we mainly focus on the textual information available from the text of tweet bodies in noise reduction. We start from using words appearing in tweets as features in orthodox Bag-of-Words representations, as in similar tasks. In this *word tokens* feature, we replace title strings with the same symbol (placeholder) so as to prevent our model from overfitting. Since the number of book titles are growing, the classifier should be able to distinguish TMBs based on how books are mentioned rather than what books are mentioned.

In addition, we further propose three post-processing options for *word tokens* in order to optimise this feature for the TMB identification task. The first is to distinguish tokens from before and after title strings (`distinct`). That is, when the identical tokens appear in both left and right sides of title strings, we regard tokens from different sides as different words. While Bag-of-Words representations do not keep the order of words in text, we found that characteristic words, such as reading-related verbs and quotation marks to denote titles,[9] cluster around title strings. The typical way to consider word order is Bag-of-Ngrams representations. However, Ngrams cause sparsity in features and increase the computational load. The `distinct` option can provide denser features with a faster speed. The *word tokens* feature with the `distinct` option enabled (or simply, the `distinct` *feature*) works similar to the 'window size' in Aramaki, Maskawa and Morita (2011) (cf. Section 7.2). Unlike this study, we do not filter the words outside of the window size. The difference in the number of words between TMBs and noise tweets may contribute to the performance, since TMBs are about 20% longer in average than noise tweets in our data set.

The second option is to abstract bibliographic information (`bib`). Some TMBs also contain bibliographic information besides titles, e.g. authors and publishers. It would be effective to give the classifier explicit signals that some tokens are bibliographic information of mentioned book titles. This also corresponds to utilising domain knowledge, which is popular in related tasks (Section 7.2).

The third option is to include Part-of-Speech (PoS) tags in the title strings (`titleness`). We decided to replace title strings with placeholders to alleviate over-fitting, which may omit some useful information about the typical forms of books. For instance, title strings consisting of mainly nouns are more likely to be book titles than those consisting of interjections or verbs (e.g. *A Brief History of Time* vs. *See Me*). We take into account the extent of formal typicality of book titles experimentally from PoS information.

In addition, we exploit two features from bot filtering: `app` and `url`. The `app` feature will work because some TMBs are posted via SRSs. TMBs containing URLs directed to book-related websites, such as bookstores and book reviews, can be distinguished by `url` from noise tweets with links to book-unrelated websites.

In summary, the following features are defined in noise reduction:

Metadata

Application name (`app`): The same as in bot filtering.

Tweet Body Text

URL host name tokens (`url`): The same as in bot filtering.

Word tokens: As the main feature in noise reduction, this generates the Bag-of-Words vector representation of the tweet weighted by IDF.[10] We applied the following preprocessing, including Japanese-specific operations:

[9] In Japanese, book titles are supposed to be enclosed by double quotation marks (i.e. 『』).

[10] Note that all values in every vector are normalised with its L2 norm.

i. replace URLs and mentioned user names with the @ symbol into `%URL%` and `%REPLY%` respectively;
ii. replace all emojis with `%EMOJI%` so as to reduce the noise that might result from distinguishing between emojis;
iii. convert title strings into `%TITLE%` in order to abstract them to focus on contextual words around titles;
iv. normalise similarly shaped characters inconsistently used in Japanese text into one instance;[11]
v. replace all blank characters (such as blank spaces and line breaks) into `%S%` to capture the visual format in tweets[12]; and
vi. apply the Japanese morphological analyser `MeCab`[13] to split Japanese text into words (tokens).

We applied the following optional post-processing to extracted word tokens.

Distinguish tokens before/after titles (`distinct`):

To distinguish tokens that appear before/after the title string in each tweet, the suffix `-/+` is added respectively. Furthermore, to capture neighbouring words around the title string, the distance from the title strings is appended to tokens within the 3-word window. For example, the sentence "I have finished reading `%TITLE%`! It's really interesting!" becomes ['I-', 'have-3', 'finished-2', 'reading-1', '`%TITLE%`', '!+1', "It's+2", 'really+3', 'interesting+', '!+'].[14]

Abstracting bibliographic information (`bib`):

If any words in a TMB matches bibliographic elements (other than book titles) of a mentioned book, the words are replaced with the corresponding names of the bibliographic elements. For example, if 'Haruki' and 'Murakami' appear in a TMB that mentions the book *1Q84* authored by Haruki Murakami, these two tokens will be converted into `%CREATOR%`. We consider six bibliographic elements: creators (authors), publisher, edition, volume, publication year, and titles of short stories if the mentioned book is a collection. The corresponding bibliographic information of mentioned titles is retrieved from a Japanese bibliographic database called Webcat Plus[15].

[11] We adopted a popular normalisation scheme for Japanese informal texts (https://github.com/neologd/mecab-ipadic-neologd/wiki/Regexp).

[12] Some tweets use blank spaces and line breaks to be more readable.

[13] https://taku910.github.io/mecab/

[14] This example does not consider any preprocessing which is typically applied in English text processing. Note that we target Japanese tweets.

[15] Webcat Plus is a unified bibliographic database run by the Japan National Institute of Informatics. http://webcatplus.nii.ac.jp/

***Adding PoS tags of title tokens* (`titleness`):**
This option adds PoS tags of the title string tokens into the *word tokens* feature. We used MeCab to obtain PoS tags of Japanese title strings.

7.3.3 Classification algorithms

As reviewed in Section 7.2, related research applied different machine learning (ML) algorithms such as naive Bayes, MaxEnt, random forests, and SVM. These algorithms showed different results, depending on task and parameter settings. Since the TMBs we deal with appear less frequently than non-TMBs, as we will show in Section 7.4.1, the proportion of classes within data is imbalanced in comparison to general text classification tasks. The methods we choose thus should be robust enough to handle imbalanced data.

We chose two algorithms that have been widely used and proved to have good performance in related studies: **MaxEnt** and **SVM**.[16] These methods outperformed other algorithms such as naive Bayes and random forests in our previous study (Yada and Kageura, 2015). We will consider both the parameters and the proposed features in our experiments, since these methods have hyper parameters.

MaxEnt: We consider the following two hyper parameters: The inverse regularisation strength C and the norm used in the penalisation, which is chosen from L1 and L2.

SVM: We use the radial basis function (RBF) kernel for SVM, because our preliminary experiment shows better result than linear, polynomial, and sigmoid kernels. We take into account the penalty parameter of the error term (C) and the RBF kernel coefficient (γ).

In this work, we do not adopt neural network (NN) models that recent studies apply to tweet classification (Prusa and Khoshgoftaar, 2017; Severyn and Moschitti, 2015). It is usually harder to interpret the linguistic features learned with NN models in compared with those learned with non-NN models (Belinkov and Glass, 2019). The interpretability of feature effects is preferable for our task because it will help us understand the linguistic characteristics of book mentions online. This is useful for a future descriptive analysis. Besides, most NN models require much larger corpora than the non-NN algorithms to achieve a good performance. For specific NE recognition tasks or specialised IR tasks, it is often difficult to collect a sufficient amount of annotated data from which to learn (cf. (Crichton et al., 2017)). When the target data is limited to a particular text type, size issues become even more critical.

[16] We adopted implementations of `scikit-learn` (Pedregosa et al., 2011) (v0.20.2) for both algorithms.

Table 7.3: TMB identification dataset with bot and noise tweets annotated

	Num. of tweets	Num. of accounts
TMB	450	441
Bot	4,740	3,439
Noise	5,601	5,553
Total	10,791	9,483

7.4 Data and experiments

7.4.1 Data

We used the TMB identification dataset (Section 5.2.1) for evaluation experiments. Following the definition of bot and noise tweets, they were further distinguished from non-TMBs in the dataset. Through this additional finer annotation, we obtained **450** TMBs, **4,740** bot tweets, and **5,601** noise Tweets, as shown in Table 7.3. Under the identical setup in Section 5.4.1, the inter-coder agreement of this three-class annotation was 0.888, which can be interpreted as a high-level agreement Artstein and Poesio (2008). In this data set, there are 441 distinct users in TMBs, 3,439 in bot tweets, and 5,553 in noise tweets. Also note that this dataset is published as Yada (2019).

7.4.2 Experimental setups

We conducted two experiments to evaluate our method. In **Experiment 1**, we investigated the best combinations of features along with algorithms and their hyper parameters separately for each step. This experiment is also a preparatory step to build the pipeline. That is, we use the best classifiers for each step in our two-step pipeline. We also compared the performance of the best classifiers with relevant baselines.

In **Experiment 2**, we built the two-step pipeline using the best combinations found in Experiment 1 and applied them to TMB identification. In this experiment, we also applied one-step classifiers as baselines to evaluate whether the two-step method we adopted was effective or not.

Experiment 1

Procedures For both bot filtering and noise reduction, we conducted exhaustive grid searches for all combinations of algorithms, their hyper parameters, and proposed features. The candidate values of hyper parameters to search for are listed in Table 7.4. In the grid

Table 7.4: Candidate values of hyper parameters to search for in Experiment 1

Algorithm	Candidate values
MaxEnt	$C = \{0.01, 0.1, 1.0, 10.0, 100.0\}$, norm = {L1, L2}
SVM	$C = \{0.01, 0.1, 1.0, 10.0, 100.0\}$, $\gamma = \{0.01, 0.1, 1.0, 10.0, 100.0\}$

search for bot filtering, we used the entirety of the annotated data, whereas only the data consisting of TMBs and noise tweets was used in the grid search for noise reduction. For each combination of parameters and features, classifiers trained and predicted the data in a three-fold cross validation (CV) manner (3CV). All three splits of tweets were randomly sampled with the proportion of labels kept (i.e. so-called *shuffled and stratified* CV).

The best classifiers were selected by a single measure: the mean F1-score for the class of interest over 3CV, i.e. the bot class in bot filtering and the TMB class in noise reduction. For these best classifiers, we report mean values of precision, recall, F1-scores, and area under the precision-recall curves (AUC)[17] over newly conducted 10CV. Additionally, the proportion of TMBs wrongly rejected in bot filtering was calculated as **TMB loss**.

In bot filtering, these scores for both bot tweets and non-bot tweets (i.e. TMBs and noise tweets) are reported, while those of TMBs are shown in noise reduction. Most binary classification experiments consider the scores of the class that is of interest at that point in the classification (or the *positive* class) only. Since bot filtering is the first step of the pipeline, it should achieve high scores in both labels (positive and negative) to filter as many bot tweets as possible, and to let through as many TMBs as possible at the same time.

Baseline Methods For bot filtering, we adopted a simple rule-based method and the state-of-the-art bot (account) detector as baselines.

Baseline A (rule-based): This filters non-human application names, because they should be a strong indicator of tweet automation (see Section 7.3.2). This also shows the bare proportion of tweets posted by non-human application names in our data set. We created a list of 30 popular Twitter client applications that do not have any functions to automate tweeting by investigating the top-ranked mobile and desktop applications for Twitter distributed in the application stores of iOS, Android, ma-

[17] While the area under the curve is often calculated in receiver operating characteristic curves, precision-recall curves is preferred in imbalanced data sets (J. Davis and Goadrich, 2006; Saito and Rehmsmeier, 2015).

cOS, and Windows in 2015. Note that we also added SRSs found in our data set to the list. This rule-based baseline identifies application names not found in the list as bot tweets.

Baseline B (Botometer): Among the two publicly available state-of-the-art bot account detection methods introduced in Section 7.2, we adopted Botometer, because it provides a high capacity API rate limit for free[18]. The comparative evaluations are carried out as follows. As a preparation, we fed 200 recent tweets from the user and 200 recent mentions of the user collected on 26 January 2019 to Botometer, as they are required for Botometer to operate. In comparative evaluation, we used a subset of our data, consisting of 1,887 bot tweets and 3,517 non-bot tweets, which in turn consist of 331 TMBs and 3,179 noise tweets. This is because 62.5% of the accounts were not available on Twitter at the time of the experiment, and Botometer requires accounts to be alive.

Botometer returns the complete automation probability (CAP) for the given account, the higher value of which means a higher possibility for the account to be bot. We set the threshold of CAP to maximise the F1-score for the bot class, which was 0.034. In order to conduct a fair comparison, we also applied our method, i.e. the best bot filtering classifiers we had found in the grid search, to the same subset of our data, instead of using the overall performance. In the re-experiment, we conducted 5CV on the subset, by training the classifiers on 4/5 of the subset plus the rest of our data set and tested them on 1/5 of the subset, with the train/test split rotated. In this scheme, the part of our original data which were not used for CV are regarded as a pool of training data. Note that Botometer is pre-trained. The performance was evaluated from the point of view of bot *tweet* detection, as this was our original task. To make it possible to evaluate the performance of Botometer from this point of view, we judged that all tweets of the accounts that were judged as bot by Botometer were regarded as bot tweets.

For the further comparison, the rule-based baseline was also applied to this subset.

For noise reduction, we used a rule-based method and three ML-based methods.

Baseline C (rule-based): We utilised a book-or-reading related word list which contains words appearing more frequently in TMBs than noise tweets, since contextual words are expected to hold fundamental information to identify TMBs. The baseline for noise reduction treats tweets that contain both title strings and any word in the list as TMBs. Words in the list were ranked by the log relative ratio of word frequencies (Tang and H.-H. Chen, 2012) of TMBs over noise tweets, and the top n

[18] https://rapidapi.com/OSoMe/api/botometer-pro

ranked words were chosen so as to result in the maximum F1-score. Through a 10CV experiment, the mean value of n was 79.7 (std: 14.2).

Baseline D (TFIDF): This is an SVM classifier with Ngrams ($N \leq 3$) of tweet texts weighted by TFIDF. Ngrams are generated by following the same preprocessing as in the *tweet tokens* feature, except the step iii (replacing title strings to `%TITLE%` tokens). Most related work of tweets that mention certain entities (Section 7.2) adopted the combination of a classification algorithm and TFIDF feature. This baseline simulates a typical setting of tweet classification without any task-specific features, which enables us to observe the effectiveness of our features. We separately conducted a grid search for this baseline, considering the following parameters: C and γ of SVM, candidates of which are the same as in Table 7.4, and the proportion of features kept by mutual information-based feature selection (Chandrashekar and Sahin, 2014) among $\{40, 60, 80, 100\}$%. The best parameters were: $C = 100$, $\gamma = 0.1$, and 60% of features.

Baseline E (D + Title): This is a modified version of Baseline D, considering title string information. While Baseline D uses Ngrams of plain tweet texts, Baseline E uses Ngrams of tweet texts in which title strings are replaced with the placeholders (`%TITLE%` tokens). In other words, this baseline can distinguish the position of title strings in TCTSs like the `distinct` feature does. We separately conducted a grid search for this baseline as well, considering the same hyper parameters as Baseline D. The best parameters were: $C = 1000$, $\gamma = 0.001$, and with 80% of features.

Baseline F (E + All): This is a modified version of Baseline E, in which all features for noise reduction except `distinct` are included. Comparing with our method will give the performance difference between Ngrams and `distinct` in relation to other proposed features. We separately conducted a grid search for this baseline, considering the same hyper parameters as Baseline D and E. The best performance was obtained with `url`, `app`, and `bib`. The best parameters were: $C = 1000$, $\gamma = 0.01$, and with 100% of features.

Experiment 2

Procedures We made two-step pipelines by combining the best classifiers of each classification step ($2 \times 2 = 4$), and applied them to the entirety of the annotated data with the shuffled and stratified 10CV. The two-step pipelines conducted bot filtering of the whole input tweets at first, and then ran noise reduction of the bot-filtered tweets. We evaluated the performance by the same metrics as Experiment 1, except AUC. While AUC is defined in binary classification, this experiment conducts multi-class classification.

Baseline Methods In order to evaluate our two-step architecture, we compared its results with the one-step (or so called 'one-versus-the-rest') method as the baseline. Four baselines with different schemes were compared with our two-step pipelines:

Baseline D' This baseline has the same architecture as Baseline D for noise reduction in Experiment 1, i.e. an SVM using Ngrams ($N \leq 3$) features weighted by TFIDF with mutual information-based feature selection. After a 3CV grid search, Baseline D' was configured as follows: $C = 10$, $\gamma = 0.1$, and with 80% of features.

Baseline E' This baseline has the same architecture as Baseline E for noise reduction in Experiment 1, i.e. Baseline D' aware of title strings. After a 3CV grid search, Baseline E' was configured as follows: $C = 10$, $\gamma = 0.1$, and with 80% of features.

Baseline F' This baseline has the same architecture as Baseline F for noise reduction in Experiment 1, i.e. Baseline E' considering all noise reduction features except `distinct`. After a 3CV grid search, Baseline F' was configured as follows: $C = 1000$, $\gamma = 0.001$, and with 100% of the features in which `app`, `url`, `bib` and `titleness` were enabled.

Baseline G This baseline adopts all features we proposed for both steps, and not using Ngrams. We carried out another grid search (3CV) for algorithms, their hyper parameters, and features. As a result, the baseline was set to an SVM classifier of $C = 100$ and $\gamma = 0.01$ with the `hashtag`, `app`, `url`, `distinct`, and `bib` features enabled.

7.4.3 Results

Experiment 1

After the grid searches, we found the best combinations for the features and hyper parameters for both algorithms.

Bot Filtering Table 7.5 shows the best feature settings and classification scores in bot filtering. The best hyper parameters, which are omitted from the table, were as follows: $C = 10.0$ and L1 norm for MaxEnt; $C = 10.0$ and $\gamma = 0.1$ for SVM. The best combination of features were different in the two algorithms. While the best MaxEnt used only two features, SVM was best with all features included. Nevertheless, both algorithms performed similarly in scores.

All scores exceeds 0.97, which is a highly satisfactory performance as the first step of the pipeline. Our methods performed better than Baseline A, especially in TMB loss (0.047

Table 7.5: Mean precision, recall, F1-scores, AUC, and TMB loss over 10CV of the two ML algorithms with the best feature combinations (found in the prior 3CV grid searches) and Baseline A in bot filtering. Check marks (✓) denote the state of the corresponding feature that is enabled. The numbers in parentheses mean the standard deviation. (Other tables follow these notations.)

Algorithms	**Features**				**Noise and TMB**				**Bot**				**TMB loss**
	`app`	`f/f`	`hashtag`	`url`	**P**	**R**	**F**	**AUC**	**P**	**R**	**F**	**AUC**	
MaxEnt	✓	–	–	✓	0.983 (0.007)	0.982 (0.006)	0.982 (0.003)	0.995 (0.005)	0.977 (0.008)	0.978 (0.009)	0.978 (0.004)	0.993 (0.003)	0.047 (0.031)
SVM	✓	✓	✓	✓	0.982 (0.009)	0.985 (0.006)	0.984 (0.005)	0.993 (0.004)	0.981 (0.007)	0.977 (0.011)	0.979 (0.006)	0.992 (0.003)	0.034 (0.028)
Baseline A	Rule-based (application names)				0.971	0.937	0.954	–	0.923	0.965	0.944	–	0.073

Table 7.6: Precision, recall, F1-scores, AUC, and TMB loss of the same two classifiers as in Table 7.5 and the two baselines, on the subset data for Baseline B (Botometer). The scores of our two classifiers are mean values over 5CV.

Algorithms	**Features**				**Noise and TMB**				**Bot**				**TMB loss**
	`app`	`f/f`	`hashtag`	`url`	**P**	**R**	**F**	**AUC**	**P**	**R**	**F**	**AUC**	
MaxEnt	✓	–	–	✓	0.984 (0.006)	0.976 (0.005)	0.980 (0.002)	0.994 (0.004)	0.956 (0.008)	0.970 (0.011)	0.963 (0.004)	0.987 (0.004)	0.047 (0.023)
SVM	✓	✓	✓	✓	0.983 (0.006)	0.980 (0.008)	0.981 (0.003)	0.991 (0.005)	0.963 (0.014)	0.969 (0.012)	0.966 (0.006)	0.983 (0.004)	0.038 (0.024)
Baseline A	Rule-based (application names)				0.983	0.923	0.952	–	0.872	0.971	0.919	–	0.082
Baseline B	Botometer				0.819	0.786	0.802	0.801	0.629	0.676	0.652	0.685	0.326

Table 7.7: Mean precision, recall, F1-scores, and AUC over 10CV of the two ML algorithms with the best feature combinations (found in the prior 3CV grid searches) and the four baselines in noise reduction.

Algorithms	**Features**					**TMB**			
	`app`	`url`	`distinct`	`bib`	`titleness`	**P**	**R**	**F**	**AUC**
MaxEnt	–	✓	✓	✓	✓	0.806 (0.061)	0.716 (0.053)	0.757 (0.047)	0.819 (0.045)
SVM	✓	–	✓	✓	–	0.898 (0.033)	0.662 (0.049)	0.761 (0.038)	0.821 (0.051)
Baseline C	Rule-based (top frequent words in TMBs)					0.495 (0.058)	0.569 (0.066)	0.527 (0.048)	–
Baseline D	SVM with TFIDF-weighted Ngrams					0.660 (0.032)	0.680 (0.072)	0.669 (0.048)	0.722 (0.045)
Baseline E	Baseline D + title					0.851 (0.042)	0.631 (0.065)	0.723 (0.051)	0.806 (0.053)
Baseline F	✓	✓	Ngrams	✓	–	0.831 (0.053)	0.702 (0.067)	0.760 (0.050)	0.835 (0.044)

and 0.034 ≤ 0.073). These results implies our bot filters can cover more bot tweet instances and avoid rejecting fewer TMBs than the rule-based filtering.

The comparison with Botometer (Baseline B) is presented in Table 7.6. We found our method outperformed the state-of-the-art bot account detector (e.g. 0.963 and 0.966 ≥ 0.652 in F1-score). The app feature, or Twitter client application names, seems effective to detect bot tweets, as Baseline A is superior to Botometer which does not use such features.

Noise Reduction The results of noise reduction are shown in Table 7.7. The hyper parameters of the best classifiers were: $C = 10.0$ and L1 norm for MaxEnt; $C = 100.0$ and $\gamma = 0.01$ for SVM. Again, the best combination of features were different in the two algorithms. The `distinct` and `bib` features (options for the *word tokens* feature) were shared. We will discuss the detailed effects of features in Section 7.5.2. Unlike bot filtering, the two algorithms worked differently; the balance between precision and recall was more balanced in MaxEnt than SVM which achieved a high precision score with a relatively lower recall value.

We can see that our noise reduction classifiers outperformed both baselines. In particular, the lower scores of Baseline D means that our features can capture the difference between TMBs and noise tweets better than general-purpose tweet classifiers. We can see Baseline F, which use Ngram instead of our `distinct` feature and all other features, worked comparable to our methods. This means that the `distinct` feature can perform successfully well with less features than those produced by Ngrams.

Experiment 2

From the results of Experiment 1, we built four combinations of two-step pipelines that consist of the best bot filtering followed by the best noise reduction. The hyper parameters and features are the same as reported in Experiment 1 for each algorithm. Table 7.8 shows the classification scores of our pipelines and baselines. All of our pipelines outperformed all of the baselines, with a performance of at least 0.74 by ours vs. 0.72 by the best baselines (F' and G) The scores for the TMB class are almost identical to the stand-alone noise reduction because the performance of bot filtering is very high. The best combination was SVM bot filtering and SVM noise reduction, achieving 0.76 F1-score in the TMB class. The F1-scores of the similar tasks (Section 7.2) were distributed from 0.50 to 0.77 against simpler data sets than ours. We can conclude that our method performed comparably to the methods that achieved the highest performance in related tasks. In other words, our final score can be regarded as the state-of-the-art, and we consider that our pipeline is ready for practical use.

Table 7.8: Mean precision, recall, F1-scores, and TMB loss over 10CV of two-step pipelines consisting of the best classifiers (Tables 7.5 and 7.7) and the four baselines in the TMB identification task.

Classification steps		Bot			TMB loss	Noise			TMB		
Bot filtering	**Noise reduction**	**P**	**R**	**F**		**P**	**R**	**F**	**P**	**R**	**F**
MaxEnt	**MaxEnt**	0.979 (0.010)	0.977 (0.010)	0.978 (0.007)	0.047 (0.038)	0.961 (0.010)	0.973 (0.009)	0.967 (0.007)	0.803 (0.064)	0.689 (0.067)	0.740 (0.058)
MaxEnt	**SVM**	0.979 (0.010)	0.977 (0.010)	0.978 (0.007)	0.047 (0.038)	0.958 (0.012)	0.981 (0.006)	0.969 (0.007)	0.898 (0.042)	0.649 (0.082)	0.751 (0.063)
SVM	**MaxEnt**	0.983 (0.008)	0.976 (0.009)	0.979 (0.006)	0.033 (0.030)	0.960 (0.009)	0.976 (0.008)	0.968 (0.006)	0.803 (0.063)	0.700 (0.072)	0.747 (0.060)
SVM	**SVM**	0.983 (0.008)	0.976 (0.009)	0.979 (0.006)	0.033 (0.030)	0.957 (0.011)	0.983 (0.005)	0.970 (0.006)	0.896 (0.042)	0.656 (0.083)	0.755 (0.064)
Baseline D' (SVM w/ Ngrams)		0.683 (0.013)	0.916 (0.011)	0.782 (0.010)	–	0.886 (0.017)	0.651 (0.020)	0.751 (0.017)	0.779 (0.039)	0.536 (0.040)	0.634 (0.031)
Baseline E' (SVM w/ Ngrams + title)		0.700 (0.013)	0.907 (0.013)	0.790 (0.010)	–	0.886 (0.016)	0.679 (0.022)	0.769 (0.017)	0.792 (0.041)	0.613 (0.038)	0.691 (0.036)
Baseline F' (SVM w/ Ngrams + all noise features)		0.942 (0.011)	0.911 (0.016)	0.926 (0.009)	–	0.907 (0.011)	0.954 (0.010)	0.930 (0.007)	0.874 (0.070)	0.609 (0.053)	0.716 (0.047)
Baseline G (SVM w/ all spam + noise features)		0.964 (0.006)	0.943 (0.013)	0.953 (0.009)	–	0.929 (0.009)	0.970 (0.007)	0.949 (0.007)	0.898 (0.055)	0.600 (0.076)	0.716 (0.059)

7.5 Analysis and diagnosis

In this section, we analyse and diagnose the performance of our method from the following points of view: the relationship between precision and recall, the effects of each feature, prediction errors, and the relationship between scores and data size.

7.5.1 Precision and recall

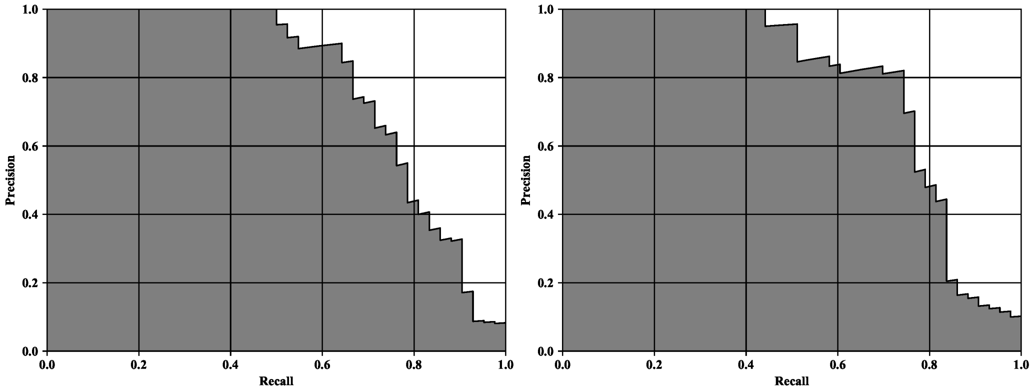
Our classifiers output predicted probabilities of each class per input tweet. When we change the threshold to determine the predicted class, we obtain different results. Figure 7.2 shows the relationship of prediction results produced by our two-step pipelines in the precision-recall curve. The prediction results were generated by 10-fold shuffled and stratified CV of the pipelines, and the CV rounds that produced the median AUC value were chosen for drawing the curves.

From the figure, where the curves form a *shoulder* towards the upper right corner, we can see that taking high recall does not overly spoil precision in our pipeline.

7.5.2 Feature effects

In order to evaluate the detailed effects of each feature, we investigate the results of all the 3CV results executed in Experiment 1 from the following perspectives.

Single Effects: The case where only the baseline feature was used is compared to cases with one additional feature enabled. The baseline feature for bot filtering is app, while that for noise reduction is *word tokens* with all options disabled.



(a) MaxEnt (AUC = 0.78)

(b) SVM (AUC = 0.78)

Figure 7.2: The Precision-Recall curves of the classification results with the median AUC values in 10CV experiments

- **Multiple Effects:** Differences of scores are calculated in a controlled manner. That is, cases where a targeted feature was enabled are compared to cases in which it was disabled, with the condition of all the other features remaining the same.
- **Top Five Combinations:** The top five combinations of features are observed in order to grasp the high-performance feature combinations.

In all perspectives, F1-scores are used as the standard metric. For each combination of features, many combinations of hyper parameters were also applied in grid searches. We chose the best-performing parameters for each feature combination.

Bot filtering

Single Effects The F1-score of the baseline feature, or `app`, was 0.972889 both in MaxEnt and in SVM. From the result shown in Table 7.9, the proposed features seem to make almost no contribution to either algorithm in terms of the *absolute* value of difference from the baseline feature. For a fair comparison, we should take into account the fact that the baseline feature already achieves a high F1-score. In *relative* terms, `hashtag` and `url` were seen as effective. They both are extracted from the tweet text, whereas `f/f` is a user-wise metadata feature. They may perform better than `f/f` in bot *tweet* detection.

Multiple Effects Results are shown in Table 7.10. We can see that (i) adding one feature to `app` improves performance (except `f/f`) and that (ii) two or three features in addition to `app` cause inconsistent effects to different algorithms. Our features for bot filtering may be related in a complex manner.

Table 7.9: The performance contribution (ΔF1-score) of each single feature in comparison to the baseline feature in bot filtering.

Enabled feature	**ΔF1-score**	
	MaxEnt	SVM
`f/f`	.000000	.000000
`hashtag`	.000491	.002553
`url`	.002031	.002943

Table 7.10: Effects of multiple features in bot filtering. ΔF1-score is the difference in the F1-score when a feature is enabled with the combination of other features fixed.

+f/f		**ΔF1-score**		**+hash**		**ΔF1-score**		**+url**		**ΔF1-score**	
`hashtag`	`url`	MaxEnt	SVM	`f/f`	`url`	MaxEnt	SVM	`f/f`	`hashtag`	MaxEnt	SVM
✓	✓	−.000315	.001279	✓	✓	−.000422	.000535	✓	✓	−.000219	.000926
✓	–	.001024	−.000211	✓	–	.001515	.002342	✓	–	.001717	.002732
–	✓	−.000314	−.000211	–	✓	−.000420	−.000955	–	✓	.001120	−.000564
–	–	.000000	.000000	–	–	.000491	.002553	–	–	.002031	.002943

Top Five Combinations The top five best cases shown in Table 7.11 performed very similarly, but we may say that `url` made a relatively higher contribution to both algorithms, which reflects the difference in the presence of URLs between bot tweets and human tweets.

Noise reduction

Single Effects The baseline performance of noise reduction classifiers, which use the *word tokens* feature only, was 0.663266 in MaxEnt and 0.678787 in SVM. Table 7.12 provides the results of both algorithms. In both algorithms, the most effective feature when used alone was `distinct`, and `bib` also contributed substantially to the performance. For MaxEnt,

Table 7.11: The five best-performing settings of bot filtering

(a) MaxEnt

`f/f`	`hashtag`	`url`	F1-score
–	–	✓	.974920
✓	–	✓	.974606
–	✓	✓	.974499
✓	✓	–	.974404
✓	✓	✓	.974184
–	–	–	.972889

(b) SVM

`f/f`	`hashtag`	`url`	F1-score
✓	✓	✓	.976156
–	–	✓	.975832
✓	–	✓	.975621
–	✓	–	.975441
✓	✓	–	.975230
–	–	–	.972889

Table 7.12: The performance contribution (ΔF1-score) of each single feature in comparison to the baseline feature in noise reduction.

Enabled feature	**ΔF1-score**	
	MaxEnt	SVM
`app`	.018073	−.010102
`url`	.023740	−.001474
`distinct`	.058880	.048740
`bib`	.039859	.029970
`titleness`	.008545	−.002943

other features improved the scores by around 1 or 2 points, but for SVM, they had negative effects. The plain *word tokens* feature seems to provide dominant information for SVM to identify TMBs.

Multiple Effects Table 7.13 shows the results of multiple effects on noise reduction. One notable finding is that `distinct` and `bib` always contributed to the performance in both algorithms, which means they are key features for noise reduction. The following features brought different results to different algorithms:

(a) Adding `url` or `titleness` to other features contributed to MaxEnt but not to SVM.

(b) Combining `app` with other features did not work in MaxEnt but did in SVM.

Top Five Combinations Table 7.14 lists the top five cases for the two algorithms. In both algorithms, the classifier with all features enabled performed very close to the best setting. Besides, combinations of three or more features contributed to the higher performance. We conclude that noise reduction seems to require several types of information.

7.5.3 Error analysis

Bot filtering

For bot filtering, we investigate TMBs that were mistakenly labelled bot tweets. They were all tweets that promoted the tweet authors' own book reviews posted to other media or ones about newly published books by the tweet authors. They were tweeted via minor applications of tweet automation. This case can be handled with a list of such applications as a positive (white) list obtained by surveying SRSs and book publishing services.

Table 7.13: Effects of multiple feature in noise reduction. ΔF1-score means differences in the F1-score when a feature is enabled with the combination of other features fixed.

+app				**ΔF1-score**	
url	distinct	bib	titleness	MaxEnt	SVM
✓	✓	✓	✓	−.003507	.004284
✓	✓	✓	–	.005453	−.005938
✓	✓	–	✓	.000366	.002803
✓	✓	–	–	−.004668	.001162
✓	–	✓	✓	.003040	.002128
✓	–	✓	–	.003757	.006620
✓	–	–	✓	.008191	.000413
✓	–	–	–	.001001	.001019
–	✓	✓	✓	−.001872	.005817
–	✓	✓	–	−.005793	.010034
–	✓	–	✓	−.005935	.013570
–	✓	–	–	−.005198	.006803
–	–	✓	✓	−.003644	.005056
–	–	✓	–	.010263	.004557
–	–	–	✓	.014363	.005871
–	–	–	–	.018079	−.010102

+url				**ΔF1-score**	
app	distinct	bib	titleness	MaxEnt	SVM
✓	✓	✓	✓	.007186	.006154
✓	✓	✓	–	.011777	−.007267
✓	✓	–	✓	.009484	−.002869
✓	✓	–	–	.009716	.000200
✓	–	✓	✓	.006531	−.004015
✓	–	✓	–	.000983	−.000642
✓	–	–	✓	.011423	−.001479
✓	–	–	–	.006663	.009648
–	✓	✓	✓	.008822	.007686
–	✓	✓	–	.000530	.008705
–	✓	–	✓	.003183	.007898
–	✓	–	–	.009186	.005842
–	–	✓	✓	−.000153	−.001087
–	–	✓	–	.007489	−.002705
–	–	–	✓	.017595	.003978
–	–	–	–	.023740	−.001474

+distinct				**ΔF1-score**	
app	url	bib	titleness	MaxEnt	SVM
✓	✓	✓	✓	.021102	.027175
✓	✓	✓	–	.025628	.025994
✓	✓	–	✓	.029292	.057528
✓	✓	–	–	.038657	.056197
✓	–	✓	✓	.020447	.017005
✓	–	✓	–	.014834	.032619
✓	–	–	✓	.031231	.058918
✓	–	–	–	.035604	.065645
–	✓	✓	✓	.027649	.025019
–	✓	✓	–	.023932	.038552
–	✓	–	✓	.037117	.055139
–	✓	–	–	.044327	.056055
–	–	✓	✓	.018675	.016245
–	–	✓	–	.030891	.027142
–	–	–	✓	.051529	.051219
–	–	–	–	.058880	.048740

+bib				**ΔF1-score**	
app	url	distinct	titleness	MaxEnt	SVM
✓	✓	✓	✓	.013924	.009976
✓	✓	✓	–	.013335	.004136
✓	✓	–	✓	.022113	.040330
✓	✓	–	–	.026364	.034339
✓	–	✓	✓	.016221	.000953
✓	–	✓	–	.011274	.011604
✓	–	–	✓	.027005	.042866
✓	–	–	–	.032044	.044629
–	✓	✓	✓	.017797	.008495
–	✓	✓	–	.003214	.011236
–	✓	–	✓	.027264	.038615
–	✓	–	–	.023608	.028738
–	–	✓	✓	.012158	.008707
–	–	✓	–	.011870	.008373
–	–	–	✓	.045012	.043681
–	–	–	–	.039859	.029970

+titleness				**ΔF1-score**	
app	url	distinct	bib	MaxEnt	SVM
✓	✓	✓	✓	.000813	.006430
✓	✓	✓	–	.000224	.000589
✓	✓	–	✓	.005339	.005249
✓	✓	–	–	.009589	−.000741
✓	–	✓	✓	.005403	−.006992
✓	–	✓	–	.000456	.003658
✓	–	–	✓	−.000209	.008622
✓	–	–	–	.004829	.010385
–	✓	✓	✓	.009773	−.003793
–	✓	✓	–	−.004809	−.001052
–	✓	–	✓	.006056	.009740
–	✓	–	–	.002400	−.000136
–	–	✓	✓	.001482	−.002774
–	–	✓	–	.001193	−.003109
–	–	–	✓	.013698	.008123
–	–	–	–	.008545	−.005588

Table 7.14: The five best-performing settings of noise reduction

(a) MaxEnt

app	url	distinct	bib	titleness	F1-score
–	✓	✓	✓	✓	.744320
✓	✓	✓	✓	✓	.740812
✓	✓	✓	✓	–	.739999
–	–	✓	✓	✓	.735498
–	✓	✓	✓	–	.734546
–	–	–	–	–	.663266

(b) SVM

app	url	distinct	bib	titleness	F1-score
✓	–	✓	✓	–	.745933
✓	✓	✓	✓	✓	.745095
–	✓	✓	✓	–	.744604
–	✓	✓	✓	✓	.740811
✓	–	✓	✓	✓	.738941
–	–	–	–	–	.678787

Noise reduction

We investigated false positives and false negatives by observing the text in the tweet body because they were the main features of our noise reduction.

False Positives A noticeable group of false positives are tweets that mention titles of songs, films, and TV programs. These titles were often the same strings as book titles, and put inside Japanese quotation marks (「」 or 『』). Although our `distinct` feature was partly intended to identify quotation marks around book titles, the classifier seemed to fit overly to the marks. Because these tweets contained book-unrelated words such as '歌詞 (lyrics)' and '上映 (showing a film)', we expect that this issue can be fixed by increasing training data or creating negative word lists. Larger training data sets can generate more difference in the word distribution between TMBs and noise tweets, so that ML models can learn book-related/unrelated words more precisely. Book-unrelated words (like 'lyrics') can be manually collected by considering non-book domains found in type-(b) noise (Section 7.3.2). For example, we can use a collection of film-related words (e.g. extracted from film domain corpora) to add a binary feature which tells the existence of such words in tweet texts.

Other than these, a very small portion of bot tweets (1 tweet in 4/10 CV rounds) that were not rejected in bot filtering were labelled TMB. These tweets were posted from the

Twitter site and were about book promotion. Their rarity indicates that we can safely ignore them.

False Negatives Over half of the false negatives per CV round mentioned books without quotation marks around the title. This was the opposite of the first case of false positives above. This case also seems to come from over-fitting to the marks, as these false negatives contained book-related words such as '本屋 (book store)', '連載 (serial)', and '小説 (novel)'. Whereas these words were less frequent in our data set, they are likely to appear more if tweets mention books. Hence, enlarging training data and augmenting positive word lists will be effective in solving these cases.

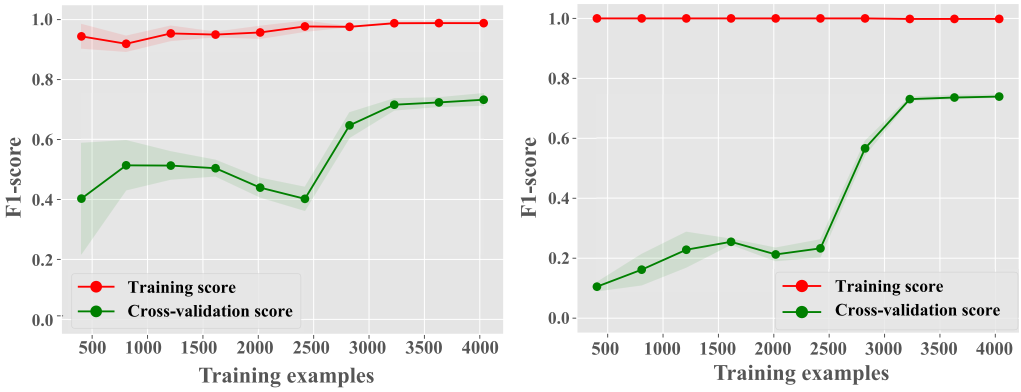
In addition, the TMBs that take the form of replies to other users tend to be mislabelled. The cause seemed to be that many noise tweets are conversations, and our classifier overly attributed reply tokens as the cue of noise tweets. Reducing the weight of reply tokens in training stages of the classifier is a possible work-around.

Summary Both of the false positives and negatives we observed were related to sparseness of vocabulary. The size of our TMB data set seems not sufficient to cover all possible variations of ways to mention books. Increasing the data size will thus improve the performance of our method without modification. We will confirm this point in Section 7.5.4.

Another solution may be to take into account the sense of words, phrases, or sentences. For instance, we can create lists of words related/unrelated to reading and increase/decrease their weight accordingly. Such lists can be created by corpus analyses, word similarity databases, or pre-trained word embeddings. Furthermore, NN-based models that suit embedded representation of texts may also be adapted for this task. For example, character-level NN models for document classification (e.g. X. Zhang, Zhao and LeCun, 2015) are applicable because many book-related but infrequent Japanese words consist of book-related Chinese characters. Besides, the performance of a NN model for a task can be improved by simultaneously learning other related tasks (multi-task learning; e.g. Crichton et al., 2017; P. Liu, Qiu and X. Huang, 2016). As mentioned in Section 7.3.3, however, we are aware that NN-based models may need very large data sets, and render a detailed feature analyses difficult in general.

7.5.4 Scores and data size

Our error analysis suggested that more data could increase the performance of noise reduction. This is confirmed by the learning curves, which show fluctuation of training scores and test scores of our classifiers in relation to the size of the data.



(a) MaxEnt **(b)** SVM

Figure 7.3: The bias-variance tests on noise reduction of the best classifiers

If both training scores and test scores converge to a low score, the model cannot learn significant information from data and the model is under-fitting (high bias). If training scores remain high and test scores increase as data size grows, the model is over-fitting (high variance) and data augmentation will contribute to improving the model.

Figure 7.3 presents F1-scores on both training and test according to the data increase. We can see that our classifiers currently over-fit the training data. In other words, they can be improved by increasing the size of training data.

7.6 Conclusions

7.6.1 Summary

In this chapter, we tackled the task of identifying Japanese tweets that mention books (TMBs). In order to make this task feasible, we focused on identifying *manually generated Japanese tweets that mention specific books by their titles.* Starting from tweets that contain title strings (TCTSs), we re-defined this task as the classification of TCTSs into TMBs, bot tweets, and noise tweets. To solve this multi-class classification task, we proposed a two-step pipeline consisting of bot filtering followed by noise reduction. That is, we first filtered out automated tweets about books that are not so useful for possible real-world applications of TMBs, and then removed noise TCTSs that do not mention any books. Our pipeline is based on the different characteristics among the three classes. Bot tweets are characterised by shallow metadata in contrast to TMBs and noise, whereas the linguistic content of the tweet is required to distinguish TMBs from noise.

Our evaluation experiments showed that both steps of the pipeline outperformed the rule-based baselines, and that our two-step architecture also outperformed the one-step ('one-versus-the-rest') baseline. Furthermore, our method achieved performance comparable to the highest-scored methods among related tasks (0.76 F1-score).

We analysed our method in detail from four perspectives: the precision-recall performance, feature effects, error samples, and the training data size. The proposed method maintained the balance between precision and recall. The feature analysis showed the most effective features as follows: in bot filtering, application names (`app`); in noise reduction, considering the positions of words in relation to book title strings (`distinct`). Through observing error samples, we believe that the performance can be further improved by taking into account word senses as features. Our method also has room for improvement by increasing the training data.

7.6.2 Limitations

One limitation in this study is the range of the target TMBs.[19] Our target was TMBs using book titles, and we used a comprehensive book title list to retrieve them, thus starting from TCTSs. That is, TMBs excluded from TCTSs were not covered by our framework. There are three types of such TMBs: (a) TMBs using abbreviated or incorrect titles, (b) TMBs using bibliographic information other than titles, and (c) TMBs without any formal bibliographic information.

We did not cover the type (a) because the title list to collect TCTSs include full book titles only. However, abbreviated or wrong book titles are often used in TMBs, since social media texts are basically informal. We may address this problem with creating another list of titles, or by applying rules to obtain abbreviated or incorrect book titles, although the way to do this is not obvious.

Possible examples of the type (b) are 'the first novel of Dan Simmons' or "Iwanami's book I purchased yesterday". They are not negligible, although we reasonably assumed TMBs using book titles are the majority. If phrases such as the above are parsed appropriately, we can take them into account in combination with bibliographic databases or users' past tweets.

In the real Twitter stream, we sometimes observed the type-(c): TMBs that state or just allude to the content of specific books, e.g. 'Do you know how "Cogito, ergo sum" was originally said? You must refer to the context'. This mock example uses the exact quote to specify René Descartes' *Principles of Philosophy*. If we take into account the content of

[19] In this section, the term *TMBs* is mentioned in its general meaning.

books (e.g. full texts or information from book reviews), we might be able to handle this type of TMBs.

Finally, this study focused on Japanese tweets. Although our proposed method and features are language independent and thus can be applicable to other languages with some adaptations, the performance can differ from language to language, as there may well be differences in a range of factors such as conventions in making book titles and circumstances of online book services.

7.6.3 Future directions

We plan to improve our method in three ways. First, we will take better word sense representations of TMBs into consideration. In order to process book-related and book-unrelated words on a larger scale while maintaining accuracy, we will consider incorporating lexical databases and pre-trained word embeddings into our method.

Second, we are going to extend the range of the target TMBs, looking at how to handle a part of the uncovered TMBs just introduced. For type (a), if seed samples of abbreviated or mistaken titles are available, rules to produce such titles may be extracted or learned. Pre-defined linguistic patterns can collect some of the type-(b) TMBs and even identify the mentioned books. TMBs of the type (c) are the hardest. Some of these implicit TMBs are often one in a sequence or a discourse of TMBs by the same Twitter user. For instance, a tweet stating a review of a book without clarifying the book can be posted right after a short TMB with the book title, which is our target TMBs in this study. That is, our method can contribute to identify them.

Third, in order to augment the data size, we will implement a *human-in-the-loop* architecture (Holzinger, 2016) for our task. This will involve developing a user interface for human users to give feedback to correct mis-labelling produced by our method. Since our task is derived from the development of the system of an online surrogate to book exposure, this architecture can be easily incorporated into the overall system by implementing a user feedback channel.

8 iTMB scoring

In this chapter, we define and solve the tasks for the NLP modules of the iTMB scoring step. We have two modules: the pleasantness scorer and the considerateness scorer. For these tasks, we use the TMB inspiringness dataset (Section 5.2.2; i.e. dataset L in Chapter 6). We devote one section for each task, in which we formulate the task, propose a baseline method, carry on experiments, and finally summarise the outcomes and possible future directions.

8.1 Pleasantness scorer

The pleasantness scorer aims to understand pleasantness of TMBs, i.e. TMB authors' attitudes towards mentioned books. In Chapter 4, we confirmed that TMBs with positive pleasantness have better inspiringness than neutral or negative ones, as we hypothesised when we defined pleasantness (Section 3.1.3).

8.1.1 Task definition

As we already formalised in Chapters 5 and 6, the attitudes correspond to *opinions* in opinion mining: one's subjective evaluation towards a certain entity in a message, where the entity can be a product, service, person, group, event, or topic (B. Liu and L. Zhang, 2012). Pleasantness in a TMB is defined as the opinion in the TMB towards the mentioned book. In particular, since a TMB can mention several books, the actual data format is TMB records, i.e. pairs of one TMB and one of its mentioned books. For each TMB record, thus, the pleasantness scorer module aims to measure pleasantness.

While the definition of the polarity, valence, or orientation of pleasantness can be chosen from several options including discrete classes and continuous values in a range, we use the standard ternary schemes: positive, negative, and neutral. The reason why we reduce class sizes from Section 5.3 is that more complex classes, e.g. incorporating the balance of positive and negative opinions we did for inspiringness annotation, were very few (see Section 6.4.3). Henceforth, we refer to this task form as the *pleasantness classification task*.

Following the pipeline architecture of the exposure system (Section 3.2.1), the input data for this module is TMBs output from the TMB identifier (Chapter 7). For the pleasantness classification task, we assume that the all input consists of TMBs, or that no non-TMBs are contained.

8.1.2 Methods

We have just formulated pleasantness scoring into the pleasantness classification task, which technically forms a standard opinion-mining task where messages are classified into positive, negative, and neutral. Opinion mining (or sentiment analysis) is a popular research field in NLP and is included in benchmark datasets for natural language understanding (e.g. SST-2 in GLUE(A. Wang et al., 2018)). Since such a standardised scheme is applied to various task setups including Twitter (e.g. R. Wang et al., 2019), we know state-of-the-art methods robust to different tasks.[1] Since most high-performing methods are based on supervised machine learning, the rest matter we have to deal with is crafting high-quality annotated data, which we have already achieved in Chapter 5.

We apply BERT to the pleasantness classification task, which achieved the state-of-the-art performance in various tasks of natural language understanding (Devlin et al., 2019). While class imbalance is attributed to our data Section 6.4.3, some studies suggest that BERT performs relatively robust to unbalanced data too (Tayyar Madabushi, Kochkina and Castelle, 2019; Wei and Zou, 2019). That high performance comes from the language-model pre-training on large corpora and pre-trained BERT models for Japanese language are available.[2] Since pre-training on relevant domains results in better performance (J. Lee, W. Yoon et al., 2019), we use a model pre-trained on Japanese Twitter corpora (Takeshi, Sakae and Naoyuki, 2019).

In contrast to TMB identification, we do not examine different algorithms. Whereas the TMB identification task intrinsically needs external knowledge about books (e.g. book titles), pleasantness classification involves language-intrinsic features. That is, pleasantness can be inferred less dependently on entity types, and existing general methods are expected to be applicable to the classification task. For this setup, compiling high quality data is crucial while algorithms can be interchangeable based on the progress of the state-of-the-art in the corresponding task (i.e. opinion mining). Tasks like TMB identifica-

[1] Websites that track state-of-the-art methods are available, for instance: PapersWithCode (https://www.paperswithcode.com/task/sentiment-analysis) and GLUE benchmark (https://gluebenchmark.com/leaderboard).

[2] While improved model architectures based on BERT are being proposed in a rapid pace (X. Liu et al., 2019; Z. Yang et al., 2019), BERT's pre-trained models have better availability.

tion, on the other hand, always require domain adaptation, which means that features and algorithms for other entities cannot necessarily be adopted as is, to the interested entity.

Note that, although we adopt the state-of-the-art method as is, we still need to train a model for our task due to the lack of publicly available opinion-mining models for Japanese TMB data. See also the reason why we did not apply existing opinion mining models to bare TMB data, which mentioned in Section 6.2.1.

8.1.3 Experiments

We used the TMB inspiringness dataset and aggregated 'Positive > Negative' to 'Positive', 'Positive < Negative' to 'Negative', and 'Positive = Negative' to 'Neutral'. To generalise the information of book titles mentioned in TMB records, we replace the target book title in the text with "タイトル (title)", instead of leaving the book title string as is. This also provide the BERT model with information about the positions of book titles in TMBs.[3]

As described in Section 8.1.2, we applied the $\text{BERT}_{\text{BASE}}$ uncased model pre-trained on Japanese Twitter corpora (Takeshi, Sakae and Naoyuki, 2019). We follow the 10-fold cross-validation manner for the experiment: for each round, we used the three splits of the dataset with the ratio of 8:1:1; the model was fine-tuned with 80% of the dataset and validated on 10% of the dataset; then, the model is applied to 10% of the dataset for the prediction; the prediction splits over the 10 rounds have no overlaps.

The BERT's fine-tuning for pleasantness classification, we added a fully connected layer on top of the BERT model,[4] which outputs a predicted distribution of the three classes by the softmax function from a sentence-embedding token produced by the BERT model for each TMB record. The fine-tuning process jointly optimises the weights inside the BERT model and the top classification layer.

We set hyper parameters as follows:

- input text length = 80
- batch size = 32
- learning rate = 2.0×10^{-6}
- number of epochs = 6.

The input text length roughly corresponds to the word count in tweets. Then we found the largest batch size for our computing resource,[5] which was 32. This is because the larger batch size results in shorter training time and we preferred a short training time to

[3] Given that BERT can take up to two input sentences simultaneously, we also tried feeding the pair of the raw TMB text and the string of the mentioned book title, for each TMB record. In our preliminary experiment, we found that this task setup performed slightly worse than the current one.

[4] This is the standard setup for text classification tasks suggested by the original paper (Devlin et al., 2019).

[5] We used an NVIDIA GeForce GTX 1080 (8GB memories) GPU.

Table 8.1: Values of median over 10 fold cross validation of the pleasantness classification task. Parenthesised values provide median absolute deviation (MAD).

(a) BERT

	Precision	**Recall**	**F1-score**
Positive	0.807 (0.004)	0.892 (0.008)	0.846 (0.007)
Neutral	0.756 (0.024)	0.665 (0.018)	0.702 (0.012)
Negative	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)

(b) Random prediction based on the class distribution

	Precision	**Recall**	**F1-score**
Positive	0.623 (0.010)	0.620 (0.012)	0.626 (0.010)
Neutral	0.344 (0.018)	0.339 (0.013)	0.337 (0.012)
Negative	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)

conduct 10-fold cross validation. Finally, we explored the combination of the learning rate and the number of epochs, considering the balance between the convergence speed and the performance. According to S.L. Smith et al. (2018), the batch size and the learning rate maintains a trade-off relation for the final score performance under the same number of epochs.

Table 8.1 shows the result of the experiment in median and median absolute deviation (MAD) values over 10-fold cross validation. It also brings, for a reference, the result of 'random prediction' where a dummy classifier randomly assigns a label based on the class distribution of the training split. The comparison between these results indicates that our BERT model predicted pleasantness utilising linguistic characteristics in text. In particular, the 0.7–0.8 F1-scores for the 'Positive' and 'Negative' classes are regarded as a high performance, in comparison to similar opinion-mining tasks (R. Wang et al., 2019; Zimbra et al., 2018). However, the model was not able to detect the 'Negative' class, which is probably because of the extreme class imbalance (around 2.2%).

8.1.4 Conclusion

Considering the importance of 'Positive' class in pleasantness classification, our model achieved a practical-level performance. It can identify 'Positive' samples from the close amount of 'Neutral' samples with high precision and recall. Although it cannot label 'Negative' samples correctly on contrary to suggested robustness to skewed data, the contamination ratio of them in the predicted 'Positive' samples almost keeps the original probability of its appearance, as shown in Table 8.2.

However, since 'Negative' pleasantness strongly decreases inspiringness as we saw in Chapter 4, the next step should be to improve the performance of detecting 'Negative' samples. Negative pleasantness in TMBs appear infrequently, which causes a class imbal-

Table 8.2: Ratio of true 'Negative' in the predicted 'Positive' samples by BERT

Median	0.025197
MAD	0.006687

ance problem. Common countermeasures to this issue in machine learning are divided largely into two types: data re-sampling and imbalance-aware modification of algorithms (Johnson and Khoshgoftaar, 2019). The former has two directions, i.e increasing samples of small classes (over sampling) and decreasing samples of large classes (under sampling). Recently, over-sampling studies (or data augmentation) have been showing promising results for NLP tasks (Kobayashi, 2018; Sennrich, Haddow and Birch, 2016), but some papers argue that some methods do not improve BERT's performance very well (Tayyar Madabushi, Kochkina and Castelle, 2019; Wei and Zou, 2019). Also because the effect of data augmentation varies among different tasks (Johnson and Khoshgoftaar, 2019), this field still remains to be investigated for practical applications.

8.2 Textual considerateness scorer

Textual considerateness scorer, defined in Section 3.2.3, aims to detect considerateness in TMB text. We have been handling textual considerateness as the strength of book recommendation since Section 5.3. As we saw in Chapter 4, a strong or excessive recommendation decrease inspiringness while weak ones can increase it on the other hand. The goal of this module is thus to detect the strength of recommendation phrases in TMBs.

8.2.1 Task definition

We follow the levels we set for the strength of recommendation: none, weak, strong, and excess (see Section 5.3). Textual considerateness scoring can thus be formulated as *recommendation strength classification.* Based on the result of Section 4.3.2, *excess* strongly decreases inspiringness of TMBs with positive pleasantness, whereas *weak* recommendations increase inspiringness most amongst the four levels. Therefore, we can put importance on filtering the excess-level samples and finding the weak-level samples.

We also adopt the same input-data format as for pleasantness classification, i.e. input data is TMB records.

Algorithm 1 Recommendation strength classification

```
 1: $t \leftarrow$ a TMB text string
 2: if $t$ contains any of the phrases unique to strong recommendations then
 3:     classify $t$ as strong
 4: else
 5:     if $t$ contains any of the emphasising patterns to ‘strengthen’ weak phrases then
 6:         if $t$ contains any of the phrases unique to weak recommendations then
 7:             classify $t$ as strong
 8:         else
 9:             classify $t$ as weak
10:         end if
11:     else
12:         classify $t$ as none
13:     end if
14: end if
```

8.2.2 Methods

The strength of recommendation belongs to linguistic characteristics, and its detection could be handled by machine learning models. For TMB identification and pleasantness classification, we had a certain amount of labelled data to apply machine learning. However, our datasets contain less than 500 TMBs that recommends books, among which the strength further distributes in an unbalanced manner. Machine learning methods may over-fit this data, which makes them hard to be applied and evaluated.

In Section 6.5.1, we found that frequent recommendation phrases of weak and strong levels have explicit linguistic patterns. Instead using machine learning, therefore, we build a rule-based classifier for this task. We make use of the following facts of recommendation phrases:

- major patterns in weak recommendations are ‘recommendation’, ‘hope’, and ‘suggestion’
- strong recommendations have largely two types:
  - using unique expressions to the strong level, e.g. ‘order’, and ‘obligation’
  - emphasising the patterns for weak recommendations (i.e. strengthened) with adverbs, adjectives, and exclamation marks

We designed the rules to classify the recommendation strength as shown in Algorithm 1. We implemented this algorithm using a set of regular expressions listed in Table 8.3. At the line 6 of the algorithm, we search for emphasising words only around the expressions for the weak level if they exist. We set the span to search for emphasising words to 10 characters behind and ahead of the appearing weak phrases.

Table 8.3: Regular expressions for weak and strong recommendation phrases

<table>
<tr><th>Phrase type</th><th>Regular expression</th></tr>
<tr><td><b>Weak level</b></td><td></td></tr>
<tr><td>Recommend</td><td>(?:おすすめ|[おオ]ススメ|お[勧薦奨]め)(?:\W|です|かも|の|し|す|本)</td></tr>
<tr><td>Recommend 2</td><td>[勧薦奨]め(?:た[いく]|る)</td></tr>
<tr><td>‘Zehi’ suggest</td><td>(?:ぜひ|是非|ゼヒ)\W</td></tr>
<tr><td>Try</td><td>読んで(?:[ほ欲]しい|[み見]て|(?:貰|もら)いたい|(?:いただ|頂)きたい)</td></tr>
<tr><td>Suggest</td><td>[お御]読み(?:くだ|下)さ</td></tr>
<tr><td><b>Strong level</b></td><td></td></tr>
<tr><td><i>Unique phrases</i></td><td></td></tr>
<tr><td>Must read</td><td>必[読見]</td></tr>
<tr><td>Norm</td><td>読むべ[きし]</td></tr>
<tr><td>Had better</td><td>読ん(?:だ|でおいた|どく)(?:方|ほう)が[い良]い</td></tr>
<tr><td>Plea</td><td>読んで(?:(?:くだ|下)さい|[^、，「「\(（\w])</td></tr>
<tr><td><i>Emphasising words</i></td><td></td></tr>
<tr><td>Emphasis</td><td>断然|超絶?|本当に|ほんと|ホント|まじ|マジ|絶対|とにかく|めっ?ちゃ|メッ?チャ</td></tr>
<tr><td>Exclamation</td><td>[!！　]{2,}</td></tr>
</table>

Table 8.4: Performance of our rule-based method for recommendation strength classification

	Precision	Recall	F1-score
None	0.984	0.982	0.983
Weak	0.680	0.696	0.688
Strong	0.410	0.576	0.479
Excess	0.000	0.000	0.000

Although these rules cannot detect complex or creative phrases, most frequent phrases should be covered with high precision. We do not actively target excess-level phrases either, due to the shortness of their known patterns and to the possibility of their rich linguistic diversity in the real world. Building rules from the current 12 examples may also increase the risk of false positives. We expect that the rules for weak and strong phrases may not overlap possible patterns for the excess level.

8.2.3 Experiments

Performance

Table 8.4 reports precision, recall, and F1-scores of the classification performance for our rule-based classifier applied to the TMB inspiringness dataset. The rules we designed achieved reasonable performance (i.e. close to 0.7 F1-score) in the detection of ‘weak’ recommendations, while the scores for the ‘strong’ level shows difficulty of the detection. As we designed, our rules did not detect any ‘excess’ samples.

Table 8.5: Confusion matrix of our rule-based method for recommendation strength classification

Rule-based method	True labelling			
	None	Weak	Strong	Excess
None	8510	100	22	12
Weak	125	272	3	0
Strong	30	19	34	0
Excess	0	0	0	0

Table 8.5 is a confusion matrix on the output of our method and the true label in the recommendation strength. The highest counts of the rule outputs for each level appear in corresponding levels, which suggests that our rules successfully detect the majority cases of each level. Furthermore, our rules did not catch excess samples at all in any way. Though this resulted in zero scores in the classification performance, our rules will not mistakenly pass excess-level recommendations to the latter pipeline steps of the exposure system. This behaviour is preferable since excess-level recommendations ruin inspiringness of positive TMBs (see Chapter 4).

Error analysis

We conducted error analysis of this classification result. We start from general errors found across different strength levels, and then have closer looks into each strength level. Italics in examples below mean the corresponding English phrases detected by our regular-expression rules.

The missed samples (i.e. weak, strong, or excess TMB records classified as 'none') were complex or creative phrases, which we did not intend to cover by the proposed rules. Such patterns are hard to be defined as rules due to the risk of increasing the number of false detection (e.g. 'I hope you know this trick' for mystery novels; for example phrases, see also Section 6.5.1). The examples of false detection (i.e. non-recommending TMB records but classified as weak or strong recommendations) mostly come from recommending TMBs that mention multiple books, which happened because our rules do not consider the target book of recommendation phrases.

Next, we summarise the errors of weak labels output by our rules. The non-recommending TMB records classified as weak contained the mentions of the fact that other people (not the TMB author) recommended the books, e.g. 'The book *recommended* by `@user` has just arrived to my place!'. Some rare emphasising words like 'without missing any single line' were used in the three strong-strength samples falsely labelled weak.

Finally, we explain the error cases of strong labels classified by our rules. Some of the non-recommending TMB records wrongly marked as strong include one's own regrets, for instance, '(I) *should* have *read* this book before' where the subject 'I' was sometimes omitted in the sentence.[6] In addition, similar to the false pattern in the weak level, the third person's obligation to read certain books were sometimes mentioned, e.g. 'someone said this book should be read'. The false labelling to weak in true strong samples occasionally shares the same recommendation phrases (e.g. 'I hope you will read this book very much!'), which implies that the context can affect the strength.

8.2.4 Conclusion

We tackled the recommendation-strength classification task by designing linguistic rules based on the quantitative analysis of recommendation phrases (Section 6.5.1) and achieved a fairly well performance, which is close to 0.7 F1-score. This initial result has room for improvement: to distinguish which book is the concern of the recommendation phrase and to understand to whom the recommendation phrase is directed. Whereas our rules consider the existence of recommendation phrases in TMBs, utilising syntax or dependency information may contribute to these issues.

Suggestion mining, which aims to identify messages that suggest some ideas or plans in industrial customer reviews or customer support responses, is similar to detection of recommendation phrases. Recently, one shared task for it was run in SemEval 2019 (Negi, Daudert and Buitelaar, 2019), which implies the growing interest on this matter. The setup of the shared task roughly corresponds to the classification between recommending TMBs or not. According to its evaluation, rule-based methods performed surprisingly well: e.g. one rule-based system achieved the top score in a subtask and is also placed at the fifth rank in the other subtask. Although BERT-based ensemble classifiers were popularly used and performed highly well in the shared task, the best BERT-based system gained only 4 points in F1-score from the the best rule-based system. This result supports our decision to adopt a rule-based method.

Another difference from the suggestion-mining shared task is that we aim to identify the levels of the recommendation strength from our extremely skewed data (i.e. the smallest class 'excess' only contains 12 samples). More samples for recommendation phrases are demanded not only for enhancing our rules, but also for applying machine learning methods in future. We do not believe that our samples of the 'excess'-level recommendation represent the diversity of existing expressions. Before applying any bootstrap-like

[6] Japanese text often omits the subject, but even in English, for example, informal casual text allows the omission as well.

methods for data augmentation, i.e. generating or finding similar samples to given labelled data, we still need to collect relevant data for the ‘excess’ expression.

‘Excess’-level recommendations include aggressive expressions, some of which may be related to offensive language. While its detection tasks are also gaining attention (Tuarob and Mitrpanont, 2017; Zampieri et al., 2019), they rely on a certain amount of labelled data. Besides, the definition of offensive language differs significantly among individual tasks or studies (Chakrabarty, K. Gupta and Muresan, 2019; Razavi et al., 2010). To polish the conceptualisation of ‘excess’ in recommendations in relation to these kinds of language use will be one of our next research question.

Part V

Conclusions

9 Summary

In this thesis, we proposed the necessity of a digital surrogate system for passive exposure to books that has traditionally been supported in physical environments, by propagating online social mentions of books. Towards the realisation of such a system, we identified the requirements for inspiring users to read, and designed the feasible architecture of the system. Since the core NLP modules of the system involve novel applied tasks, we primarily focused on defining and solving them, starting from collecting online social mentions of books followed by careful analyses thereof. Through this way, we surely answered the three principal RQs defined in Section 1.7:

- how do we formulate the desirable attributes of the digital surrogate system that exposes users (including infrequent readers) to social mentions of books (i.e. inspiringness)?
- what is the feasible design of the digital surrogate system with inspiringness embedded?
- how can the system modules be implemented?

Here we make sure again the contributions of this thesis, listed in Section 1.7.2.

- based on the current situation of reading environments, we revealed that the necessity of the digital surrogate system for physical passive exposures to books to deal with a potential educational gap in the future
- we identified and confirmed the requirements for inspiring infrequent readers online to read
- we designed the feasible system architecture and formulated novel tasks required to be solved in order to build the system
- we solved the tasks for the core NLP modules of the system with the practical performance, starting from compiling datasets and careful analyses of them, ending with thorough error analyses
- we summarised the future paths for building the UI/UX of the system by showing the tasks and problems remaining to be solved

Note that the last contribution will be made in Chapter 10.

Henceforth, we review our achievements for each chapter so far.

Part I: Background and research questions

Chapter 1: Introduction

We clarified why the book-exposure system is required in this chapter. By citing key literature, we confirmed the importance of reading itself in relation to literacy and pointed out that passive exposure has played an important role. We summarised the current statuses of physical and digital environments in term of books and reading: physical environments as a traditional source of passive exposure to books are decreasing, while popularising digital environments currently do not support it well due to their on-demand nature and personalisation. We believe that online social media can provide unintentional encounters to books even for infrequent readers, but it is known that filter bubbles are formed there. A digital system that propagates online social mentions can surrogate passive exposure to books in physical environments. To this end, we also claimed the necessity of inspiringness to be implemented to the system, i.e. the desirable attributes of exposure to inspire users to read.

This chapter gave an overview and structure of this thesis as well. In addition, we declared the technical scopes of our research: focusing on *Japanese tweets that mention books* for the target type of online social mentions of books.

Chapter 2: Related work

We summarised related work from the following perspectives.

Promoting reading in digital environments: we reviewed existing services and activities of reading promotions in digital environments aware of decreasing unintended encounters with books, and found that they aim to help those who have a strong will to read but find difficulty with choosing books. Our system will support people in the one-step behind of them, i.e. enhancing the desire to read

Book information systems and tasks: we compared the digital surrogate system we proposed with the existing book-information applications like book search engines and book recommendation systems. Significant differences lie on whether provided information is optimised to ‘relevance’ or not, and whether the target users include non-readers or not.

Concepts related to inspiringness: we surveyed close concepts to inspiringness which aims to attract those who have not yet been showing active interests in certain entities, such as affordances, serendipity, attitude change, and WoM effects. We made use of the findings in these fields later in Chapter 3

Part II: Conceptual framework

Chapter 3: Requirements for the digital exposures to books

This chapter answered the following questions:

- what features are required for the digital surrogate system in order to inspire infrequent readers to read?
- how can we design the system architecture with inspiringness embedded?
- what tasks and modules will be solved and implemented in this research?

Based on related work, we identified the four necessary components of inspiringness:

Dailiness the degree to which the digital environments to which the message is delivered are being used by the user

Proximity the closeness from the user to the message author

Pleasantness the joyfulness of the message author towards the mentioned target (i.e. books) perceivable from the message

Considerateness the moderateness of the exposure for the user, or how less forceful the exposure is to the user.

Then, we designed the architecture of the digital surrogate system with these four components embedded. The system consists of three-step pipeline: TMB collection, inspiringness TMB (iTMB) scoring, and iTMB exposure. We proposed two technical modules per step, i.e. six modules in total.

The overall objectives of this thesis comprised two parts: to confirm the necessity of inspiringness components and to implement the *core NLP modules*.

Chapter 4: Validation of the inspiringness components

This chapter confirmed whether the four components of inspiringness do contribute to inspiring infrequent readers. Whereas some of them were able to be shown as effective according to related work, we found that the interaction of pleasantness and considerateness had remained to be clarified. We carried out a psychological experiment and its statistical analyses showed that considerate recommendation messages enhance the desire to read books, but that forceful messages, even if they are positively framed, decrease inspiringness of TMBs.

Part III: TMB corpora

Chapter 5: TMB corpus creation

We created corpora of TMBs in order to investigate their characteristics and to prepare labelled data for development of the core NLP modules. We used two heuristic criteria to collect TMBs: whether book-title strings are contained and whether book-related hashtags are included. We conducted annotation of TMBs and inspiringness, starting from designing a guideline. As a result, nearly 20 thousand labelled tweets was obtained including non-TMBs, with high inter-coder agreements.

Chapter 6: Descriptive analysis of TMB corpora

We analysed our TMB corpora in terms of the purpose of tweeting about books, pleasantness, and textual considerateness. In terms of inspiringness components, the descriptive statistics showed, e.g. the majority of pleasant attitude towards books and the infrequent existence of inconsiderate book-recommending activities in the wild. In addition to this quantitative analysis, we carried out a qualitative analysis of considerateness based on the phrases that recommend books and the targeted audience of recommendations. We found linguistic patterns of casual recommendation in reality, and popular ways to state the target audience.

Part IV: Core NLP modules of the system

Chapter 7: TMB identification

We defined the *TMB identification* task as ternary text classification: TMB, noise, and bot. This is because we adopted the book title-based collection method for our system because of their expected naturalness in obtained TMBs. This collection method pass through not only irrelevant tweets that do not mention any books (noise), but also mechanically generated tweets that appear frequently in Twitter (bot). Since the exposure system aims to amplify the mentions around online users' social networks, both types of non-TMBs should be rejected. We proposed a two-step pipeline, i.e. bot filtering followed by noise rejection, rather than an end-to-end ternary classifier. We showed that our model outperformed baselines and achieved the practical performance.

Chapter 8: iTMB scoring

In the exposure system's pipeline, the iTMB scoring step consists of two modules: the pleasantness scorer and the textual considerateness scorer. Based on the result of the descriptive analysis in Chapter 6, we defined pleasantness scoring as ternary text classification where classes are positive, negative, and neutral. This formalisation allowed us to apply existing opinion-mining methods to the task. Our model achieved the practical performance although there are room for improvement in terms of class imbalance.

Considerateness scoring, on the other hand, was formed as forceful phrase detection. Following the findings in Chapter 6, we compiled rules for detecting the expressions to recommend books and their strengths. In the evaluation, our rules performed reasonably well, while understanding syntactic information was raised as a possible way for improvement.

10 Outlooks

In this chapter, we mention future directions of this research. We can group them broadly into three: NLP performance improvement, UI/UX implementation, and the system operation.

10.1 NLP performance improvement

As summarised in Chapter 9, we solved the tasks for core NLP modules at the practical level. However, our methods still have room for improvement in terms of their performance. Better performance of the core NLP modules can bring finer-grained control over TMBs to be exposed to users by the system.

Now that we formulated the tasks for each module into NLP frameworks, our work has made it easier to refer to the state-of-the-art methodologies in relevant NLP fields. For each core NLP module, we have already showed possible paths for technical enhancement in the last section of the corresponding chapter, as follows:

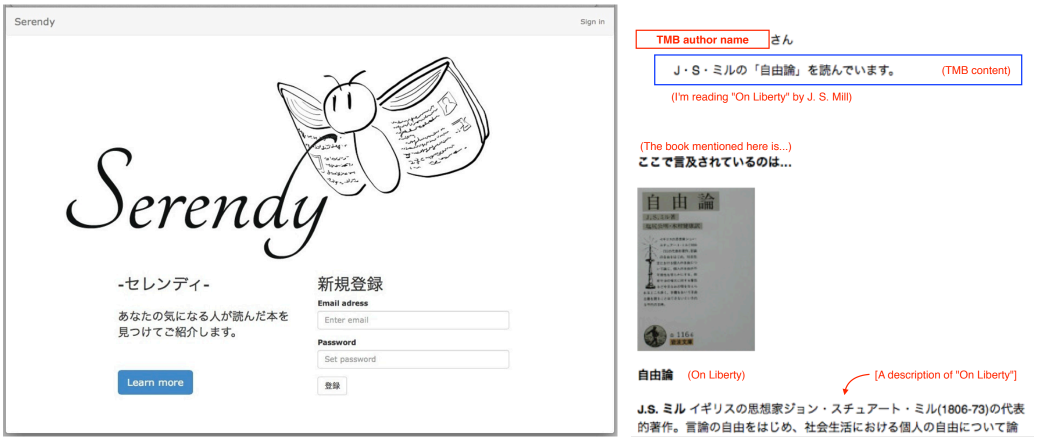
TMB identifier: Section 7.6.3

Pleasantness scorer: Section 8.1.4

Textual considerateness scorer: Section 8.2.4

10.2 UI/UX implementation

In this thesis, we focused on implementing the core NLP modules of the exposure system. Towards a digital surrogate system for passive exposure to books, another significant part to implement is the module for UI/UX, i.e. *presentation interface on media of dailiness* (Section 3.2.4). As we described in Section 3.2.1, the pipeline architecture outputs iTMBs around the user at the second step. In the third *iTMB exposure step*, the system exposes TMBs to the user based on their inspiringness. The *presentation interface* module can take several different formats as long as it meets dailiness, such as emails, messengers, and OSM client applications. How much (or rich) information should be delivered is another



Serendy

Sign in

Serendy

-セレンディ-

あなたの気になる人が読んだ本を
見つけてご紹介します。

Learn more

新規登録

Email adress

Enter email

Password

Set password

登録

TMB author name さん

J・S・ミルの「自由論」を読んでいます。 (TMB content)

(I'm reading "On Liberty" by J. S. Mill)

(The book mentioned here is...)
ここで言及されているのは...

自由論 (On Liberty) [A description of "On Liberty"]

J.S. ミル イギリスの思想家ジョン・スチュアート・ミル(1806-73)の代表的著作。言論の自由をはじめ、社会生活における個人の自由について論

(a) The top page where users can register to the system

(b) An example exposure to a TMB, formed as a mail notification

Figure 10.1: An experimental implementation of the digital surrogate system for physical passive exposure to books, named *Serendy*

issue. For instance, emails may contain more detailed text and images, but messages in messenger applications should be concise, due to their media characteristics.

We shall introduce our experimental implementation of the interface, for example, which is being built as a web-based UI with email notifications. The reason of this setup is because this can be preliminary operated in-house, i.e. just on one self-hosting server. This system sample allows users to register their Twitter accounts (Figure 10.1a), and then the system will occasionally send the users emails of detected TMBs (Figure 10.1b) after the processing of internal modules such as the TMB identifier and the iTMB scorer. While these UI/UX designs are still immature, we are experimentally implementing an exposure style with rich content, where extra book information (i.e. plots) is added within the email notification.

10.3 System operation

Once we implement the interface, we can conduct an experiment using the exposure system for the investigation of the inspiringness effect: how the values or degrees of inspiringness components affect users' attitude to reading. In other words, this would be the attempt to reveal the mechanism of passive exposure to books within digital environments—how can digital passive exposure to books inspire people to read?

Since the effect of passive exposure is expected to be subtle in comparison to, e.g. the actual action of reading, a long-term experiment will be required to observe it. The exposure system enables such experiments by collecting logs of user behaviours. Moreover, if we could retrieve users' personal traits under their permissions using online questionnaires, we can examine the relationships between the personal traits and inspiringness of exposed TMBs.

From the technical point of view, the *inspiringness coordinator* module in the iTMB exposure step (Section 3.2.4) plays the key role, because it adjusts the degrees of inspiringness components of TMBs to be exposed based on users' feedback. While this functionality aims to personalise inspiringness to individual users, we can control or fix the inspiringness of TMBs to be exposed at a certain level for experimental purposes. In this way, we can measure attitude towards reading among different groups separated by personal traits. In the future, a digital surrogate system for passive exposure to books may also contribute to fundamental research questions in LIS, e.g. how reading environments establish people's reading habits.

Aside from the research-centred view, we should mention the social value of the system operation. We proposed the necessity of a digital surrogate system for physical passive exposure to books. The popularisation of such systems is absolutely essential, in order to achieve the ultimate goal of increasing passive exposure to books in digital environments, which will alleviate the potential educational gap in future. One of the impacts of this thesis comes from the fact that we publicly opened up the technology to implement the TMB identification step and the iTMB scoring step of the system's pipeline architecture. That is, this research has made it feasible to implement a variety of exposure systems for anyone who realises the issue of less exposure to books.[1] While the content of such exposure should be unbiased as much as possible, it has a critical meaning that the methodology for the system has been developed free from conflict of interest to publishing or any other book/media-related business.[2] We conclude this thesis by hoping that our research outcome encourages the society to implement various book-exposure systems.

[1] Our research also contributed to announcing the importance of passive exposure to books especially in digital environments.

[2] This can also answer a question why this sort of system-oriented work should have been carried out in the research sector.

Appendices

A Original questionnaire questions used in Chapter 4

(Parenthesised text like this paragraph is additional notes for readers of this thesis. English translation is written in *italic* font. The original questionnaire includes Japanese only.)

A.1 TMB-like messages

(Messages are randomly sorted in the actual questionnaire. See other setups in Section 4.2.)

Instruction text

以下に提示する文章を、あなたの知っている人がSNSに投稿した発言だと思ってください。それぞれの発言を受けて、

- 話題に上がっている本をどれくらい読みたくなったか
- その表現をどれくらい押し付けがましく感じたか

をお答えください。なお、発言中の『TITLE』は本のタイトルを表し、AUTHORは著者名を表します。

Please imagine that the text proposed below is the mentions posted on online social media by your acquaintances. For each message, please answer:

- *how much do you come to want to read the book mentioned in the message*
- *how pushy do you feel about the message*

“TITLE” stands for book titles, AUTHOR denotes author names.

Neutral-none

- (the reference message; this appears twice)
  『TITLE』を読んでる。

  I’m reading “TITLE”.

- 『TITLE』はもう収集[a]つかない感じだけど一応斜め読みはしてる。

  “TITLE” can’t be settled anymore, but anyway I read it diagonally.

[a] 収集 (to collect) is a typo of 収拾 (settling). This example came from a real Japanese tweet example, and we kept it to maintain the taste of OSM posts.

Negative-none

• 『TITLE』AUTHOR 著、読了。なんか唐突に終わってしまって、おいてけぼりを食った気分。ちゃんと説明してほしい。。	*Finished reading “TITLE” by AUTHOR. It’s ended abruptly, me feeling like left behind. I want to be explained properly…*
• 『TITLE』読む。生理的に全く合わず。	*Have read “TITLE”. It doesn’t fit physiologically at all.*
• 『TITLE』なる本を微妙だなーって思いながら最後まで読んだけど、結局何も残らない感じだった。ざんねーん	*I read the book “TITLE” to the last minute, thinking that it was subtle, but after all it was like nothing left. I’m sorry*

Negative-weak

• みんなは『TITLE』みたいな変な本、あんまり読まない方がいいよ	*Everyone, shouldn’t read weird books like “TITLE”*
• 『TITLE』は本当に内容薄いからおすすめしない。	*“TITLE” is not recommended because it is really thin.*
• 『TITLE』なんて読んでたら時間無駄にするよ、やめときなよ	*If you read “TITLE”, you’ll waste time, should stop.*

Negative-strong

• 『TITLE』読むと気分悪くなる。。絶対読んじゃダメ	*Reading “TITLE” makes me feel sick…Never read*
• 書店で見つけた『TITLE』、タイトルに惹かれたけどゴミもいいところ。みんなは読むなよ！	*“TITLE” I found at a bookstore is trashy though I was attracted to the title. Don’t read it, everyone!*
• 決して読んではいけない本『TITLE』。中身のひどさが段違い	*The book “TITLE” you should never read. The terrible content is uneven*

Negative-excess

• いつまでも『TITLE』とか読んでる奴はどうかしてると思う。まさか読んでないよね？	*I think guys still reading “TITLE” are insane. Surely not are you reading it?*
• え、『TITLE』読む人いるの？ 人格疑うわ	*Huh, is there anyone reading “TITLE”? I doubt their personality*
• 『TITLE』を書いたやつもアホだけど読むやつも本当にアホ。	*The guy who wrote “TITLE” is stupid, but guys who read it are really stupid.*

Positive-none

- 『TITLE』って N 巻で完結だったのか……。明日買って帰ろう
  Was “TITLE” completed in volume N. . . ? Will buy it on the way home tomorrow
- AUTHOR『TITLE』読了。なんてバカなんだ。なのになんで最後ちょっと感動するんだ。電車で読むとニヤニヤしたりウルウルしたりで大変だった。面白かった。
  Finished reading AUTHOR’s “TITLE”. What an idiot! But why am I so impressed at the end? When I was reading it on the train, it was hard to hold my grin and tears. Was a nice book.
- 『TITLE』が面白かった
  “TITLE” was interesting

Positive-weak

- そんなあなたには AUTHOR さんの『TITLE』をおすすめしよう。
  I recommend AUTHOR’s “TITLE” to the whom like you.
- 最近出た『TITLE』は読んでおいて損はないと思うよ。いい意味で何が正解かなんて、まるで分からなくなるから。
  I think there’s no loss to read the recent “TITLE”. You may lose the idea of what’s the correct answer.
- 『TITLE』って本、同じ系統だから読んでみて
  Try “TITLE”, because it’s a book of the same kind.

Positive-strong

- AUTHOR さんの『TITLE』をめっっっっちゃおすすめする!!!!!
  I realllly recommend AUTHOR’s “TITLE”!!!!!
- 『TITLE』は学生なら必ず読むべき本。
  “TITLE” is a book that students must read.
- 当然『TITLE』は読んでるよね。
  Of course you read “TITLE”, don’t you?

Positive-excess

- 大学生にもなって『TITLE』読んだことないとかクソ。
  It’s shit that you have never read “TITLE” even though you’ve entered a college.
- ふつう『TITLE』は高校生の間に 3 回は読んでおくべきだろ。
  Normally, you should have read “TITLE” at least three times during your high school days.
- 『TITLE』も知らないの?! 早く読んで底辺抜けなよ〜
  Oh man, don’t you know even “TITLE”?! Read sooner, or stay in an underclass, haha

A.2 Reading attitude and behaviour

Instruction text

あなたの読書習慣についてお聞きします。読書は「本を読むこと」とし、「本」にはコミック（マンガ）や雑誌を含みません。

We will ask you about your reading habits. Reading here means “reading a book”, and “books” do not include comics (manga) or magazines.

Questions

• 読書は好きですか？	*Do you like reading books?*
• 読書は大事だと思いますか？	*Do you think reading books is important?*
• 普段どれくらい読書しますか？	*How often do you read books?*
• 本を読む日を平均すると、1 日あたりおよそ何時間を読書に費やしますか？	*How many hours for a day do you spend in average when you read books?*
• 今の読書量を変えたいと思いますか？	*Do you want to change your amount of reading?*

Acronyms

API	application programming interface
ELM	elaborate likelyhood model
ELPD	expected log pointwise predictive density
EWoM	electronic word-of-mouth
ISBN	international standard book number
LIS	library and information science
LOO	leave one out
LOOIC	leave-one-out information criterion
MAD	median absolute deviation
MaxEnt	maximum entropy modelling
ML	machine learning
NE	named entity
NER	named entity recognition
NN	neural network
OSM	online social media
SNS	social networking site
SRS	social reading service
SVM	support vector machine
TCTS(s)	tweet(s) that contain (book) title strings
TMB(s)	tweet(s) that mention books
UI	user interface
UX	user experience
WAIC	widely applicable information criterion
WoM	word-of-mouth

Glossary

Considerateness The degree to which the encounter of the TMB and the TMB itself is considerate to the user

Dailiness The degree to which the encounter of TMBs are embedded in the user's daily life

Infrequent reader Those who desire to have a reading habit but have not yet achieved so.

Inspiringness The extent to which a reader of the TMB is inspired to read the mentioned book. The four principal components are identified; see also dailiness, proximity, pleasantness, considerateness.

Pleasantness The apparently pleasant manner of the TMB author toward the mentioned book.

Proximity The degree to which the author of the TMB is familiar to the user

Bibliography

ACSI, 2018. *ACSI E-business report 2018*, (technical report).

Adobe, 2013. *Click here: the state of online advertising*. Adobe Systems Incorporated, (technical report).

Akl, E.A., Guyatt, G.H., Irani, J., Feldstein, D., Wasi, P., Shaw, E., Shaneyfelt, T., Levine, M. and Schünemann, H.J., 2012. "Might" or "suggest"? No wording approach was clearly superior in conveying the strength of recommendation. *Journal of Clinical Epidemiology*, 65(3), pp.268–275.

Alharthi, H., Inkpen, D. and Szpakowicz, S., 2018. A survey of book recommender systems. *Journal of Intelligent Information Systems*, 51(1), pp.139–160.

Alothali, E., Zaki, N., Mohamed, E.A. and Alashwal, H., 2018. Detecting social bots on twitter: a literature review. *International Conference on Innovations in Information Technology (IIT).* IEEE, pp.175–180.

Amigó, E., Carrillo de Albornoz, J., Chugur, I., Corujo, A., Gonzalo, J., Martín-Wanton, T., Meij, E., de Rijke, M. and Spina, D., 2013. Overview of RepLab 2013: evaluation of online reputation monitoring systems. In: Forner, P., Navigli, R., Tufis, D. and Ferro, N. eds. *Working Notes for CLEF 2013 Conference.* Vol. 1179, CEUR Workshop Proceedings. CEUR-WS.org, pp.1–20.

Amigó, E., Corujo, A., Gonzalo, J., Meij, E. and de Rijke, M., 2012. Overview of RepLab 2012: evaluating online reputation management systems. In: Forner, P., Karlgren, J. and Womser-Hacker, C. eds. *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes.* Vol. 1178, CEUR Workshop Proceedings. CEUR-WS.org, pp.1–24.

Andersen, I.G. and Jæger, M.M., 2015. Cultural capital in context: heterogeneous returns to cultural capital across schooling environments. *Social Science Research*, 50, pp.177–188.

Anger, I. and Kittl, C., 2011. Measuring influence on Twitter. *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies*. New York, New York, USA: ACM Press, pp.1–4.

Aral, S. and Walker, D., 2014. Tie strength, embeddedness, and social influence: a large-scale networked experiment. *Management Science*, 60(6), pp.1352–1370.

Aramaki, E., Maskawa, S. and Morita, M., 2011. Twitter catches the flu: detecting influenza epidemics using Twitter. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp.1568–1576.

Ari, E. and Yildiz, Z., 2014. Parallel lines assumption in ordinal logistic regression and analysis approaches. *International Interdisciplinary Journal of Scientific Research*, 1(3), pp.8–23.

Aron, A., Aron, E.N. and Smollan, D., 1992. Inclusion of other in the self scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology*, 63(4), pp.596–612.

Artstein, R. and Poesio, M., 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4), pp.555–596.

Austin, J.L., 1962. *How to do things with words: second edition.*

Bakshy, E., Rosenn, I., Marlow, C. and Adamic, L., 2012. The role of social networks in information diffusion. *Proceedings of the 21st International Conference on World Wide Web*, WWW '12. New York, NY, USA: ACM, pp.519–528.

Belinkov, Y. and Glass, J., 2019. Analysis methods in neural language processing: a survey. *Transactions of the Association for Computational Linguistics*, 7, pp.49–72.

Benesse Educational Research and Development Institute, 2018. **小学生の読書に関する実態調査・研究** *[An actual condition survey of elementary school children's reading; in Japanese]*, (technical report).

Bhattacharya, P., Ghosh, S., Kulshrestha, J., Mondal, M., Zafar, M.B., Ganguly, N. and Gummadi, K.P., 2014. Deep Twitter diving: exploring topical groups in microblogs at scale. *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing - CSCW '14*. New York, New York, USA: ACM Press, pp.197–210.

Bogers, T. and Björneborn, L., 2013. Micro-serendipity: meaningful coincidences in everyday life shared on Twitter. *iConference 2013 Proceedings.* iSchools, pp.196–208.

Bond, C.(U., Ferraro, C.(U., Luxton, S.(U. and Sands, S.(U., 2010. Social media advertising: an investigation of consumer perception, attitudes and preferences for engagement. *Proceedings of 2010 Australian and New Zealand Marketing Academy*, pp.1–9.

Bosco, C., Patti, V. and Bolioli, A., 2013. Developing corpora for sentiment analysis: the case of irony and Senti-TUT. *IEEE Intelligent Systems*, 28(2), pp.55–63.

Bourdieu, P., 1986. The forms of capital. *Handbook of theory and research for the sociology of education.* New York: Greenwood Press.

Brehm, J.W., 1966. *A theory of psychological reactance.* New York: Academic Press, p.135.

Brehm, S.S. and Brehm, J.W., 1981. *Psychological reactance: a theory of freedom and control.* New York: Academic Press, p.432.

Breiman, L., 1996. Bagging predictors. *Machine Learning*, 24(2), pp.123–140.

Breiman, L., 2001. Random forests. English. *Machine Learning*, 45(1), pp.5–32.

Brin, S., 1999. Extracting patterns and relations from the World Wide Web. In: Atzeni, P., Mendelzon, A.O. and Mecca, G. eds. *Selected Papers from the International Workshop on The World Wide Web and Databases*, WebDB '98. London, UK: Springer, pp.172–183.

Briñol, P., Petty, R.E., Valle, C., Rucker, D.D. and Becerra, A., 2007. The effects of message recipients' power before and after persuasion: A self-validation analysis. *Journal of Personality and Social Psychology*, 93(6), pp.1040–1053.

Brown, J.J. and Reingen, P.H., 1987. Social ties and word-of-mouth referral behavior. *Journal of Consumer Research*, 14(3), p.350.

Bruns, A. and Burgess, J., 2011. The use of Twitter hashtags in the formation of ad hoc publics. *Proceedings of the 6th European Consortium for Political Research General Conference*, pp.1–9.

Buller, D.B., Burgoon, M., Hall, J.R., Levine, N., Taylor, A.M., Beach, B., Klein Buller, M. and Melcher, C., 2000. Long-term effects of language intensity in preventive messages on planned family solar protection. *Health Communication*, 12(3), pp.261–275.

Burgoon, M., Alvaro, E., Grandpre, J. and Voloudakis, M., 2002. Revisiting the theory of psychological reactance: communicating threats to attitudinal freedom. *The persuasion handbook: Developments in theory and practice.*

Bürkner, P.-C., 2017. brms: an R package for bayesian multilevel models using stan. *Journal of Statistical Software*, 80(1), pp.1–28.

Bus, A.G., van IJzendoorn, M.H. and Pellegrini, A.D., 1995. Joint book reading makes for success in learning to read: a meta-analysis on intergenerational transmission of literacy. *Review of Educational Research*, 65(1), pp.1–21.

Carter, S., Weerkamp, W. and Tsagkias, M., 2013. Microblog language identification: overcoming the limitations of short, unedited and idiomatic text. *Language Resources and Evaluation*, 47(1), pp.195–215.

Chaiken, S., 1980. Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology*, 39(5), pp.752–766.

Chakrabarty, T., Gupta, K. and Muresan, S., 2019. Pay 'attention' to your context when classifying abusive language. *Proceedings of the Third Workshop on Abusive Language Online*, 2017. Stroudsburg, PA, USA: Association for Computational Linguistics, pp.70–79.

Chandrashekar, G. and Sahin, F., 2014. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1), pp.16–28.

Chavoshi, N., Hamooni, H. and Mueen, A., 2016. DeBot: Twitter bot detection via warped correlation. *Proceedings of IEEE 16th International Conference on Data Mining*. IEEE, pp.817–822.

Chen, K., Luo, P. and Wang, H., 2017. An influence framework on product Word-of-Mouth (WoM) measurement. *Information & Management*, 54(2), pp.228–240.

Cheung, C.M. and Thadani, D.R., 2012. The impact of electronic word-of-mouth communication: a literature analysis and integrative model. *Decision Support Systems*, 54(1), pp.461–470.

Chevalier, J.A. and Mayzlin, D., 2006. The effect of word of mouth on sales: online book reviews. *Journal of Marketing Research*, 43(3), pp.345–354.

China Audio-video and Digital Publishing Association, 2019. *2018 Digital Reading White Paper [in Chinese]*, (technical report).

Choi, Y.K., Seo, Y. and Yoon, S., 2017. E-WoM messaging on social media: social ties, temporal distance, and message concreteness. *Internet Research*, 27(3), pp.495–505.

Chu, Z., Gianvecchio, S., Wang, H. and Jajodia, S., 2012. Detecting automation of Twitter accounts: are you a human, bot, or cyborg? *IEEE Transactions on Dependable and Secure Computing*, 9(6), pp.811–824.

Clark, C. and Rumbold, K., 2006. *Reading for pleasure: a research overview*, (technical report).

Cohen, J., 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), pp.37–46.

Cohn, N., 2014. Building a better ‘comic‘ theory: shortcoming of theoretical research on comics and how to overcome them. *Studies in Comics*, 5(1), pp.57–75.

Cortes, C. and Vapnik, V., 1995. Support-vector networks. *Machine Learning*, 20(3), pp.273–297.

Council on Library and Information Resources, 2005. *Library as place: rethinking roles, rethinking space.* Council on Library and Information Resources.

Courtois, C., Slechten, L. and Coenen, L., 2018. Challenging Google Search filter bubbles in social and political information: disconforming evidence from a digital methods case study. *Telematics and Informatics*, 35(7), pp.2006–2015.

Crano, W.D. and Prislin, R., 2006. Attitudes and persuasion. *Annual Review of Psychology*, 57(1), pp.345–374.

Crichton, G., Pyysalo, S., Chiu, B. and Korhonen, A., 2017. A neural network multi-task learning approach to biomedical named entity recognition. *BMC Bioinformatics*, 18(1), p.368.

Cronbach, L.J., 1951. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), pp.297–334.

Dagoula, C., 2019. Mapping political discussions on Twitter: where the elites remain elites. *Media and Communication*, 7(1), p.225.

Davis, C.A., Varol, O., Ferrara, E., Flammini, A. and Menczer, F., 2016. BotOrNot: a system to evaluate social bots. *Proceedings of the 25th International Conference Companion on World Wide Web*. New York, New York, USA: ACM Press, pp.273–274.

Davis, J. and Goadrich, M., 2006. The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd International Conference on Machine Learning*. New York, New York, USA: ACM Press, pp.233–240.

de Gemmis, M., Lops, P., Semeraro, G. and Musto, C., 2015. An investigation on the serendipity problem in recommender systems. *Information Processing and Management*, 51(5), pp.695–717.

Derczynski, L., Maynard, D., Rizzo, G., Van Erp, M., Gorrell, G., Troncy, R., Petrak, J. and Bontcheva, K., 2014. Analysis of named entity recognition and linking for tweets. *Information Processing and Management*, 51(2), pp.32–49.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K., 2019. BERT: pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Stroudsburg, PA, USA: Association for Computational Linguistics, pp.4171–4186.

Dey, K., Kaushik, S., Garg, K. and Shrivastava, R., 2018. Topic lifecycle on social networks: analyzing the effects of semantic continuity and social communities. In: Pasi, G., Piwowarski, B., Azzopardi, L. and Hanbury, A. eds. *Advances in Information Retrieval*. Cham: Springer International Publishing, pp.29–42.

Diakopoulos, N. and Koliska, M., 2017. Algorithmic transparency in the news media. *Digital Journalism*, 5(7), pp.809–828.

Dillard, J.P. and Shen, L., 2005. On the nature of reactance and its role in persuasive health communication. *Communication Monographs*, 72(2), pp.144–168.

DiMaggio, P., 1982. Cultural capital and school success: the impact of status culture participation on the grades of U.S. high school students. *American Sociological Review*, 47(2), pp.189–201.

Dow, P.A., Adamic, L. and Friggeri, A., 2013. The anatomy of large Facebook cascades. *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, pp.145–154.

Downey, D., Broadhead, M. and Etzioni, O., 2007. Locating complex named entities in web text. *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence*, pp.2733–2739.

Dronkers, J., 1992. Parents, love, and money: the relations between parental class, cognitive skill, educational attainment, occupation, marriage, and family income. *International Perspectives on Education and Society*, 2, pp.277–293.

Duhan, D., Johnson, S., Wilcox, J. and Harrell, G., 1997. Influences on consumer use of word-of-mouth recommendation sources. *Journal of the Academy of Marketing Science*, 25(4), pp.283–295.

Eady, G., Nagler, J., Guess, A., Zilinsky, J. and Tucker, J.A., 2019. How many people live in political bubbles on social media? evidence from linked survey and Twitter data. *SAGE Open*.

Eager, C. and Roy, J., 2017. Mixed effects models are sometimes terrible. *Linguistic Society of America*, pp.1–24.

Erdelez, S., Heinström, J., Makri, S., Björneborn, L., Beheshti, J., Toms, E. and Agarwal, N.K., 2016. Research perspectives on serendipity and information encountering. *Proceedings of the Association for Information Science and Technology*, 53(1), pp.1–5.

Erdelez, S., 1999. Information encountering: it's more than just bumping into information. *Bulletin of the American Society for Information Science and Technology*, 25(3), pp.26–29.

Erdmann, M., Ward, E., Ikeda, K., Hattori, G., Ono, C. and Takishima, Y., 2013. Automatic labeling of training data for collecting Tweets for ambiguous TV program titles. *Proceedings of 5th International Conference on Social Computing*, pp.796–802.

Evans, M.D.R., Kelley, J. and Sikora, J., 2014. Scholarly culture and academic performance in 42 nations. *Social Forces*, 92(4), pp.1573–1605.

Evans, M.D., Kelley, J., Sikora, J. and Treiman, D.J., 2010. Family scholarly culture and educational success: books and schooling in 27 nations. *Research in Social Stratification and Mobility*, 28(2), pp.171–197.

Flaxman, S., Goel, S. and Rao, J.M., 2016. Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly*, 80(S1), pp.298–320.

Fleming, M.A. and Petty, R.E., 2000. *Identity and persuasion: an elaboration likelihood approach.* Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Fletcher, R. and Nielsen, R.K., 2018. Are people incidentally exposed to news on social media? A comparative analysis. *New Media and Society*, 20(7), pp.2450–2468.

Freund, Y. and Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1), pp.119–139.

Fukada, H., 2002. *説得心理学ハンドブック: 説得コミュニケーション研究の最前線 [Persuasive psychology handbook: the frontier of persuasive communication research; in Japanese].* Ed. by H. Fukada.

Gelman, A. and Rubin, D.B., 1992. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4), pp.457–472.

Geschke, D., Lorenz, J. and Holtz, P., 2019. The triple-filter bubble: using agent-based modelling to test a meta-theoretical framework for the emergence of filter bubbles and echo chambers. *British Journal of Social Psychology*, 58(1), pp.129–149.

Ghosh, J., Li, Y. and Mitra, R., 2018. On the use of Cauchy prior distributions for Bayesian logistic regression. *Bayesian Analysis*, 13(2), pp.359–383.

Gibson, J.J., 1979. *The ecological approach to visual perception.* Houghton Mifflin.

Go, A., Bhayani, R. and Huang, L., 2009. *Twitter sentiment classification using distant supervision.* Stanford, (technical report), pp.1–12.

Gonzales, L., 2014. An analysis of Twitter conversations at academic conferences. *Proceedings of the 32nd ACM International Conference on The Design of Communication.* New York, New York, USA: ACM Press, pp.1–8.

Graells-Garrido, E., Baeza-Yates, R. and Lalmas, M., 2019. How representative is an abortion debate on Twitter? *Proceedings of the 10th ACM Conference on Web Science.* New York, New York, USA: ACM Press, pp.133–134.

Gridach, M., 2017. Character-level neural network for biomedical named entity recognition. *Journal of Biomedical Informatics*, 70, pp.85–91.

Grier, C., Thomas, K., Paxson, V. and Zhang, M., 2010. @Spam: the underground on 140 characters or less. *Proceedings of the 17th ACM Conference on Computer and Communications Security*, pp.27–37.

Gupta, P. and Harris, J., 2010. How e-WoM recommendations influence product consideration and quality of choice: a motivation to process information perspective. *Journal of Business Research*, 63(9-10), pp.1041–1049.

Guzman, E., Alkadhi, R. and Seyff, N., 2017. An exploratory study of Twitter messages about software applications. *Requirements Engineering*, 22(3), pp.387–412.

Habib, M.B. and Van Keulen, M., 2016. TwitterNEED: a hybrid approach for named entity extraction and disambiguation for tweet. *Natural Language Engineering*, 22(3), pp.423–456.

Hafezieh, N. and Eshraghian, F., 2017. Affordance theory in social media research: systematic review and synthesis of the literature. *Proceedings of the 25th European Conference on Information Systems*, pp.3155–3166.

Haim, M., Graefe, A. and Brosius, H.B., 2018. Burst of the filter bubble?: effects of personalization on the diversity of Google News. *Digital Journalism*, 6(3), pp.330–343.

Halberstam, Y. and Knight, B., 2016. Homophily, group size, and the diffusion of political information in social networks: evidence from Twitter. *Journal of Public Economics*, 143, pp.73–88.

Hashimoto, Y., 2016. **日本人の情報行動** *2015 [Information behaviour 2015; in Japanese]*. Ed. by Y. Hashimoto. University of Tokyo Press.

Himelboim, I., McCreery, S. and Smith, M., 2013. Birds of a feather Tweet together: integrating network and content analyses to examine cross-ideology exposure on Twitter. *Journal of Computer-Mediated Communication*, 18(2), pp.40–60.

Himelboim, I., Smith, M. and Shneiderman, B., 2013. Tweeting apart: applying network analysis to detect selective exposure clusters in Twitter. *Communication Methods and Measures*, 7(3), pp.169–197.

Hoffman, M.D. and Gelman, A., 2014. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15, pp.1593–1623.

Holzinger, A., 2016. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics*, 3(2), pp.119–131.

Huang, J., Cheng, X.-Q., Shen, H.-W., Zhou, T. and Jin, X., 2012. Exploring social influence via posterior effect of word-of-mouth recommendations. *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12. New York, NY, USA: ACM, pp.573–582.

Imajo, S., 2012. Do high-pressure communications cause reactance or compliance? [in Japanese]. *Annual Bulletin of Institute of Psychological Studies in Showa Women's University*, 14, pp.1–9.

Ivanovic, E., 2005. Dialogue act tagging for instant messaging chat sessions. *Proceedings of the ACL Student Research Workshop*, pp.79–84.

Izyan, N., Saat, Y., Azman, S., Noah, M. and Mohd, M., 2018. Towards serendipity for content-based recommender systems. *International Journal on Advanced Science, Engineering and Information Technology*, 8(4-2), pp.1762–1769.

J. O'Keefe, D., 2012. Persuasion. *The International Encyclopedia of Communication*. Chichester, UK: John Wiley & Sons, Ltd.

Jamieson, K.H. and Cappella, J.N., 2008. *Echo chamber: rush Limbaugh and the conservative media establishment*. Oxford University Press.

Japan MEXT, 2015. *高校生の読書に関する意識等調査報告書 [Survey report on awareness of reading by high-school students; in Japanese]*. Japanese Ministry of Education, Culture, Sports, Science and Technology, (technical report).

Japan MEXT, 2016. *平成 28 年度「地域における読書活動推進のための体制整備に関する調査研究」[Survey study 2016 on system development for promoting reading activities in local communities; in Japanese]*. Japanese Ministry of Education, Culture, Sports, Science and Technology, (technical report).

Japan MEXT, 2019. *平成 30 年度「子供の読書活動推進計画に関する調査研究」[Survey Study 2018 on children's reading-activities promotion plan; in Japanese]*. Japanese Ministry of Education, Culture, Sports, Science and Technology, (technical report).

Japan Publishing Industry Foundation for Culture, 2009. *現代人の読書実態調査 [Survey of actual reading by modern people; in Japanese]*, (technical report).

Japanese Agency for Cultural Affairs, 2019. 平成 *30* 年度「国語に関する世論調査」*[Opinion poll about Japanese language in 2018; in Japanese]*, (technical report).

Jhangiani, R. and Tarry, H., 2014. *Principles of social psychology*. 1st Intern. Victoria, B.C.: BCcampus.

Jiang, R., Chiappa, S., Lattimore, T., Agyorgy, A., Kohli, P., Sinha, A., Gleich, D.F. and Ramani, K., 2016. Deconvolving feedback loops in recommender systems. In: Lee, D.D., Sugiyama, M., Luxburg, U.V., Guyon, I. and Garnett, R. eds. *Advances in Neural Information Processing Systems*. Vol. 29. Curran Associates, Inc., pp.3243–3251.

Jiang, T., Liu, F. and Chi, Y., 2015. Online information encountering: modeling the process and influencing factors. *Journal of Documentation*, 71(6), pp.1135–1157.

Johnson, J.M. and Khoshgoftaar, T.M., 2019. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1), p.27.

Kaikati, A.M. and Kaikati, J.G., 2004. Stealth marketing: how to reach consumers surreptitiously. *California Management Review*, 46(4), pp.6–22.

Kaiser, J. and Quandt, T., 2016. Book lovers, bibliophiles, and fetishists: the social benefits of heavy book usage. *Psychology of Popular Media Culture*, 5(4), pp.356–371.

Kaminskas, M. and Bridge, D., 2016. Diversity, serendipity, novelty, and coverage: a survey and empirical analysis of beyond-Accuracy objectives in recommender systems. *ACM Transactions on Interactive Intelligent Systems*, 7(1), pp.1–42.

Karppinen, P. and Oinas-Kukkonen, H., 2013. Three approaches to ethical considerations in the design of behavior change support systems. In: Berkovsky, S. and Freyne, J. eds. *International Conference on Persuasive Technology*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp.87–98.

Kaur, P., Singhal, A. and Kaur, J., 2016. Spam detection on Twitter: a survey. *3rd International Conference on Computing for Sustainable Global Development*, pp.2570–2573.

Kazai, G., Landoni, M., Eickhoff, C. and Brusilovsky, P., 2012. BooksOnline'12: 5th workshop on online books, complementary social media and their impact. *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pp.2764–2765.

Kemp, S., 2019. *Digital 2019: global ditigtal yearbook*. Hootsuite, (technical report).

Kendall, M.G., 1938. A new measure of rank correlation. *Biometrika*, 30(1-2), pp.81–93.

Kessler, T. and Jena, E.-a.-h., 2002. Prejudice and extremism: explanations based on ingroup projection, perspective divergence, and minimal standards. *Social Psychology*, 82(359), pp.427–434.

Kim, A., Miano, T., Chew, R., Eggers, M. and Nonnemaker, J., 2017. Classification of Twitter users who tweet about e-cigarettes. *JMIR Public Health and Surveillance*, 3(3), e63.

Kim, S.N., Cavedon, L. and Baldwin, T., 2012. Classifying dialogue acts in multi-party live chats. *Proceedings of 26th Pacific Asia Conference on Language, Information and Computation*, pp.463–472.

Kim, S.N., Wang, L. and Baldwin, T., 2010. Tagging and linking web forum posts. *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pp.192–202.

Kitamura, S., Hashimoto, Y., Kimura, T., Tsuji, D., Korenaga, R., Mori, Y., Ogasahara, M. and Kawai, D., 2018. Cross-national comparison of information behavior and social attitudes: online survey in Japan, China, South Korea, Singapore, and the United States [in Japanese]. *Research Survey Reports in Information Studies in Interfaculty Initiative in Information Studies, the University of Tokyo*, 34, pp.119–211.

Kitamura, S., Sasaki, Y. and Kawai, D., 2016. **ツイッターの心理学***:* **情報環境と利用者行動** *[Psychology on Twitter: information environment and behaviour; in Japanese]*. Seishinshobo.

Kiyota, N. and Horii, T., 2017. Remarks to enhance effectiveness of persuasion in lesson environment : how to prevent psychological reactance [in Japanese]. *Journal of Education Design in Yokohama National University*, (8), pp.71–79.

Knijnenburg, B.P., Willemsen, M.C., Gantner, Z., Soncu, H. and Newell, C., 2012. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, 22(4-5), pp.441–504.

Kobayashi, S., 2018. Contextual augmentation: data augmentation by words with paradigmatic relations. *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp.452–457.

Kokubu, H., Yamazaki, H. and Nosaka, M., 2013. Japanese stopword list making for keyword extraction suitable for semantic interpretation [in Japanese]. *Transactions of Japan Society of Kansei Engineering*, 12(4), pp.511–518.

Koolen, M., Bogers, T., Kazai, G. and Kamps, J., 2014. Overview of the INEX 2014 Social Book Search Track. In: Cappellato, L., Ferro, N., Halvey, M. and Kraaij, W. eds. *Working Notes for CLEF 2014 Conference*. Vol. 1180, CEUR Workshop Proceedings. CEUR-WS.org, pp.462–479.

Kotkov, D., Wang, S. and Veijalainen, J., 2016. A Survey of serendipity in recommender systems. *Knowledge-Based Systems*, 111, pp.180–192.

Kou, Z., Cohen, W.W. and Murphy, R.F., 2005. High-recall protein entity recognition using a dictionary. *Bioinformatics*, 21(Suppl 1), pp.i266–i273.

Kunimoto, C., Miyata, Y., Koizumi, M., Kinjo, Y. and Ueda, S., 2009. Dimensions of reading: findings from a focus group interview to adults [in Japanese]. *Journal of Japan Society of Library and Information Science*, 55(4), pp.199–212.

Labrador, B., Ramón, N., Alaiz-Moretón, H. and Sanjurjo-González, H., 2014. Rhetorical structure and persuasive language in the subgenre of online advertisements. *English for Specific Purposes*, 34(1), pp.38–47.

Lai, S., Xu, L., Liu, K. and Zhao, J., 2015. Recurrent convolutional neural networks for text classification. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp.2267–2273.

Landis, J.R. and Koch, G.G., 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1), pp.159–74.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H. and Kang, J., 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. Ed. by J. Wren, pp.1–8.

Lee, J., Walker, E., Burleson, W., Kay, M., Buman, M. and Hekler, E.B., 2017. Self-experimentation for behavior change: design and formative evaluation of two approaches. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp.6837–6849.

Lee, J., Kim, S. and Ham, C.D., 2016. A Double-edged sword? predicting consumers' attitudes toward and sharing intention of native advertising on social media. *American Behavioral Scientist*, 60(12), pp.1425–1441.

Lex, E., Wagner, M. and Kowald, D., 2018. Mitigating confirmation bias on Twitter by recommending opposing views. *European Symposium Series on Societal Challenges in Computational Social Science*, pp.1–3.

Li, H., Edwards, S.M. and Lee, J.H., 2002. Measuring the intrusiveness of advertisements: scale development and validation. *Journal of Advertising*, 31(2), pp.37–47.

Li, Z. and Yin, Y., 2018. Attractiveness, expertise and closeness: the effect of source credibility of the first lady as political endorser on social media in China. *Global Media and China*, 3(4), pp.297–315.

Liebrecht, C., Hustinx, L. and van Mulken, M., 2019. The relative power of negativity: the influence of language intensity on perceived strength. *Journal of Language and Social Psychology*, 38(2), pp.170–193.

Limsopatham, N. and Collier, N., 2016. Bidirectional LSTM for named entity recognition in Twitter messages. *Proceedings of the 2nd Workshop on Noisy User-generated Text*, pp.145–152.

Liu, B. and Zhang, L., 2012. A survey of opinion mining and sentiment analysis. English. In: Aggarwal, C.C. and Zhai, C. eds. *Mining Text Data*. Springer US. Chap. 13, pp.415–463.

Liu, I.L., Cheung, C.M. and Lee, M.K., 2016. User satisfaction with microblogging: information dissemination versus social networking. *Journal of the Association for Information Science and Technology*, 67(1), pp.56–70.

Liu, P., Qiu, X. and Huang, X., 2016. Recurrent neural network for text classification with multi-task learning. *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pp.2873–2879.

Liu, X., He, P., Chen, W. and Gao, J., 2019. Multi-task deep neural networks for natural language understanding. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, pp.4487–4496.

Logan, J.A.R., Justice, L.M., Yumuş, M. and Chaparro-Moreno, L.J., 2019. When children are not read to at home. *Journal of Developmental & Behavioral Pediatrics*, 40(5), pp.383–386.

Luna-Nevarez, C. and Torres, I.M., 2015. Consumer attitudes toward social network advertising. *Journal of Current Issues and Research in Advertising*, 36(1), pp.1–19.

Luyt, B. and Heok, A., 2015. David and Goliath: tales of independent bookstores in Singapore. *Publishing Research Quarterly*, 31(2), pp.122–131.

Mackie, D.M., Worth, L.T. and Asuncion, A.G., 1990. *Processing of persuasive in-group messages*. US: American Psychological Association.

Mainichi Shimbun, 2013. 読書世論調査 *[Reading Poll; in Japanese]*. Mainichi Shimbun.

Mainichi Shimbun, 2019. 読書世論調査 *[Reading Poll; in Japanese]*. Mainichi Shimbun.

Makri, S., Blandford, A., Woods, M., Sharples, S. and Maxwell, D., 2014. "Making my own luck": serendipity strategies and how to support them in digital information environments. *Journal of the Association for Information Science and Technology*, 65(11), pp.2179–2194.

Mar, R.A., Oatley, K. and Peterson, J.B., 2009. Exploring the link between reading fiction and empathy: ruling out individual differences and examining outcomes. *Communications*, 34(4), pp.407–428.

McCallum, A. and Nigam, K., 1998. A comparison of event models for naive Bayes text classification. *Proceedings of the AAAI-98 Workshop on Learning for Text Categorization*, pp.41–48.

McCay-Peet, L. and Toms, E.G., 2017. Researching serendipity in digital information environments. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 9(6), pp.i–91.

McCord, M. and Chuah, M., 2011. Spam detection on Twitter using traditional classifiers. English. In: Calero, J., Yang, L., Mármol, F., García Villalba, L., Li, A. and Wang, Y. eds. *Autonomic and Trusted Computing*. Vol. 6906, Lecture Notes in Computer Science. Berlin, Germany: Springer, pp.175–186.

McGrenere, J. and Ho, W., 2000. Affordances : clarifying and evolving a concept. *Proceedings of Graphics Interface 2000*, pp.1–8.

McPherson, M., Smith-Lovin, L. and Cook, J.M., 2001. Birds of a feather: homophily in social networks. *Annual review of Sociology*, 27, pp.415–444.

Miller, C.H., Lane, L.T., Deatrick, L.M., Young, A.M. and Potts, K.A., 2007. Psychological reactance and promotional health messages: the effects of controlling language, lexical concreteness, and the restoration of freedom. *Human Communication Research*, 33(2), pp.219–240.

Miron, A.M. and Brehm, J.W., 2006. Reactance theory—40 years later. *Zeitschrift fur Sozialpsychologie*, 37(1), pp.9–18.

Mohammad, S.M., Kiritchenko, S. and Martin, J., 2013. Identifying purpose behind electoral tweets. *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, pp.1–9.

Möller, J., Trilling, D., Helberger, N. and van Es, B., 2018. Do not blame it on the algorithm: an empirical assessment of multiple recommender systems and their impact on content diversity. *Information Communication and Society*, 21(7), pp.959–977.

Morder Intelligence, 2019. *E-book market—segmented by geography (North America, Europe, Asia-Pacific, South America, Middle East and Africa—growth, trends, and forecast (2019–2024),* (technical report).

Myers, S.A., Zhu, C. and Leskovec, J., 2012. Information diffusion and external influence in networks. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.33–41.

Nakazawa, J., 2005. Development of manga (comic book) literacy in children. In: Shwalb, D., Nakazawa, J. and Shwalb, B.J. eds. *Applied Developmental Psychology: Theory, Practice, and Research From Japan.* Greenwich, CT: Information Age Publishing. Chap. 2, pp.23–42.

Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F. and Stoyanov, V., 2016. SemEval-2016 Task 4: Sentiment analysis in Twitter. *Proceedings of the 10th International Workshop on Semantic Evaluation*, pp.312–320.

Nakov, P., Rosenthal, S., Kozareva, Z., Stoyanov, V., Ritter, A. and Wilson, T., 2013. SemEval-2013 Task 2: Sentiment Analysis in Twitter. *Proceedings of the 7th International Workshop on Semantic Evaluation*, pp.312–320.

National Institution For Youth Education, 2013. 子どもの読書活動の実態とその影響・効果に関する調査研究報告書 *[Research report on the actual conditions of child reading behaviour and its effect; in Japanese]*, (technical report).

National Library Board Singapore, 2019. *2018 national reading habits study on adults*, (technical report).

Negi, S., Daudert, T. and Buitelaar, P., 2019. SemEval-2019 Task 9: Suggestion Mining from Online Reviews and Forums. *Proceedings of the 13th International Workshop on Semantic Evaluation*, pp.877–887.

Netherlands Institute for Social Research, 2018. *Reading: time reading in the netherlands*, (technical report).

Neubig, G. and Duh, K., 2013. How much is said in a Tweet? A multilingual, information-theoretic perspective. *Proceedings of the AAAI Spring Symposium: Analyzing Microtext*, pp.32–39.

Neuman, S.B. and Knapczyk, J.J., 2018. Reaching families where they are: examining an innovative book distribution program. *Urban Education*, (online first).

Nguyen, T.T., Maxwell, P.-m.H.F., Loren, H. and Joseph, T., 2014. Exploring the filter bubble: the effect of using recommender systems on content diversity. *Proceedings of the 23rd international conference on World wide web*, pp.677–686.

Nigam, K., 1999. Using maximum entropy for text classification. *Papers of the Workshop on Machine Learning for Information Filtering*, pp.61–67.

Norman, D.A., 1988. *The psychology of everyday things*. Basic Books.

O'Hara, K. and Stevens, D., 2015. Echo chambers and online radicalism: assessing the internet's complicity in violent extremism. *Policy and Internet*, 7(4), pp.401–422.

OECD, 2018. *Preparing our youth for an inclusive and sustainable world: The OECD PISA global competence framework*.

Oinas-Kukkonen, H. and Harri, 2013. A foundation for the study of behavior change support systems. *Personal and Ubiquitous Computing*, 17(6), pp.1223–1235.

Oraby, S., Gundecha, P., Mahmud, J., Bhuiyan, M. and Akkiraju, R., 2017. "How may I help you?": modeling Twitter customer service conversations using fine-grained dialogue acts. *Proceedings of the 22nd International Conference on Intelligent User Interfaces*. New York, New York, USA: ACM Press, pp.343–355.

Ororbia Ii, A.G., Giles, C.L. and Reitter, D., 2015. Learning a deep hybrid model for semi-supervised text classification. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp.471–481.

Packard, G. and Berger, J., 2016. How language shapes word of mouth's impact. *Journal of Marketing Research*, 54(4), pp.572–588.

Pak, A. and Paroubek, P., 2010. Twitter as a corpus for sentiment analysis and opinion mining. *Proceedings of the Seventh International Conference on Language Resources and Evaluation*. Valletta, Malta, pp.1320–1326.

Pandey, G., Kotkov, D. and Semenov, A., 2018. Recommending serendipitous items using transfer learning. *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pp.1771–1774.

Pang, B., Lee, L. and Vaithyanathan, S., 2002. Thumbs up?: sentiment classification using machine learning techniques. *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, EMNLP '02. Stroudsburg, PA, USA: Association for Computational Linguistics, pp.79–86.

Pariser, E., 2011. *The filter bubble: what the internet is hiding from you*. Penguin Press.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E., 2011. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12, pp.2825–2830.

Peeters, G. and Czapinski, J., 1990. Positive-negative asymmetry in evaluations: the distinction between affective and informational negativity effects. *European Review of Social Psychology*, 1(1), pp.33–60.

Petty, R.E. and Cacioppo, J.T., 1988. Communication and persuasion: central and peripheral routes to persuasion. *The Public Opinion Quarterly*, 52(2), pp.262–265.

Philander, K. and Zhong, Y., 2016. Twitter sentiment analysis: capturing sentiment from integrated resort tweets. *International Journal of Hospitality Management*, 55, pp.16–24.

Ponnusamy, R., Degife, W.A. and Alemu, T., 2018. Recommender frameworks outline system design and strategies: a review. *Knowledge Computing and its Applications*. Singapore: Springer Singapore, pp.261–285.

Pozzi, G., Pigni, F. and Vitari, C., 2014. Affordance theory in the IS discipline: a review and synthesis of the literature. *Proceedings of the Twentieth Americas Conference on Information Systems.*

Prasetyo, P.K., Lo, D., Achananuparp, P., Tian, Y. and Lim, E.-P., 2012. Automatic classification of software related microblogs. *Proceedings of the 28th International Conference on Software Maintenance*, pp.596–599.

Priester, J.R. and Petty, R.E., 2003. The influence of spokesperson trustworthiness on message elaboration, attitude strength, and advertising effectiveness. *Journal of Consumer Psychology*, 13(4), pp.408–421.

Prior, M., 2013. Media and political polarization. *Annual Review of Political Science*, 16(1), pp.101–127.

Prusa, J.D. and Khoshgoftaar, T.M., 2017. Deep neural network architecture for character-level learning on short text. *Proceedings of the Thirtieth International Florida Artificial Intelligence Research Society Conference*, pp.353–358.

Quick, B.L. and Considine, J.R., 2008. Examining the use of forceful language when designing exercise persuasive messages for adults: a test of conceptualizing reactance arousal as a two-step process. en. *Health Communication*, 23(5), pp.483–491.

R Core Team, 2019. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria.

Rains, S.A., 2013. The nature of psychological reactance revisited: a meta-analytic review. *Human Communication Research*, 39(1), pp.47–73.

Razavi, A.H., Inkpen, D., Uritsky, S. and Matwin, S., 2010. Offensive language detection using multi-level classification. *Proceedings of the 23rd Canadian Conference of Artificial Intelligence*, pp.16–27.

Rejón-Guardia, F. and Martínez-López, F.J., 2014. Online advertising intrusiveness and consumers' avoidance behaviors. In: Martínez-López, F.J. ed. *Handbook of strategic e-business management*, pp.565–586.

Reviglio, U., 2017. Serendipity by design? how to turn from diversity exposure to diversity experience to face filter bubbles in social media. In: Kompatsiaris, I., Cave, J., Satsiou, A., Carle, G., Passani, A., Kontopoulos, E., Diplaris, S. and McMillan, D. eds. *Proceedings of the 4th International Conference on Internet Science*. Cham: Springer International Publishing, pp.281–300.

Reviglio, U., 2019. Serendipity as an emerging design principle of the infosphere: challenges and opportunities. *Ethics and Information Technology*, 21(2), pp.151–166.

Ritter, A., Clark, S., Mausam and Etzioni, O., 2011. Named entity recognition in tweets: an experimental study. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. ACL, pp.1524–1534.

Saito, T. and Rehmsmeier, M., 2015. The recision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, 10(3), pp.1–21.

Salton, G. and Yang, C.-S., 1973. On the specification of term values in automatic indexing. *Journal of Documentation*, 29(4), pp.351–372.

Sang, E.F.T.K. and De Meulder, F., 2003. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. *Proceedings of the Seventh Conference on Natural Language Learning*. Vol. 4, pp.142–147.

Sasidharan, L. and Menéndez, M., 2014. Partial proportional odds model—an alternate choice for analyzing pedestrian crash injury severities. *Accident Analysis and Prevention*, 72, pp.330–340.

Sato, T., 2015. *Neologism dictionary based on the language resources on the web for MeCab.*

Scholastic, 2019. *Kids & family reading report: 7th edition*, (technical report).

Scott, K., 2015. The pragmatics of hashtags: inference and conversational style on Twitter. *Journal of Pragmatics*, 81, pp.8–20.

Searle, J., 1975. A taxonomy of illocutionary acts. *Language, Mind, and Knowledge*, 7, pp.344–369.

Sekine, S. and Nobata, C., 2004. Definition, dictionaries and tagger for extended named entity hierarchy. *Proceedings of the 11th International Conference on Language Resources and Evaluation*, pp.1977–1980.

Sennrich, R., Haddow, B. and Birch, A., 2016. Improving neural machine translation models with monolingual data. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, pp.86–96.

Severyn, A. and Moschitti, A., 2015. UNITN : Training Deep Convolutional Neural Network for Twitter Sentiment Classification. *Proceedings of the 9th International Workshop on Semantic Evaluation*, pp.464–469.

Sharma, M. and Mann, S., 2013. A survey of recommender systems: approaches and limitations. *International Journal of Innovations in Engineering and Technology*, pp.1–9.

Sikora, J., Evans, M.D. and Kelley, J., 2019. Scholarly culture: how books in adolescence enhance adult literacy, numeracy and technology skills in 31 societies. *Social Science Research*, 77, pp.1–15.

Smith, S.L., Kindermans, P.-J., Ying, C. and Le, Q.V., 2018. Don't decay the learning rate, increase the batch size. *Proceedings of the Sixth International Conference on Learning Representations*.

Spearman, C., 1904. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1), p.72.

Standing, C., Holzweber, M. and Mattsson, J., 2016. Exploring emotional expressions in e-word-of-mouth from online communities. *Information Processing & Management*, 52(5), pp.721–732.

Statistics Bureau of Japan, 2016. *平成 28 年 社会生活基本調査 [Domestic survey of social life in 2016; in Japanese]*, (technical report).

Steede, G.M., Meyers, C., Li, N., Irlbeck, E. and Gearhart, S., 2018. A sentiment and content analysis of Twitter content regarding the use of antibiotics in livestock. *Journal of Applied Communications*, 102(4), pp.1–16.

Steffes, E.M. and Burgee, L.E., 2009. Social ties and online word of mouth. *Internet Research*, 19(1), pp.42–59.

Steindl, C., Jonas, E., Sittenthaler, S., Traut-Mattausch, E. and Greenberg, J., 2015. Understanding psychological reactance: new developments and findings. *Zeitschrift fur Psychologie / Journal of Psychology*, 223(4), pp.205–214.

Strauss, B., Toma, B.E., Ritter, A., De Marneffe, M.-C. and Xu, W., 2016. Results of the WNUT16 named entity recognition shared task. *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pp.138–144.

Sullivan, A. and Brown, M., 2015. Reading for pleasure and progress in vocabulary and mathematics. *British Educational Research Journal*, 41(6), pp.971–991.

Takeshi, S., Sakae, M. and Naoyuki, G., 2019. BERT pre-trained model trained on large-scale Japanese social media corpus. *Github: hottolink/hottoSNS-bert*.

Tang, Y.-j. and Chen, H.-H., 2012. Mining sentiment words from microblogs for predicting writer-reader emotion transition. *Proceedings of the Eight International Conference on Language Resources and Evaluation*, pp.1226–1229.

Tavakol, M. and Dennick, R., 2011. Making sense of Cronbach's alpha. eng. *International journal of medical education*, 2, pp.53–55.

Tayyar Madabushi, H., Kochkina, E. and Castelle, M., 2019. Cost-sensitive BERT for generalisable sentence classification on imbalanced data. *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*. Stroudsburg, PA, USA: Association for Computational Linguistics, pp.125–134.

Tuarob, S. and Mitrpanont, J.L., 2017. Automatic Discovery of Abusive Thai Language Usages in Social Networks. In: Choemprayong, S., Crestani, F. and Cunningham, S.J. eds. *Digital Libraries: Data, Information, and Knowledge for Digital Lives*. Vol. 10647, Lecture Notes in Computer Science, pp.267–278.

Tuarob, S., Tucker, C.S., Salathe, M. and Ram, N., 2014. An ensemble heterogeneous classification methodology for discovering health-related knowledge in social media messages. *Journal of Biomedical Informatics*, 49, pp.255–268.

Tutz, G. and Hennevogl, W., 1996. Random effects in ordinal regression models. *Computational Statistics and Data Analysis*, 22(5), pp.537–557.

Twenge, J.M., Martin, G.N. and Spitzberg, B.H., 2019. Trends in U.S. adolescents' media use, 1976–2016: the rise of digital media, the decline of TV, and the (near) demise of print. *Psychology of Popular Media Culture*, 8(4), pp.329–345.

UNESCO, 1965. Recommendation concerning the international standardization of statistics relating to book production and periodicals. *Records of the General Conference, thirteenth session, Paris, 1964: Resolutions*, pp.143–147.

Uzawa, H., 2005. *Economic analysis of social common capital.* Cambridge University Press, p.406.

Uzawa, H., 2009. Lecture "Social Common Capital". *2009 Blue Planet Prize Commemorative Lectures.* The Asahi Glass Foundation. The Asahi Glass Foundation, pp.6–21.

Uzawa, H., 2010. Social common capital, imputed price, and sustainable development. *Macroeconomic Dynamics*, 14(2), pp.149–165.

Van de Kauter, M., Breesch, D. and Hoste, V., 2015. Fine-grained analysis of explicit and implicit sentiment in financial news articles. *Expert Systems with Applications*, 42(11).

van den Berg, J.A., 2014. The story of the hashtag(#): A practical theological tracing of the hashtag(#) symbol on Twitter. *HTS Teologiese Studies / Theological Studies*, 70(1), pp.1–6.

van Bergen, E., van Zuijen, T., Bishop, D. and de Jong, P.F., 2017. Why are home literacy environment and children's reading skills associated? What parental skills reveal. *Reading Research Quarterly*, 52(2), pp.147–160.

Varol, O., Ferrara, E., Davis, C., Menczer, F. and Flammini, A., 2017. Online human-bot interactions: detection, estimation, and characterization. *International AAAI Conference on Web and Social Media*, pp.280–289.

Vehtari, A., Gelman, A. and Gabry, J., 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5), pp.1413–1432.

Verma, M., Divya, D. and Sofat, S., 2014. Techniques to detect spammers in Twitter—A survey. *International Journal of Computer Applications*, 85(10), pp.27–32.

Vosoughi, S. and Roy, D., 2016. Tweet Acts: a speech act classifier for Twitter. *Proceedings of the 10th AAAI Conference on Weblogs and Social Media.* Cologne, Germany, pp.1–4.

Wade, S. and Kidd, C., 2019. The role of prior knowledge and curiosity in learning. *Psychonomic Bulletin & Review*, 26(4), pp.1377–1387.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O. and Bowman, S., 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Stroudsburg, PA, USA: Association for Computational Linguistics, pp.353–355.

Wang, A., 2010. Detecting spam bots in online social networking sites: a machine learning approach. English. *Proceedings of the 24th IFIP Annual Conference on Data and Applications Security and Privacy*, pp.335–342.

Wang, R., Zhou, D., Jiang, M., Si, J. and Yang, Y., 2019. A survey on opinion mining: from stance to product aspect. *IEEE Access*, 7, pp.41101–41124.

Waxman, L., Clemons, S., Banning, J. and McKelfresh, D., 2007. The library as place: Providing students with opportunities for socialization, relaxation, and restoration. *New Library World*, 108(9-10), pp.424–434.

Wei, J. and Zou, K., 2019. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Stroudsburg, PA, USA: Association for Computational Linguistics, pp.6383–6389.

Weiser, P., Bucher, D., Cellina, F. and De Luca, V., 2015. A taxonomy of motivational affordances for meaningful gamified and persuasive technologies. *Proceedings of EnviroInfo and ICT for Sustainability 2015*. Paris, France: Atlantis Press.

Weng, R. and Takaku, M., 2019. Book recommender system using linked data for improving serendipity. *Journal of Japan Society of Information and Knowledge*, 29(2), pp.164–169.

Willis, C. and Efron, M., 2013. Finding information in books: characteristics of full-text searches in a collection of 10 million books. *Proceedings of the 76th ASIS&T Annual Meeting*, pp.1–10.

Wojdynski, B.W., Evans, N.J. and Hoy, M.G., 2018. Measuring sponsorship transparency in the age of native advertising. *Journal of Consumer Affairs*, 52(1), pp.115–137.

Wojdynski, B.W. and Golan, G.J., 2016. Native advertising and the future of mass communication. *American Behavioral Scientist*, 60(12), pp.1403–1407.

Wolf, M., 2008. *Proust and the squid: the story and science of the reading brain.* Harper Perennial.

Wu, M., Scholer, F. and Thom, J.A., 2009. The Impact of Query Length and Document Length on Book Search Effectiveness. *Proceedings of the International Workshop of the Initiative for the Evaluation of XML Retrieval*, pp.172–178.

Wu, T., Wen, S., Xiang, Y. and Zhou, W., 2018. Twitter spam detection: survey of new approaches and comparative study. *Computers and Security*, 76, pp.265–284.

Yada, S., 2014. Development of a book recommendation system to inspire 'infrequent readers'. In: Tuamsuk, K., Jatowt, A. and Rasmussen, E. eds. *The Emergence of Digital Libraries – Research and Practices.* Vol. 8839, LNCS. Springer International Publishing, pp.399–404.

Yada, S., 2019. *Tweets that mention books 2015.* Mendeley Data.

Yada, S. and Kageura, K., 2015. Identification of Tweets that Mention Books: An Experimental Comparison of Machine Learning Methods. In: Allen, R.B., Hunter, J. and Zeng, M.L. eds. *Digital Libraries: Providing Quality Information.* Vol. 9469, Lecture Notes in Computer Science. Cham: Springer International Publishing, pp.278–288.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R. and Le, Q.V., 2019. XLNet: generalized autoregressive pretraining for language understanding. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. and Garnett, R. eds. *Advances in Neural Information Processing Systems.* Vol. 32. Curran Associates, Inc., pp.5754–5764.

Zajonc, R.B., 1968. Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology*, 9(2, Pt.2), pp.1–27.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N. and Kumar, R., 2019. Predicting the Type and Target of Offensive Posts in Social Media. *Proceedings of the 2019 Conference of the North.* Stroudsburg, PA, USA: Association for Computational Linguistics, pp.1415–1420.

Zhang, R., Li, W., Gao, D. and You, O., 2013. Automatic Twitter topic summarization with speech acts. *IEEE Transactions on Audio, Speech, and Language Processing*, 21, pp.649–658.

Zhang, X., Zhao, J. and LeCun, Y., 2015. Character-level convolutional networks for text classification. *Proceedings of the 28th International Conference on Neural Information Processing Systems*, pp.649–657.

Zimbra, D., Abbasi, A., Zeng, D. and Chen, H., 2018. The state-of-the-art in Twitter sentiment analysis: a review and benchmark evaluation. *ACM Transactions on Management Information Systems*, 9(2), 5:1–29.