University of Wolverhampton

7LN009 Dissertation

Blue or Yellow?

An Investigation on Hong Kong Political Short Texts

Hei Man, Haley FONG

Supervisor: Dr. Le An Ha

February 2021

Dedication

To my family, especially my mother for encouraging me to empower myself with skills to survive in a fast-changing world. Thank you for supporting me fully and unconditionally on my path towards learning languages and studying overseas.

To my dear friends, thank you for being with me through ups and downs, even across borders and time zones. You have been crucial for keeping me company through this great, bumpy adventure. I love you all.

Finally, to Hong Kong, my hometown that inspired me to work on this dissertation. Let us hope that dark times pass and the city regains its light. Glory to Hong Kong.

Acknowledgements

First of all, I should express my gratitude towards all my instructors in the MA Computational Linguistics. It is a pleasure to have Dr. Ha as my supervisor, who provided me with professional advice and has stayed by my side for the past couple of months. I would especially like to show my appreciation towards Dr. Oakes, the programme leader for his care for students. In addition, I am delighted to have met Jiani Fan, my classmate and floormate who motivated me with thought-provoking discussions, as well as delicious and homey food while we were on campus.

On the other hand, I am thankful to have the opportunity to learn about language technologies as an undergraduate, which gave me a solid foundation to pursue this MA programme. I am grateful to have met Dr. Caesar Lun, who offered me insights towards possible paths of a linguistic graduate. I am also pleased to have sat in Dr. Chun Yu Kit and Dr. John Lee' s lectures during my gap months to deepen my knowledge in various applications of language technologies. Dr. Kit gave me a big push in continuing my study in the field of Computational Linguistics, and it is my pleasure to be Dr. Lee's research assistant upon completing this MA programme. Meanwhile, I should take this opportunity to show my gratitude to Phoebe Lam, my senior who led my way towards linguistics. Thank you for sharing your inspiring journey of studying languages. I would not have reached where I am today without you.

Last but not least, I should acknowledge Dr. Pamela Leung, my aunt who offered me critical advice and assisted me in collecting quality Cantonese-related references while completing this dissertation.

Abstract

The Internet has been providing numerous platforms for people to express their views for decades. While the platforms were intended to facilitate communication, the human nature of preference over like-minds and advancement of SNS content suggesting algorithms seemed to have created echo-chambers that separated people into different, often opposing, opinion groups. Recent years, severe real-life effects of such online phenomenon have been coming into sight. In the context of Hong Kong, such effects have been increasingly evident through the past few years.

This dissertation aims to look into such political stratification in Hong Kong from a computational linguistics angle, which is, to study how people holding different political opinions write online using computational means. Beginning with a literature review on Cantonese as a distinctive Chinese variety and opinion mining, a survey on the state-of-the-art text classification technologies was conducted. This was followed by an investigation on Hong Kong online discussion forums, namely Discuss.com and LIHKG. Around 15000 forum comments revolving recent, local affairs were crawled and classified with Naive Bayes, LSTM, CNN, XLMRoBERTA and XLNet models, incorporating conventional Chinese NLP resources and innovative Cantonese materials. Although the deep learning models did not outperform the baseline traditional machine learning models in classifying the comments, the baselines still offered plenty of insights on how netizens from respective forums write. Through learning the limitations and proposing improvements for this dissertation, it is suggested that better versions of the project could be applied to identifying varied ideas expressed in political short texts. By incorporating such technology into content suggesting algorithms, it could be possible to regulate opinionated content that are being fed to users, eventually melting the walls of echo chambers and restoring the vision of uniting people with the Internet.

Contents

1 Introduction

With the increasing popularity of the Internet as a platform for people to express themselves, user-created content has been gaining worth with its potential to reflect people' s attitudes. For instance, companies could make use of customer reviews to improve their products. It is undoubted that there is a vast amount of such information on the Internet, to the extent that manpower is insufficient to study the entire mass of user content, not to mention the fact that people could easily append new content to the collection of existing content. Computational efforts and artificial intelligence has been progressively employed to investigate online content. Opinion mining is a form of text classification, which could be automated using machine learning or deep learning techniques. In addition to its application in the commercial world, it could come in handy in academia in terms of social and political studies.

As suggested by Dalal & Zaveri (2011), text classification in regional languages should be further researched. While both English and standard Chinese are widely researched languages, along with the long-complained lack of data in Cantonese, studying online discussions in the variety could be significant to both academia and its users. Despite Cantonese being a majorly spoken variety, it has in fact been written more and more, especially in informal texts on the Internet. Meanwhile, since the 2014 Umbrella Movement, polarized political views have been observed in Hong Kong. The situation has escalated through 2019 with the Extradition Law Amendment Bill protests, and numerous events in 2020, including the coronavirus pandemic crisis and the introduction of the National Security Law. Generally speaking, *yellow ribbons* and *blue ribbons* are used to label those who support and oppose the government respectively. Back in the 2014, yellow ribbons were used by Umbrella

Movement supporters, and blue ribbons, in the same colour of the local police uniform, were used by those who supported the police to restore social order.[1] Inspired by TaDay's observations [2] on LIHKG and Hong Kong Discuss Forum (referred as Discuss in the following), as well as Liu (2015)' s analysis on a student-government discourse during the Umbrella Movement, this dissertation aims to perform automatic text classification with written Cantonese data mined from the forums, with the goal to provide insights concerning how people at opposite ends of the political spectrum write online.

In fact, looking into demographics of government supporters and opposers, it could be hypothesized that people of different stances write in dissimilar manners as well. From a sociolinguistic point of view, factors such as age, gender, education level and birthplace are related to one' s language use. According to the Hong Kong Public Opinion Research Institute' s Survey on the 2019 political movement (Chung 2019), it was revealed that yellow ribbons tended to be younger and were recipients of higher education. When looking into detailed opinions held by the respondents, it was learnt that there was a high degree of distrust towards the Chinese government among yellow ribbons. With the strong belief that "Hong Kong is not China" , yellow ribbons have been striving to strengthen the HongKonger identity through various means. Advocating the use of Cantonese –for instance, writing in the variety and boycotting "Mainland Chinese vocabulary" –have been one of the linguistic trends noticed among the group. Hence, to be specific, it is expected that netizens from the two forums make different lexical choices as they write.

[1]The Independent on coloured ribbons back in the 2014 Umbrella Movement `https://www.independent.co.uk/news/world/asia/hong-kong-protests-a-guide-to-yellow-ribbons-blue-ribbons-and-all-the-other-colours-9775324.html`

[2]TaDay is a company under GLOs, founded by Dr. Simon Shen, a recognized expert in international affairs in Hong Kong. TaDay has been advocating the use of big data for daily and research purposes, especially for analysing local and global social affairs. Articles and findings of the company could be found at `https://www.thestandnews.com/author/taday-glos/`

In order to uncover such differences, this dissertation would be experimenting with traditional text classification solutions and innovative standard Chinese and Cantonese natural language processing resources. The next chapters would first include a detailed literature review, followed by a comprehensive introduction of the experiments to be carried out, and finally an in-depth analysis on the experimental results. The literature review would begin with defining the Cantonese language, and move on to take a brief look into the dividing online communities, eventually landing onto the survey of text classification techniques. The subsequent chapter then presents the methodology of this dissertation, including the data being exploited, investigation procedures, as well as the baseline and experimental models. The computational outcome, error analysis and a thorough reflection on the limitations of the experiments would be given in another chapter.

2 Literature Review

2.1 Hong Kong Cantonese

Cantonese has long been a spoken variety of Chinese in Southern coastal China. It could broadly refer to the collection of dialects spoken in Guangdong and Guangxi areas, as well as any dialect spoken in these areas (Groves 2010). It is a major variety spoken by Chinese communities overseas in North America, Europe and Australasia too (Bauer & Matthews 2003). The debate on whether Cantonese is a language or a dialect has been ongoing for decades. Bauer (1988) pointed out that the Hong Kong community regarded Cantonese as a dialect because it was not supposed to be written like standard Chinese, which in the opposite, was treated as a language. Returning to Groves (2010)' s work, from a linguist' s perspective, mutual intelligibility is the key towards telling languages and dialects apart. It was reckoned that Chinese varieties are similar to Romance languages. However with political consideration, plus the fact that there is a standard script for Chinese, while being partially mutually intelligible, various Chinese varieties should be labelled as *topolects* (the speech variety of an area) or *regiolects* (varieties with mixed standard language and dialectal features), as suggested by Groves (2010).

This section aims to provide a brief introduction of Cantonese, with a focus on its written form, along with an overview of the variety in the context of Hong Kong.

2.1.1 Defining Cantonese

First of all, Chinese could be divided into standard and non-standard forms. Also, spoken and written forms should be investigated separately. The standard written form for Chinese nowadays is *baihua* (in contrast with *wenyan*, Classical Chinese), which could be written in either traditional or simplified characters. Characters' pronunciations vary across different spoken varieties. For standard spoken Chinese – *Putonghua*, which literally means common language – is based on the pronunciation and lexicon of Mandarin, the collection of northern Chinese dialectal varieties. In short, standard Chinese today uses the orthography of simplified Chinese characters, with a pronunciation and lexicon based on Mandarin, and employs the grammar of *baihua*. To avoid the language-dialect controversy, Cantonese and standard Chinese (Putonghua) would be treated as two *varieties* in the rest of this dissertation.

Differences in Cantonese and standard Chinese at several levels of language contributed to why the two varieties are mutually unintelligible. On the phonological level, Cantonese has more tones than standard Chinese – the former having six and the latter having four. What is more, Cantonese is one of the Chinese varieties that retains the syllable-final, unreleased stops (specifically *-p, -t, -k* finals, which are often considered as three stopped tones in additional to the six, resulting in nine tones in total) in their pronunciation. The stopped tone was found in medieval Chinese. For instance, the *Qieyun* rhyme dictionary written during medieval Chinese documented such pronunciation, but it no longer exists in modern Mandarin. (Norman 2003) This is one of the indicators of Cantonese being older than present standard Chinese. There are other minor phonological differences between the two languages, such as consonants, vowels and syllable structure.

On the lexical level, Cantonese is said to be heavily influenced by English, as reflected in loanwords and the norm of code-mixing with English. It was taken note that as a major variety in the Southeastern coast of China, Cantonese had contacted European languages as early as of the mid 16th Century (Bauer 1988). With Hong Kong becoming a British colony in 1841, English interacted with Cantonese vigorously through colonial administration. The development of the city into an international trade hub created the optimal environment for English to thrive in the community and continue influencing the Cantonese lexicon even after the 1997 handover. Furthermore, Bauer & Matthews (2003) highlighted that English words borrowed into Cantonese are "so assimilated that they have written forms with Chinese characters" and that Cantonese speakers have taken those words for granted as Chinese words. At the same time, numerous words in Cantonese cannot be written into Chinese characters and some do not have semantically identical words in standard Chinese. It is possible that such words hold roots in non-Sinitic languages (Bauer 2018).

On the grammatical level, as brought up by Yue (2003), Chinese varieties could be divided into three main groups – Northern, Central and Southern. The groups could be put into a scale where Northern and Southern varieties show prominent discrepancies and Central varieties show characteristics from both ends on the scale. The most noticeable difference between standard Chinese (based on Northern Chinese varieties) and Cantonese (a Southern Chinese variety) is that they do not share the same closed class words (as well as characters). For example, Cantonese has more aspect markers, classifiers (Yue 2003) and utterance-final particles (which encode speakers' attitudes or emotions) (Bauer & Matthews 2003) than that of standard Chinese. The perfective and past aspects could be disambiguated using suffixes in Cantonese but not standard Chinese. Looking into sentence structures, for instance, passive and comparative constructions are not parallel in the two varieties. Word ordering and the presence of certain sentence components could make an utterance grammatical in a variety but not the other.

2.1.2 Cantonese in Hong Kong

According to a survey by the Census and Statistics Department of the Hong Kong government on language use, Cantonese has still been widely used in the local community. Among roughly 5.6 million respondents, over 90% reported that they must or often use Cantonese in daily communication, for entertainment and at work (Census And Statistics Department 2019). A relatively small-scale study on students in a local university showed that local students felt positively about Cantonese (Liu 2018). Yet, despite efforts to advocate biliteracy (in written Chinese and English) and trilingualism (Cantonese, English and Putonghua), Cantonese in fact received the least attention from the local government (Lee & Leung 2012). To be specific, the least funding was utilized to promote better Cantonese pronunciation and that Cantonese linguistics was never taught in schools. Most importantly, Lee & Leung (2012) pointed out that the government did not clearly define what "written Chinese" should people be "biliterate" in. As stated by Snow (2013), schoolchildren are taught to write in standard Chinese.

In spite of passive institutional attitudes towards Cantonese, a boom of written Cantonese has been observed through the past decades. Snow (2013) attempted to draw a theory of vernacularization, in which a low variety takes over a high variety in a diglossic society, by investigating four cases – *Baihua* (today' s standard Chinese), written Wu, Cantonese and Minnan. Under the diglossic situation of written Chinese in Hong Kong – where the low variety is written Cantonese and the high variety is standard Chinese – although vernacularization is yet to be completed, there has been an obvious expansion of written Cantonese growing from a variety that is mainly used by the lower class to being used by the government (mainly on

social media and promotional materials, see fig) and studied by scholars (Snow 2004).

In addition, Snow (2008) explained how written Cantonese has gone on the path of standardization. It is clear that Cantonese is spoken in both daily and official contexts, indicating that Cantonese has an abundant vocabulary, not to mention users' norm of code-mixing with English. Since the 1980s, written Cantonese works in Hong Kong have been observed to be inclining to Cantonese, instead of standard Chinese, vocabulary and syntax. Despite the fact that there are no official or standard rules towards writing Cantonese, having the main assumption that the reader speaks Cantonese (Bauer 1988), as long as the written form corresponds to an utterance that sounds meaningful and grammatical in spoken Cantonese, it is an acceptable form. Meanwhile, Cantonese is an essential part of the *Hongkonger* identity. Being perceived as an "intimate and lively" variety, youngsters in particular, has been seen to write Cantonese more frequently, especially on the Internet. Throughout the past few decades, written Cantonese has expanded from sub-cultural products to more mainstream domains including advertisement and entertainment. This eventually leads to the gradual use of written Cantonese in promotional materials even by the local government too. Even if the variety does not seem to have grown into a regional standard of the Cantonese-speaking area of China, it has undoubtedly taken more and more roles in the Hong Kong society.

Figure 2.1: *news.gov.hk* Facebook page, with Cantonese expressions highlighted

Figure 2.2: *news.gov.hk* web page, with standard Chinese expressions highlighted

In short, regardless of the fact that Cantonese is not taught to be written in schools nor intensively promoted by government policy, it is surprising to see the variety spread from spoken to written domains of language use, solely by speakers' effort. The following section further explores how is Cantonese written in Hong Kong.

2.1.3 Written Cantonese

Having said that the Cantonese vocabulary largely overlaps with standard Chinese' s (Snow 2008), and that it is written in Chinese characters, there are in fact unique patterns that could guide one in distinguishing written Cantonese and standard written Chinese. Bauer (1988)' s pioneering work in drawing orthographic conventions of written Cantonese inspired a number of researchers to study the language. Recently, Bauer (2018) better concluded his observations, saying that Cantonese is written by following 12 principles which could be further summarized into 5 processes.

A piece of written Cantonese text should contain at least one Cantonese lexical item, and is targeted to a Cantonese-speaking audience (Bauer 2018). To begin with, the Cantonese lexicon could be divided into 3 layers – the first that overlaps with standard Chinese, the second that is natively, exclusively Cantonese and the third containing English loanwords. Cantonese words can be written using Chinese characters by utilizing either of the processes – adhering to traditional norms, phoneticization, indigenization or semanticization – or by English alphabets through alphabetization. The table on the following page summarizes Bauer (2018)' s work.

Process	Principle	Example(s) in Written Cantonese	Pronunciation in Cantonese	Meaning in English	Corresponding expression in standard Chinese
Traditional usage of standard Chinese characters		香港	*heong1 gong2*	Hong Kong	香港 *xiānggǎng*
Indigenization of Chinese characters	Using standard Chinese characters in a non-standard way	飲	*jam2*	to drink	喝 *hē*
	Reviving old Chinese characters	搵	*wan2*	to find	找 *zhǎo*
	Using non-standard Chinese characters	梘	*gaan2*	soap	碱 *jiǎn*
	Creating Cantonese-only characters	佢嘅 and 嘢	*koei5 ge3* and *mak1*	his and mug/(trade)mark	他的 *tāde* and 馬克杯 *mǎkebēi*/ 商標 *shāngbiāo*
Borrowing standard Chinese characters to phonetically transcribe…	… Cantonese words	錫	*sek3*	to kiss	吻 *wěn* （錫 is a kind of metal in standard Chinese）
	… English words	士多	*si6 do1*	store	小店 *xiǎodiàn*
Semanticization of standard Chinese characters	Reading Chinese characters for their meanings but not pronunciations	泊(車)	*paak3 (ce1)*	to park (a car)	停(車) *tíng(chē)* （泊 means the anchoring of boats in standard Chinese）
Alphabeticizion: Using English alphabets to …	Ad hoc romanization	Hea	*he3*	(to be) lazy	懶 *lǎn*
	… represent Cantonese morphosyllables	D（啲 in Cantonese character）	*di1*	some	些 *xiē*
	… write English loanwords	M 巾	*em1 gan1*	Menstrual pads	衛生巾 *wèishēngjīn*
	Retaining English words and their spellings and reading with Cantonese syllables	Print	*pin1*	to print	打印 *dǎyìn*

While the above framework facilitates the identification of Cantonese texts, the root issue for not having a standard for writing the variety could give rise to problems in natural language processing. As a result, heavy data-cleaning could be expected. For instance, when multiple characters transcribe the same morphosyllable, normalization is required to maintain one-to-one character-morphosyllable correspondence. At the same time, if a character has multiple pronunciations and is used to represent different morphosyllables, it is crucial to keep track of the collocations of characters for effective pronunciation and meaning disambiguation. Finally, if there is no suitable character for a morphosyllable, it would be left as an empty box. This could make written Cantonese data error-prone if it is not preprocessed with suitable substitutes, for example romanizing the morphosyllable with English alphabets. However, biases could be introduced if there are more than one grammatical morphosyllables that could fit into the empty box, distorting the original, intended message.

2.2 Social Networking Sites (SNS) and Opinion Polarization

As mentioned, diversifying political opinions have been observed in Hong Kong since the 2014 Umbrella Movement. While large scale political movements seemed to have died down the following few years, the 2019 protests have brought opinion diversion to another level. SNS is believed to be an optimal environment for the process, allowing people to easily give and take ideas and information. Discourses on SNS are usually informal and speech-like. In Hong Kong' s context, Cantonese is the prominent language of online discourses. This section dives into how opinions become polarized online, supplemented with related studies of Hong Kong' s divided opinions on SNS, along with investigation of the issue from linguistic perspectives.

2.2.1 Online Opinion Polarization

Studying Facebook pages on the Umbrella Movement, Chan & Fu (2017) concluded from previous works that interest groups are formed and divided in the virtual space with information technology. Such diversion is fuelled mainly by selective exposure, followed with social comparison hence reinforcement. Kobayashi & Ikeda (2009) explored the effect of selective political web browsing on democratic development. It was discovered that the prerequisite of selective exposure to homogenous views is a strong personal tie of an individual and the concerned issue, as the Internet provides equal tendency of one' s exposure to homogenous and heterogenous information. On the other hand, Banisch & Olbrich (2019) drew a reinforcement learning mechanism, showing that with positive feedback from people alike, one would be more convinced to believe in a certain opinion, eventually giving rise to polarized stances.

From the political viewpoint, Kobayashi (2012) worried that opinion polarization on SNS would fragment social reality as people become more reluctant to opposing attitudes, which could threaten democratic development. Edwards (2013) tracked netizens' activity on a forum for over a year, confirming the tendency for online political discussions to go homogenous and polarized. The cycle of homogenization and polarization results in "echo chambers" of extreme opinion. In addition to Kobayashi & Ikeda (2009)' s study on voluntary selective exposure, Matakos et al. (2017) pointed out that SNS algorithms created "filter bubbles" . As a result, SNS users are unconsciously fed with homogenous information, which could further consolidate their viewpoints. Proposing a quantitative measure for polarization, the three agreed with Edwards (2013) that moderation is important in reducing polarization. With these said, the rise of polarized opinion groups could be dangerous, as seen in examples in latter parts of this section.

2.2.2 SNS and Hong Kong Political Movements

Back in 2014, cyberbalkanization – the fragmentation of the online community – has been noticed in Hong Kong. Studying data from forums (UWant and Discuss) and Facebook pages, Lee, Liang & Tang (2019) suggested that online incivility, through the use of swear words, fans the fire of polarization. The amount of uncivil comments increases with that of political discussion. Prolonged, uncivil exchange of political comments was said to bring mutual distrust and disrespect, resulting in balkanization. Chan & Fu (2017) looked specifically into Facebook pages, proposing that strong-tie interactions encourages one to take a more extreme side.

Forwarding to 2019, SNS undoubtedly played a heavy role in the prolonged protests. Thanks to platforms such as LIHKG and Telegram, a leader is no longer needed in the protests (Lee, Yuen, Tang & Cheng 2019) because participants expressed their views through voting and commenting on the platforms (Purbrick 2019). The echo chamber effect on the platforms helped mobilizing participants and spread latest action plans. The protests were even seen as an outcome of a "Radically Networked Society" by Kewalramani & Seth (2020), which is rooted from the society' s hyperactive online communities.

As investigated by Lee, Liang & Tang (2019), polarized political discourse holds certain linguistic features. The next section gives several examples on linguistic analysis on polarized opinion clusters on the Internet.

2.2.3 Linguistic Insights from Polarized Opinion Groups

Rho et al. (2018) studied political comments from three opinion groups on Facebook regarding the MeToo Movement. Generally speaking, those from the opposite side of the political spectrum in fact wrote similarly when compared to those from those who held a relatively neutral stance. It was common to spot cognitive shortcuts and heuristic cues. Besides, strong and negative words were prevalent. For instance swear words – echoing Lee, Liang & Tang (2019)' s work, as well as subcultural slangs (Jaki et al. 2019). The construction of in- and out- group identity, in which out-group instances are viewed negatively, was reported too. Bäck et al. (2018)' s longitudinal examination of an online xenophobic forum recorded an increase of collective pronouns for in- and out- group identification over time in an echo chamber.

At the same time, Rho et al. (2018) reckoned that texts could reveal how one perceives the world and at the same time construct one' s reality. The difference in the top tokens associated with the Movement proved that the Movement was presented differently in the lens of different viewpoints, which agreed with Kobayashi (2012)' s idea on fragmented reality. Assuming that different opinion groups have different ways to say things, it could be possible to back-trace one' s opinion through texts they produce. Specifically, by utilizing computational methods such as automatic text classification, one could tell which side is being favoured from a piece of text. Moving on, this could help detecting unwanted information, for example fake news and hate speech, thus preventing disastrous outcomes like violent attacks (Jaki et al. 2019). Meanwhile, the same technology is also useful for other purposes. Spam mail filtering, automated chatbots, and even identification of suicidal persons on SNS are some of the uses of automatic text classification technologies. The following section further introduces how the task is performed by computers.

2.3 Text Classification

Text classification could be performed by computers in numerous ways. The essence of the process is to have the machines learn and identify patterns that determine why a text belongs to a group but not another. Basic procedures involve text collection, text preprocessing, feature selection, machine learning and evaluation. From Dalal & Zaveri (2011)' s work to Mirończuk & Protasiewicz (2018)' s work, solutions to perfecting automatic text classification have bloomed in all aspects. Not only have techniques in various steps in the automatic text classification workflow improved, it is now possible for machines to cluster texts into classes instead of having researchers provide predefined classes before classification. This section first gives an overview of what is involved in automatic text classification, following its application in classifying short, political texts, highlighting how n-grams are utilized during classification. Further attention would be given to how Chinese texts are machine-classified, supplemented with related studies on standard Chinese and Cantonese short text classification.

First and foremost, data is crucial, and it could be acquired through established sources (such as existing corpora or databases) or manually collected. In order to facilitate machine learning, the data should be preprocessed by removing noise and inconsistency, leaving a considerably representative size of training material for the machine. Common techniques in pre-processing include tokenization, sentence boundary detection and stopword elimination. In addition, for inflectional languages, lemmatization is carried out to further simplify words into their root forms. According to Dalal & Zaveri (2011), effective stemming could shrink the vocabulary of training data to about three times less compared to that of raw data. It was also reminded that script tags should be taken care of when dealing with web documents.

Upon pre-processing, the text data, through feature construction, is further simplified into numbers (Grimmer & Stewart 2013). With reference to Mirończuk & Protasiewicz (2018), features could be constructed from different levels of language. At the lexical level, simple features are drawn from n-grams. Domain specific features for example emojis could be studied too. At the semantic level, n-gram features could be grouped with respect to taxonomies or ontologies (for instance, with the help of WordNet). Embedded features and simple semantic features like named entities give more insights towards classifying texts with the meaning they carry. At the discourse level, features could be built on top of topics of discussion. And last but not least, other non-textual metadata could be used as features as well, including when the text was written and by whom. With all features constructed, they could be encoded into numbers and finally be able to be learnt and processed by computers. Balancing the weights of the features, the text data could be mapped into graphs or vector space, making the data more reader friendly by visualizing it.

Having plenty of features to train with, only important features would be selected for efficient machine learning. This process is known as dimension reduction. Again, the importance of individual features could be measured in different ways. Older techniques tend to use statistical means, in which word frequencies imply importance. Dalal & Zaveri (2011)' s case of such means is term frequency-inverse document frequency (TFIDF). TFIDF looks into the frequency of a word in relation to the document and corpus it was found in. Meanwhile, semantics come into play in dimension reduction techniques such as latent semantic indexing. In this case, the meaning a word holds determines its contribution towards the result of which class a piece of text belongs to.

With an abundant amount of data, which is represented by the optimal number of features, it is time to advance to the machine training stage. In this stage, the machine would learn how to classify texts based on the given features. Texts could be classified into either human or machine-determined classes. Grimmer & Stewart (2013) suggested a number of methods to classify texts in both ways. For predetermined classes, dictionary methods find out keywords that make texts different from one another. The amount of certain keywords a text holds would then reflect to which class the text is inclined to be. Additionally, supervised machine learning methods, say Naive Bayes, are used to divide texts into known classes.

On the other hand, when further allowing computers do decide classes for the texts, clustering methods are employed instead. Clustering could be performed in a fully automated manner. To be specific, automatic clustering models are of either single or mixed memberships. Both models consist of a definition of distance between texts, an objective function and optimization algorithm. The K-means algorithm is a kind of single membership clustering, in which clusters could be mapped in a two-dimensional manner. Multi membership clustering takes more dimensions into consideration, for instance, topics of text. These models tend to arrange classes in a hierarchical structure. Still, human effort is needed to translate the clustering results, but with unsupervised learning, it is obvious that the insight-seeking process is highly sped up.

After machine learning, it is necessary to prove the results are legitimate through evaluation. Grimmer & Stewart (2013) stressed that validation is one of the principles when carrying out automatic text analysis. Different means should be adapted for supervised and unsupervised learning. The former is evaluated to be valid if the output is similar to that of human analysis. It takes more for the latter to be valid, in which the model should pass experimental, contextual and statistical tests, and is

able to demonstrate that it is as accurate as a supervised model.

2.3.1 Automatically Classifying Political Texts

In Grimmer & Stewart (2013)' s work, political texts refer to speeches, legislation, bills, treaties and agreement etc. It could also be texts whose topics revolve around politics. Nowadays, not only politicians and legislators produce political texts, any netizen can express their political views online, either through full-length blog or brief microblog posts. By studying the sentiment expressed in these user-created content, in political contexts, legislators could gain insights towards better policy-making as well as campaign promotion.

Melville et al. (2009) offered a framework, incorporating both knowledge-based and machine learning approaches for sentiment classification. Their work was not only tested on political blogs, but product and movie reviews too. The three noticed that solely utilizing lexical knowledge or labelled examples is insufficient in accurately classifying texts. The former method classifies texts by first compiling a lexicon of classified terms, then counting the frequency of those terms in a piece of text. The more terms of a certain group is present in a text, the text is likely to be classified as that group. Contrarily, the latter method refers to classic supervised machine learning from labelled data. In their experiment, Melville et al. (2009) compared classification results from lexical and Naive Bayes classifiers, a semi-supervised learnt model (referencing Liu et al. (2004)) and their own composite multinomial models that pooled results from lexical and training based models respectively, using 10-fold cross validation. The linear pooling model performed the best among the five models. Maximum accuracy was yielded from classifying product reviews, for a near-exclusive relationship between reviewer and stance was observed. However, Melville et al. (2009) reckoned that creativity in jokes or cultural references blurred

the line between positive and negative political opinion, hence reducing the accuracy of their model. All in all, with a background knowledge from a generic lexicon, less labelled training data is required. Additionally, with supervised machine learning, domain-specific terms and word use could be identified. As machines identify heavily-weighted words, it is possible to pick up sarcastic or rhetorical use of certain words.

The investigation carried out by Kato et al. (2008) illustrates how idiosyncratic language in online discourse hinders accurate text classification. Working with online, informal discourse, the team of researchers attempted to sort posts with traditional text-based approaches – pointwise mutual information sentiment analysis (following Turney (2002)) and Naive Bayes text classification – both failing to put texts into appropriate groups for nearly 50% of the time. Expanding on Melville et al. (2009)' s point on textual inconsistencies as an inhibitor to on-point text classification, Kato et al. (2008) listed a number of causes why words and their embedded meaning are not directly related, such as sarcasm and rhetorical expressions. Besides, they highlighted an array of difficulties in processing political short texts, which is highly due to the informal nature of internet discourse. This includes typos, out-of-dictionary political jargon and domain-specific slang, along with pointed respelling of terms. With the arbitrary tie between words and their intended meaning, the team brought author information into consideration. Analysing and making use of the social network of the netizens and their individual behaviour online, a near 70% accuracy was finally obtained with Naive Bayes and clustering helped to bring the accuracy over 70%.

While it seems that lexical and probabilistic approaches are rather naive in classifying texts, recent studies which took embedded meanings in words with neural networks yielded better classification accuracies. Rao & Spasojevic (2016) made use of word

embeddings and Long Short-Term Memory (LSTM) to classify Tweets into either Democratic or Republican. With the optimal number of embedding and LSTM layers, the model achieved near 90% accuracy in scoring a Tweet' s degree of inclination towards a Democratic or Republican view. Poria et al. (2015) meanwhile employed deep convolutional neural network (CNN) to determine the polarity of utterances in video clips. In addition to the textual features extracted (word embeddings and part-of-speech) from the utterances, the model considered visual (facial expression of the speaker) and audio features. The study introduced the mixed use of CNN and support vector machines (SVM), in which the former produced training features and the latter made decisions during classification.

2.3.2 N-grams

As discussed in Section 2.2.3, lexical choices of polarized opinion groups are distinctive. In automated text classification, lexical items are represented in terms of tokens. Multiple tokens form n-grams, from which one can observe how lexical items are used in different textual contexts, as well as how different lexical items are collocated. The use of n-grams in text classification is not new. Back in the 90s, Cavnar et al. (1994) utilized n-gram frequencies to cluster documents, showing that their approach could overcome textual errors, including misspelling and OCR fails. Melville et al. (2009)' s attempt with a labelled word list could also be considered as detecting the presence of certain unigrams while classifying texts. Comparing the results generated from several models, Tripathy et al. (2016) concluded that the larger the n-gram is, the less accurate the models were. Using 1 to 3-grams in isolation and combined gave the best performance. HaCohen-Kerner et al. (2017) suggested that n-gram features are efficient for short text opinion mining, for the presence and absence of n-grams are highly noticeable.

Having said that words are fundamental to text classification, for languages like English, it is easy to extract individual lexical items because they are separated with white spaces. On the contrary, Chinese texts do not share the same story, which adds extra workload in the preprocessing stage where texts are segmented into words. Although automatic word segmentation does not lack attention among researchers, solutions are yet to be perfect. In fact, without identical common knowledge, manual word segmentation does not guarantee consistency. According to Luo et al. (2012), errors from automatic word segmentation could seriously deteriorate NLP performance. Automatic segmentation based on n-gram length instead of meaning was preferred after manual segmentation, which does not require prior knowledge or dictionary data (Wei et al. 2008).

Various kinds of features could be extracted from n-grams. Wei et al. (2008) made use of absolute and relative frequencies of unigrams and bigrams in classifying texts in a Chinese corpus. In addition, n-grams could be studied in terms of syntax and semantics. Considering that Chinese words are made up of character n-grams, Luo et al. (2011) looked into how words of part-of-speech contribute in text classification. Meaning carried by n-grams could be exploited by using vector generators such as word2vec too. On the other hand, while n-grams could be calculated at the word and character level for languages like English, for the context of Chinese, Cao et al. (2018) put forward a stroke n-gram model, standing out from conventional character level research. Their model made use of the sub-components of a character, which yielded better results by capturing orthographic thus semantic cues. Using sub-character embeddings, Cao et al. (2018) moved on to test their model in performing text classification and named entity recognition. Their proposal outperformed traditional word embeddings.

2.3.3 Classifying Chinese Short Texts

Understanding N-grams' importance in Chinese text classification, this section further looks into how short texts in the language are classified.

Bringing up that conventional text classification methods might work the same on short texts, Wang et al. (2017) tested various feature selectors and classifiers on financial short texts. It was found that TFIDF or Count Vectorizer, combining SVM or logistic regression classifiers attained over 80% accuracy. For everyday short texts, Ng et al. (2018) worked on Weibo, the Chinese Twitter, in censor-prone content. A number of features were taken into calculation, including linguistic information (psychological processes and parts of speech) and automatically calculated sentiment scores. It was found that subjective and informal expressions contribute to censorship, rather than sentiment.

Again, neural networks could be used for short texts. CNN and LSTM (Bidirectional LTSM as well) are popular options, with the former dealing with feature extraction and the latter with sequential text input (Li et al. 2018). They are frequently used as baselines for evaluating deep learning models. For example, Liao et al. (2019) made use of all character, word and conceptual information with CNNs to extract features from each of the levels respectively. Their model gave highest accuracy in terms of feature extraction and classification. Chen et al. (2019) utilized a more powerful attention mechanism, considering concepts both encoded within a piece of short text and that of its ontological set. Their model outperformed simpler CNN and BiLSTM models in classifying texts according to emotion, polarity and topic.

Despite the scarce number of studies on processing Cantonese short texts, existing works still provide great hints towards this dissertation. For instance, Zhang et al.

(2011) employed Naive Bayes and SVM classifiers with uni- to trigram features to classify Cantonese restaurant reviews by sentiment. 3000 reviews were obtained from an online Hong Kong dining guide. To the author ' s surprise, the former model returned better results than the latter. It was agreed between the models that bi-grams were best at capturing sentiments. In contrast, Shen & Chow (2020) utilized multi-layer perceptron neural network on nearly 100 times more data collected from a popular local forum, LIHKG. Although their model was not designed to classify texts but to extract temporal and locational information, their up-to-date techniques applied to process Cantonese short texts could benefit this project.

To conclude, the table on the next page sums up the best models of various studies on standard Chinese (and Cantonese) short text classification.

Summary of works on standard Chinese and Cantonese short text classification

Work	Data Source	Data Size	Best Features	Best Algorithm	Accuracy
Ng et al. (2018)	Weibo	344 posts	17 features including linguistic features, sensitive keyword frequency, sentiment scores and 1-3-grams etc.	Naïve Bayes	79.34%
Wang et al. (2017)	Shanghai and Shenzhen Stock Exchanges	409,871 short text announcements	Word vectors from count and TFIDF vectorizers	Logistic Regression	84.22%
Li et al. (2018)	Stanford Sentiment Treebank and THUCNews	11,885 movie reviews and unknown	word2vec generated vectors	Deep learning (BiLSTM with CNN)	90.8%
Liao et al. (2019)	Various Chinese news websites	228,000 news headlines	Concept and keyword, word and character features	Neural network including CNN and BiLSTM layers with Microsoft concept graph system Probase+ for feature processing	82.24%
Chen et al. (2019)	1. Weibo 2. Product reviews 3. News titles 4. Sougou news	1. 4,936 posts 2. 9,998 reviews 3. 192,299 titles 4. 7,960 topics	Skip-gram character, word and concept embeddings	Proposed Short Text Classification with Knowledge Powered Attention (STCKA) model	1. 43.20% 2. 74.30% 3. 80.11% 4. 88.14%
Zhang et al. (2011)	Openrice	3,000 reviews	Bigrams	Naïve Bayes	95.67%

3 Methodology

3.1 Data Preparation

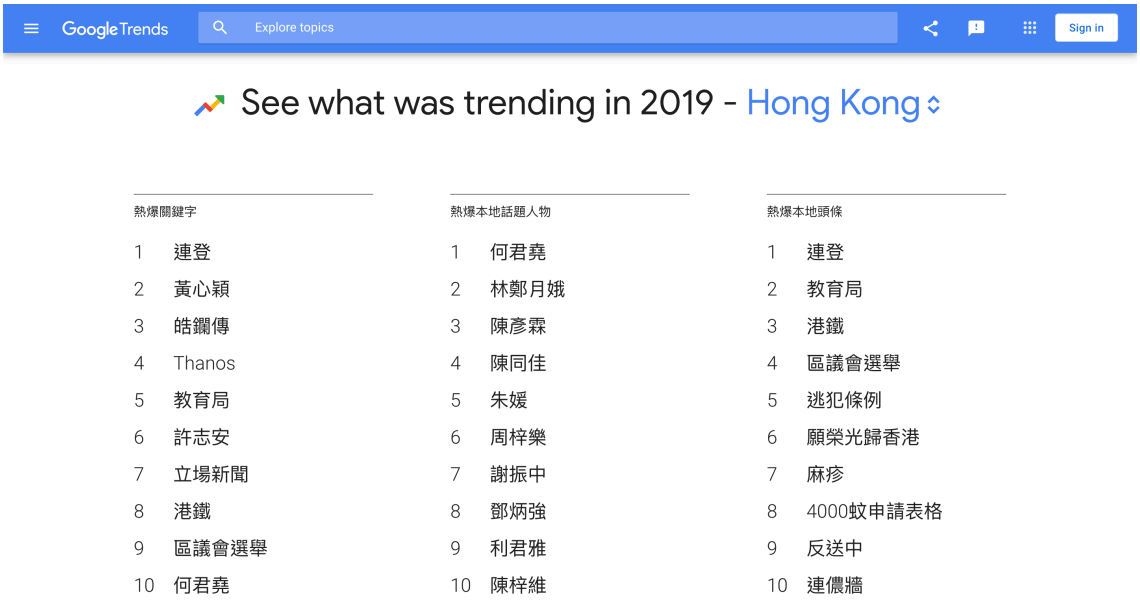


Figure 3.1: Top 10 Google Searches (Keyword, People and News) in Hong Kong

100 threads were crawled in total, 50 from LIHKG (lihkg.com) and 50 from Discuss (discuss.com.hk). Threads with more than 10 pages were selected, and comments from the first 10 pages were retrieved using a web-crawling programme. The collected threads revolved around 5 topics, in which each topic could be divided into 2 subgroups. The topics were selected for being popular searches on Google, as well as its significance in the 2019 political movement in Hong Kong. In addition, threads concerning this year's coronavirus outbreak were included. The threads were created from January 2019 to July 2020, and a total of 16234 comments were obtained. The table below gives an overview of the thread topics under study.

Topic	**Search Keyword(s)**	**Threads Retrieved/Forum**
July 21st Yuen Long Incident	721, 林卓廷, 何君堯, 元朗恐襲, 元朗黑夜	5 threads created in 2019 and 2020 respectively
2019 District Council Election	區議會選舉, 區選	5 threads created before and after the election respectively
Anti Extradition Law Amendment Bill Movement	逃犯條例, 反送中	5 threads per keyword
Coronavirus Pandemic	武漢肺炎, 新冠肺炎	5 threads per keyword
Carrie Lam (current Hong Kong Chief Executive)	林鄭月娥	5 threads created in 2019 and 2020 respectively

Table 3.1.1: Search topics and subgroups during forum thread crawling

Upon retrieval, the comments were cleaned. Following the definition for Cantonese texts by Bauer (2018), complete English sentences were removed, but not English words mixed into comments, nor those used with Cantonese norms. Comments containing only emojis and without any Chinese characters were deleted too. Hyperlinks were a target of cleaning as well. Repetitive content including quotes of other comments or external sources, along with those used to "push" the thread to make it more "popular" thus reachable on the forums were trimmed from the raw data. LIHKG, also known as the platform where protesters plan their actions, was especially noisy with comments that call for actions. For example, links to propaganda materials, online petitions and other posts about action plans were shared to threads to gain more attention by fellow *yellow ribbons*. An average of 100 comments per topic were hence eliminated from the LIHKG data. The following table summarizes the number of comments gathered. A total of 15677 comments were retained after cleaning.

Topic	LIHKG	Discuss
July 21st Yuen Long Incident	1207 (1115)	1439 (1413)
2019 District Council Election	1883 (1766)	1379 (1353)
Anti Extradition Law Amendment Bill Movement	1992 (1899)	1414 (1393)
Coronavirus Pandemic	2149 (2030)	1416 (1393)
Carrie Lam (current Hong Kong Chief Executive)	2001 (1910)	1388 (1405)
Total	**9232 (8720)**	**7036 (6957)**

Table 3.1.2: Numbers of comments collected before (and after) cleaning

3.2 Word Segmentation

As explained in the last session, word segmentation is an essential stage in Chinese automatic text classification. With the idea that Cantonese and standard Chinese texts could be represented in a spectrum –with one end with more Cantonese expressions and the other with standard Chinese expressions –the frequencies of each type of expression could be used to classify a piece of text into either variety. Classification based on n-grams is believed to be the solution towards the problem, taking into consideration of both function and content words. Assuming that people from different forums write differently, not only the degree of Cantonese-ness, but it is also expected that their lexical choices diverse, as a result of opposing political beliefs. To further look into such diversion, word segmentation is required to extract meaningful words.

Word segmentation is no new problem in the Chinese NLP field. Jieba[3] is a well-known word segmentation tool. It was used for word segmentation in recognized works, for instance, Liao et al. (2019), Li et al. (2018) and Chen et al. (2019).

[3]Jieba Chinese Text Segmentation: `https://github.com/fxsjy/jieba`

However, Jieba is supposed to be developed on simplified, standard Chinese. It is possible that using the tool on Cantonese written with traditional Chinese characters gives rise to inaccuracies. To cope with the problem, the word segmentation tool in PyCantonese[4] (Lee 2015) could be utilized. Still, the size of the training data of PyCantonese is rather small with only 150 thousand characters. Cantoseg[5] provides a solution by combining Jieba and PyCantonese.

Testing the three tools on Cantonese data, it was discovered that each of the solutions gave different results. For Jieba, it over-segmented the words into morphemes. For example, 唔好 (do not) was separated into 唔/好 (prefix showing negation and good –the meaning of 好 in isolation does not imply the action of doing). Another example is 你🄵 (plural you), which is separated into 你 (you) and 🄵 (plural suffix for personal pronouns). PyCantonse produced an output which made the most sense, yet the segments could be further grouped into larger meaningful chunks. To be exact, 票 (vote)/債 (debt)/票/償 (compensate) could be grouped into 票債 (debt of vote)/票償 (compensate by voting). Lastly, Cantoseg made a mistake at 票/債票 (financial bond)/償. Interestingly, Jieba seemed to be able to learn from previous outputs. Specifically, the segmentation results from Jieba after running Cantoseg yields 票債/票償 for 票債票償.

On the other hand, when testing on standard Chinese data (written in traditional Chinese characters), Jieba failed to recognize 到處 (everywhere) as a word. For 不單單是 (not only is), it could also be segmented into 不 (not)/單單 (only)/是 (is). Having said that there is no widely accepted, ultimate standard for Chinese word segmentation, all the segmentation results, in fact, made sense in this case. If a best output is to be chosen, it is reckoned that outputs except the one from PyCantonese

[4]PyCantonese: `https://pycantonese.org/data.html`
[5]Cantoseg: `https://pypi.org/project/cantoseg/`

could be picked. Hence, it is decided that segmentation would be performed with running Cantoseg then Jieba for the following experiments.

Segmentation Tool	Output Segments
Input (Cantonese): 泛民唔好咁樂觀，沉默的市民要你哋票債票償！	
Jieba	泛民/唔/好/咁/樂觀/，/沉默/的/市民/要/你/哋/票/債票償/！
PyCantonese	泛民/唔好/咁/樂觀/，/沉默/的/市民/要/你哋/票/債/票/償/！
Cantoseg	泛民/唔好/咁/樂觀/，/沉默/的/市民/要/你哋/票/債票/償/！
Cantoseg + Jieba	泛民/唔好/咁/樂觀/，/沉默/的/市民/要/你哋/票債/票償/！
Input (Standard Chinese): 暴徒到處犯法都沒事，香港不單單是逃犯天堂，簡直是犯罪天堂	
Jieba	暴徒/到/處/犯法/都/沒事/，/香港/不單/單是/逃犯/天堂/，/簡直/是/犯罪/天堂
PyCantonese	暴徒/到處/犯法/都/沒事/，/香港/不單/單/是/逃犯/天堂/，/簡直/是/犯罪/天堂
Cantoseg	暴徒/到處/犯法/都/沒事/，/香港/不單單是/逃犯/天堂/，/簡直/是/犯罪/天堂
Cantoseg + Jieba	暴徒/到處/犯法/都/沒事/，/香港/不單單是/逃犯/天堂/，/簡直/是/犯罪/天堂

Table 3.2: Segmentation results from various segmentation tools

3.3 Punctuation and Stopwords

It is also mentioned in the last chapter that, written Cantonese is a vernacular variation and is used mostly in informal settings. In the context of online discussion forums, it is common for users to skip lines or use spaces instead of using conventional punctuation. While the use of spaces and punctuation is a possible subject of investigation, they could introduce noise when studying other aspects of the writing norms of netizens from different forums. Therefore, it is preferred to have them removed when they are irrelevant, say, when studying lexical discrepancies of the two groups.

Meanwhile, stop words is another area of concern. Stop words are words that are commonly found in a language, too common to the extent that they could blur the distinguishing line between two groups. Function words are typical stop words that do not carry much meaning. They could be useful in showing subtle, grammatical differences, but might not give many clues on the diverse perception on the same topic by individual authors. The fact that there is no standard code for writing Cantonese –in terms of using which characters for which sounds or in which contexts –hinders the identification and elimination of stop words from a piece of Cantonese text. An example is *this*, which could be written as 呢, 哩, 尼, *lei1*, or even *lee*. Contrarily in standard Chinese –這 or 此, the variation for *this* is far narrower, and that the pronunciations of the two words are highly distinctive. Despite the fact that replacement or regular expressions could alleviate inconsistencies in the data, it is impossible to include all the variations.

3.4 Baseline Model: Multinomial Naïve Bayes

Multinomial Naïve Bayes was employed as the baseline algorithm. Bayesian algorithms are widely adopted for text classification tasks. Such algorithms basically calculate the probability of the class of a given document. For Multimnomial Naïve Bayes, instead of taking the probability of a single document, the probabilities of the occurrence of words in the sentences would be calculated. This fits the context for short text classification, in which a sentence would be tokenized and the conditional probabilities of n-grams would be taken into consideration. The algorithm is well-known for its effectiveness in dealing with huge amounts of data.

The Scikit Learn Library was used to build baseline models. First the entire dataset was split into training and testing sets with the ratio of 3:1 with the random state of 2. Features were then obtained using both the Count Vectorizer and the TFIDF Vectorizers. 1-3-gram character features and 1-2-gram segmented word features were calculated. All models using character n-grams gave at least 80% F1 scores. Bigrams gave the best accuracy with both vectorizers. For combined character n-grams, combining unigrams and bigrams gave the best results. These echoed with literature findings. Models using purely trigrams gave the worst results, which was also mentioned in the literature that the longer the n-gram length is, the less accurate the model would be. Better results were yielded using the Count Vectorizer for character n-gram features in general.

For word-level n-grams, a lower accuracy was observed. For both vectorizers, segmented unigrams gave the best, near 80%, accuracy. Removing spaces, stopwords and punctuations gave similar results. However, the accuracy of the models greatly declined using word-level bigrams. Accuracies of around 60% were recorded using bigram and combined unigram and bigram features using either of the vectorizers. Again, features extracted using the Count Vectorizer produced slightly better results compared to the TFIDF Vectorizer.

From the baseline experiments, it was decided that character unigrams and bigrams, along with both cleaned and uncleaned segmented word unigrams would be used for comparison in the following experiments with deep learning models.

Performance of Character N-gram models					
Vectorizer	**Ngram**	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Count	1	0.8031	0.8161	0.8405	0.8281
	2	0.8235	0.8191	0.8821	0.8494
	3	0.7645	0.7601	0.8518	0.8033
	1-2	0.8329	0.8366	0.8748	0.8553
	1-3	0.8327	0.8334	0.8793	0.8558
	2-3	0.8250	0.8250	0.8757	0.8496
TFIDF	1	0.7918	0.7745	0.8906	0.8285
	2	0.8074	0.7726	0.9336	0.8455
	3	0.7416	0.7066	0.9272	0.8020
	1-2	0.8000	0.7614	0.9404	0.8415
	1-3	0.7893	0.7490	0.9426	0.8347
	2-3	0.7934	0.7582	0.9309	0.8357
Performance of Segmented Word N-gram models					
Vectorizer	**Ngram**	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Count	1	0.7832	0.7813	0.8554	0.8167
	1, 2	0.6107	0.5949	0.9733	0.7384
	2	0.5730	0.5694	0.9995	0.7255
	1, cleaned	0.7788	0.7795	0.8482	0.8124
TFIDF	1	0.7755	0.7524	0.8979	0.8187
	1, 2	0.6061	0.5914	0.9783	0.7371
	2	0.5727	0.5692	0.9995	0.7254
	1, cleaned	0.7735	0.7533	0.8902	0.8161

Table 3.4: Summary of the performance of the baseline models

3.5 Deep Learning Models

3.5.1 Overview: Conventional Neural Networks

As reviewed from various literature, deep learning approaches have been rapidly refreshing state of the art standards for different NLP tasks. This is true in the domain of text classification too. In this section, the principles of deep learning, as well as how it could be applied in classifying Cantonese political short texts would be demonstrated.

First and foremost, to perform deep learning, neural networks are required. The concept of neural networks is an analogy of the human brain, where nerve cells called neurons are connected to one another and relay neurotransmitters through synapses. A neural network takes in text data, converted into numbers, as inputs, and relays the data in its hidden layers. Eventually, the network outputs the class of the text in the case of text classification. In the following, Keras with TensorFlow would be employed to implement deep learning experiments.

Input Preprocessing

Having said that text must be transformed into numbers before being fed into a neural network, Keras offered a few tools for users to preprocess their data. To be exact, the Keras Tokenizer and one-hot encoder would be used to convert characters and words into numbers in this dissertation. In programming terms, such numbers form a dense vector representation of the characters or words, which is known as embedding. Apart from creating task-specific embeddings, publicly available, pre-trained word vectors would also be exploited in latter experiments.

When experimenting with character-based classification, the Keras Tokenizer was adopted. To begin, the crawled comments were segmented into unigrams and bigrams respectively. By applying the Keras Tokenizer to the character-segmented comments, an index was assigned to every character unigram (or bigram). Upon a brief survey on the comments, it was found that the entire dataset consisted of 4046 unique unigrams and 118500 unique bigrams. Knowing that instances of low frequencies are often negligible, only the 2000 most frequent unigrams and 11850 bigrams were retained for deep learning.

Next, prior to feeding arrays of numbers into the neural network, the arrays have to be in a fixed length. Comments that are longer than a certain length would be trimmed, and those that are shorter would be padded with zeros. It was discovered that the average length of a forum comment consisted of 28 characters. Thus, the maximum length was set to 28 for unigram-based classification, and 20 for bigram- (and later, word-) based classification.

For word-based classification, embeddings were obtained by two methods, including one-hot encoding and applying pre-trained vectors. The former takes into the account of the entire vocabulary, treating each instance as a separate category. For instance, the vector for a word would be an array with the length of the vocabulary size, in which only the value for the 'category' of the word is 1 and the rest would be 0. Again, the number of unique, segmented words was observed to be 30034. Therefore, it was determined that a vocabulary of 3000 words would be investigated. Similarly, the one-hot encoded strings of segmented comments would be padded into the same length (which was 20, as aforementioned).

On the other hand, fastText word vectors were utilized to produce embeddings in the latter way. Assuming that the fastText vectors provided by Facebook (Grave et al.

2018)[6] are trained on standard Chinese, another set of pretrained fastText word vectors offered by ToastyNews [7] would also be taken into consideration. ToastyNews' fastText data is trained on Traditional Chinese Hong Kong data. Sources include Cantonese Wikipedia, forums and news websites etc. Although the vocabulary size and number of tokens in ToastyNews' training data is far less than Facebook' s, it is worth-attempting so see if variety-specific word vectors could help improve training results.

Model Architecture

All models in the coming experiments would be composed based on a sequential architecture. Such architecture is known for handling data with linear stacks of neural network layers, involving a single input and a single output tensor. For each model, there is first an embedding layer, which converts the text input into embeddings or takes in generated embeddings and relay them onto the next dense layer. The input dimensions and length of the layer was set altered depending on whether unigrams, bigrams, segmented words or cleaned segmented words as explained in the previous part. The output dimension was 64. The succeeding Dense layer is of 32 units, with ReLU as the activation function and L2 kernel regularizer of 0.1 applied. Following the Dense layer would be the LSTM or CNN layers, which would be further discussed in the further sections. Finally, the output Dense layer is of a single unit with a sigmoid activation function for binary classification.

When compiling the models, the binary cross entropy loss function would be used to measure errors made during training. The optimizer selected for the models was

[6]FastText Word vectors for 157 languages: `https://fasttext.cc/docs/en/crawl-vectors.html`
[7]Hong Kong fastText by ToastyNews: `https://github.com/toastynews/hong-kong-fastText`

Adam and the performance of the models would be calculated in terms of accuracy.

Validation and Evaluation

In the baseline models, the entire dataset was split into training and testing data for evaluation. For the deep learning models, in order to further prove that the models are effective in classifying forum comments, the dataset was further divided into training, validation and testing sets. To obtain a representative test set, all the data was once again divided by forum. Comments from separate forums were then shuffled, and 10% of the comments were selected from each group to form the test set. During model training using the remaining 90% of the data, 10-fold cross validation was implemented to divide the data again into training and validation sets.

3.5.2 LSTM with One-Hot Encoding

LSTM is a kind of recurrent neural network (RNN) architecture. In an RNN, "gates" are present to regulate the data when it is being relayed in the neural network (Li et al. 2018). In the case of the LSTM structure, a unit consists of a memory cell and three gates –one for the input, another for the output, and last but not least, a forget gate. As a result, useful information would be "remembered" and relayed to the next step, enhancing the efficiency of the computation. This is especially helpful when dealing with long textual information.

As introduced in the previous section, a LSTM layer of 16 units was inserted before the output Dense layer. Since LSTM belongs to the RNN family, result computation is performed in a repeated manner. This makes the model prone to overfitting and hence the dropout parameter for the LSTM layer was set to 0.5. The LSTM models for character-based classification would take tokenized unigrams and bigrams, and

the word-based models would make use of one-hot encoded embeddings. A total of 4 LSTM models were trained for 50 epochs with the batch size of 10.

3.5.3 CNN with FastText Embeddings

CNNs are usually used for images. When given an image, the CNN applies a filter to capture fixed-sized-bits of the image from different positions. This process is known as convolution. During convolution, redundant information is often yielded because the same area could be captured multiple times with the filter. To tackle with problem, pooling is carried out to reduce the amount of image bits, which would be relayed to subsequent layers in the neural network. In text processing, convolution occurs in a similar sliding window manner, except that the filter is in one-dimension instead of two. While convolution helps to detect unique characteristics of images, it facilitates feature extraction from texts as well. This fits perfectly with the current aim in identifying differences in word use of two groups of netizens.

6 CNN models were built and trained in total –2 using character embeddings (unigram and bigram), 2 using standard Chinese fastText and 2 using ToastyNews Hong Kong fastText (raw and cleaned segmented words). The model architecture started with an embedding layer and a Dense layer, with the specification suggested in the Overview section. Attached to the Dense layer, there was the Conv1D layer with 64 filters and the kernel size of 10. After convolution, a MaxPool1D layer was chosen for pooling. A dropout layer of 0.5 was then applied and the data was passed through a Flatten layer, so that the data' s shape could be adjusted for the output Dense layer. Each of the models were trained with a batch size of 10 for 20 epochs respectively.

3.5.4 Transfer Learning

Apart from traditional neural networks, 2 more models were trained with the Transformer architecture. In contrary to supervised learning in the LSTM and CNN models, the Transformer architecture is known for transfer learning (Malte & Ratadiya 2019). Conventional supervised learning involves the use of task-specific training data. However, with transfer learning, models are not required to train on specific data (Pan & Yang 2010). For a hypothetical task of classifying two groups of text, traditional learning would require two models for the two groups respectively. Each model would be trained and tested on texts of the same group. Meanwhile, a transfer learning model would just need to be trained on one of the groups (the source domain). After being tested on the same group, the model could reuse the knowledge from classifying the texts from one group on another group (the target domain). This type of learning is said to be able to further reduce the impact of overfitting (Mou et al. 2016) and the reliance of training data from the target domain.

The Hugging Face Transformers were utilized to create transfer learning models. The data was split into training and testing sets with the ratio of 3:1. To begin, a pre-trained tokenizer was employed to tokenize strings of comments. The tokenized comments were then encoded into tensors, padded to the maximum length of 20. They were then passed into the pre-trained models, optimized using Adam, with a learning rate of 2e-5 and epsilon of 1e-8. Last but not least, the models were compiled in the same manner as the other deep learning models –adopting the binary cross entropy loss function and evaluated by accuracy. Each model was trained for 10 epochs. Further details of the models would be introduced in the following subsections.

XLMRoBERTa

As mentioned, using a Transformer model requires training on a source task, then fine-tuning the model to work on a target task using target task training data. During pre-training with the source task, language models are built using a huge volume of unlabeled text data and tested with typical NLP tasks, for example classification, inferencing and question answering (Radford et al. 2018). Next word prediction or masked word prediction are typical ways to train a language model. The former takes into consideration of previous words and the latter takes both preceding and succeeding word contexts. In particular, Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al. 2018) is a popular solution for masked language modeling.

Cross-lingual Language Model Pretraining (XLM) (Lample & Conneau 2019) was trained to perform masked language modelling using multilingual data. When calculating the predicted word, data in a different language is exploited to project the result in the language of the masked sentence. This could greatly facilitate NLP in low-resource languages. It was then applied with the Robustly Optimized BERT approach (RoBERTa) proposed by Liu et al. (2019), which was proved to be an improved version of BERT with adjustments such as lengthen training time and expanded training data.

XLMRoBERTa is created under the marriage of the XLM and RoBERTa by Facebook Ai. It was trained using 2.5TB of CommonCrawl data, which includes 100 languages (Conneau et al. 2019). To be exact, the `tf-xlm-roberta-base` [8] pretrained model was put to use along with the XLMRobertaTokenizer and the TFXLMRoBERTAForSequenceClassification by Hugging Face in this dissertation. According to Hugging Face, the model is able to automatically determine the language of the

[8] `https://huggingface.co/jplu/tf-xlm-roberta-base`

input data. It is likely that standard Chinese data was being trained on in this model.

XLN

XLNet (XLN) is put forward by Yang et al. (2019). In contrary to BERT, XLN models language in an autoregressive manner. It was pointed out that BERT' s autoencoding language modelling along with its masking mechanism leads to distortion of training data. This gives rise to discrepancies between pre-training with the source task and fine-tuning with the target task. However, in the case of XLN, bidirectional contexts are captured by acknowledging all possible permutations of the tokens in a sentence as it is. Permutation, at the same time, processes textual data randomly rather than sequentially. As a result, XLN is capable of making maximum use of the general data and overcome the pre-training and fine-tuning gap in BERT.

While pre-trained XLNet models are available on Hugging Face, these models are trained on English texts. By happy chance, it was discovered that ToastyNews offered a pre-trained model which was developed with Hong Kong data. Similar to the word vectors used in the CNN models, the `xlnet-hongkongese-base` [9] model was trained with mostly news and blog articles, as well as the Yue (Cantonese) Wikipedia, restaurant reviews, forums threads and online fiction. It should also be noted that a ratio of 6:3:1 standard Chinese, Cantonese and English data was being trained on. This could possibly tackle with the code-mixed materials crawled for this dissertation. The last thing to note for this pretrained model is that, one of the evaluation tasks was to categorize LIHKG threads based on the first post in the thread. Despite the fact that the model was not of developed in a scholarly manner, its affinity towards the data investigated in this dissertation provides good grounds to exploit it.

[9] `https://huggingface.co/toastynews/xlnet-hongkongese-base`

Finally, upon identifying the optimal pre-trained model, the AutoTokenizer and TFXLNetForSequenceClassification were used to process crawled forum comments string by string. The following subsection summarizes the testing and validation results of all the deep learning models from all the previous sections.

3.5.5 Summary

Architecture	Embedding/Pretrain data	Accuracy	Precision	Recall	F1 Score
LSTM	**unigram + tokenizer**	**0.764**	**0.76**	**0.76**	**0.76**
	bigram + tokenizer	**0.742**	**0.75**	**0.74**	**0.73**
	segments + one-hot encoding	0.707	0.71	0.71	0.71
	cleaned segments + one-hot encoding	0.678	0.68	0.68	0.68
CNN	unigram + tokenizer	0.721	0.72	0.72	0.72
	bigram + tokenizer	0.714	0.71	0.71	0.71
	segments + fastText standard Chinese	0.714	0.71	0.71	0.71
	cleaned segments + fastText standard Chinese	0.667	0.68	0.67	0.65
	segments + Hong Kong fastText	**0.739**	**0.75**	**0.74**	**0.73**
	cleaned segments + Hong Kong fastText	0.692	0.69	0.69	0.68
XLMRoBERTa	jplu/tf-xlm-roberta-base	0.530	(unavailable due to computational constraint)		
XLNet	**toastynews/xlnet-hongkongese-base**	**0.559**			

Table 3.5: Summary of the performance of the deep learning models

Generally speaking, character-based classification models performed better than word-based classification models. Unigrams provided the maximum clues towards telling which forum a comment belongs to. The LSTM models did a better job at character-level classification.

For word-based classification, stopword removal did not seem to improve classification accuracy. Instead, it appeared to have done the opposite, bringing model accuracies below 70%. Looking into the confusion matrices of the models, it could be seen that the most falsely labelled comments were from models that took in cleaned word segments. On the other hand, the incorporation of pre-trained word embeddings demonstrated improvement on the classification accuracy. It was proven

that having the pinpoint embeddings is crucial to accurate classification as the Hong Kong fastText model achieved the best results on a new set of data.

Again, the importance of matching pre-training data was justified in the transfer learning models as the XLNet model outperformed the XLMRoBERTa model. Although the models failed to complete training on all 15 epochs due to hardware constraints, preliminary statistics were successfully captured from model validation during the first 10 epochs.

The evaluation of the deep learning models would be continued in the next chapter, with an in-depth exploration on the insights gained. In addition, analysis on the errors and possible solutions towards limitations of the experiments would be suggested as well.

4 Results and Discussion

4.1 Results

Model	Accuracy	Precision	Recall	F1 Score
MNB Count Vectorizer Unigram	0.80	0.82	0.84	0.83
MNB Count Vectorizer Bigram	0.82	0.82	0.88	0.85
MNB Count Vectorizer Segments	0.78	0.78	0.86	0.82
MNB Count Vectorizer Clean Segments	0.78	0.78	0.85	0.81
LSTM Unigram + Keras Tokenizer	0.76	0.76	0.76	0.76
LSTM Bigram + Keras Tokenizer	0.74	0.75	0.74	0.73
CNN Segments + HK fastText	0.74	0.75	0.74	0.73
CNN Clean Segments + HK fastText	0.69	0.69	0.69	0.68
XLNet HongKongese Base	0.56	n/a		

Table 4.1: Performance of the most accurate baselines and deep learning models

Comparing the accuracy matrices of the best deep learning models with the best baseline models, it seems that the deep learning models were unsuccessful in overcoming the baselines. Up to this point, the Naïve Bayes bigram model still tops the chart with the highest accuracy, surpassing 80%, as well as the greatest f1 score of 85. This agrees with the findings of Zhang et al. (2011) that Naïve Bayes classifier using bigram features are the best in labelling Cantonese data.

Character-level n-gram classification was generally better than word-level classification for both the baseline models and deep learning models. For word-level classification, while stopword cleaning did not bring much impact on the baseline models,

prediction quality was drastically deteriorated in the deep learning models. All in all, it could be concluded that multinomial Naïve Bayes is useful for classification with both character and word-based features, with a slight preference for the former. Reviewing all the baseline models, it was noticed that models using the Count Vectorizers produced generally better results with the features selected for deep learning. Therefore, the next section looks deeper into such models for any insights that helps to highlight expressions that are preferred for each of the forums.

4.2 Findings

Looking into the classifier coefficient of the features, brief characteristics of how people from Discuss and LIHKG write differently.

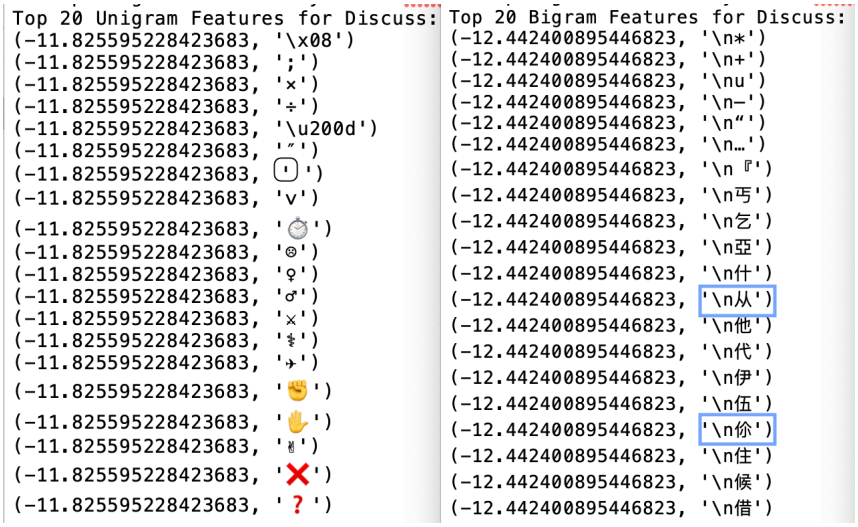


Figure 4.1: The most important unigram and bigram features from Discuss

First of all, it was within expectation to see a large number of emoji unigrams in Discuss comments, based on the observation that local elders use them frequently

and Discuss is associated with blue ribbons. For bigrams, although the top features were highly polluted with the newline character, it could still be noticed that Discuss comments were rich in punctuations. Simplified and non-standard Chinese characters (those highlighted in blue), which were likely to be produced from handwriting input, were spotted as well.

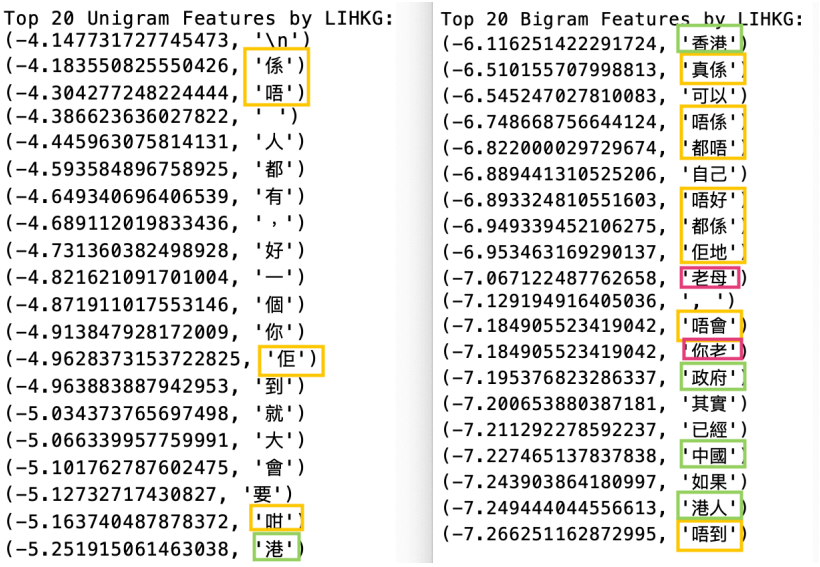
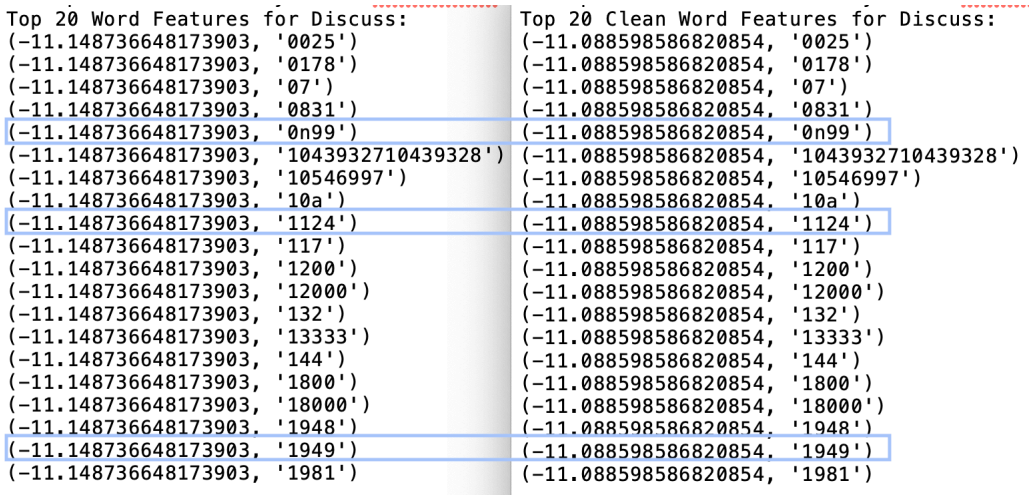


Figure 4.2: The most important unigram and bigram features from LIHKG

Moving onto the top character features from LIHKG, a clear rank difference was recorded. Instead of punctuation marks, new lines and spaces were preferred by LIHKG users. The only punctuation mark seen was comma. Several Cantonese particles and Cantonese grammatical expressions (highlighted in yellow) were on the lists, hinting that writings on LIHKG are heavily Cantonese-inclined. Other notable traits include the absence of simplified Chinese characters and vulgar expressions (highlighted in red). It was also obeserved that the focus of discussion in LIHKG often highlights China, Hong Kong and the government (highlighted in green).



Top 20 Word Features for Discuss:	Top 20 Clean Word Features for Discuss:
(-11.148736648173903, '0025')	(-11.088598586820854, '0025')
(-11.148736648173903, '0178')	(-11.088598586820854, '0178')
(-11.148736648173903, '07')	(-11.088598586820854, '07')
(-11.148736648173903, '0831')	(-11.088598586820854, '0831')
(-11.148736648173903, '0n99')	(-11.088598586820854, '0n99')
(-11.148736648173903, '1043932710439328')	(-11.088598586820854, '1043932710439328')
(-11.148736648173903, '10546997')	(-11.088598586820854, '10546997')
(-11.148736648173903, '10a')	(-11.088598586820854, '10a')
(-11.148736648173903, '1124')	(-11.088598586820854, '1124')
(-11.148736648173903, '117')	(-11.088598586820854, '117')
(-11.148736648173903, '1200')	(-11.088598586820854, '1200')
(-11.148736648173903, '12000')	(-11.088598586820854, '12000')
(-11.148736648173903, '132')	(-11.088598586820854, '132')
(-11.148736648173903, '13333')	(-11.088598586820854, '13333')
(-11.148736648173903, '144')	(-11.088598586820854, '144')
(-11.148736648173903, '1800')	(-11.088598586820854, '1800')
(-11.148736648173903, '18000')	(-11.088598586820854, '18000')
(-11.148736648173903, '1948')	(-11.088598586820854, '1948')
(-11.148736648173903, '1949')	(-11.088598586820854, '1949')
(-11.148736648173903, '1981')	(-11.088598586820854, '1981')

Figure 4.3: The most important uncleaned and cleaned word features from Discuss

For top word features in Discuss, there were no different with the lists of the most important words before and after cleaning. It was unexpected that all of the features were strings of numbers. Still, the numbers were not completely meaningless at all. For instance, 1124 was the date of the 2019 District Council election, and 1949 was the year when People' s Republic of China was established. An interesting observation was that "0n99" , which was supposedly written as "on99" , that has the same pronunciation of a vulgar expression for "someone weird"was on the list.

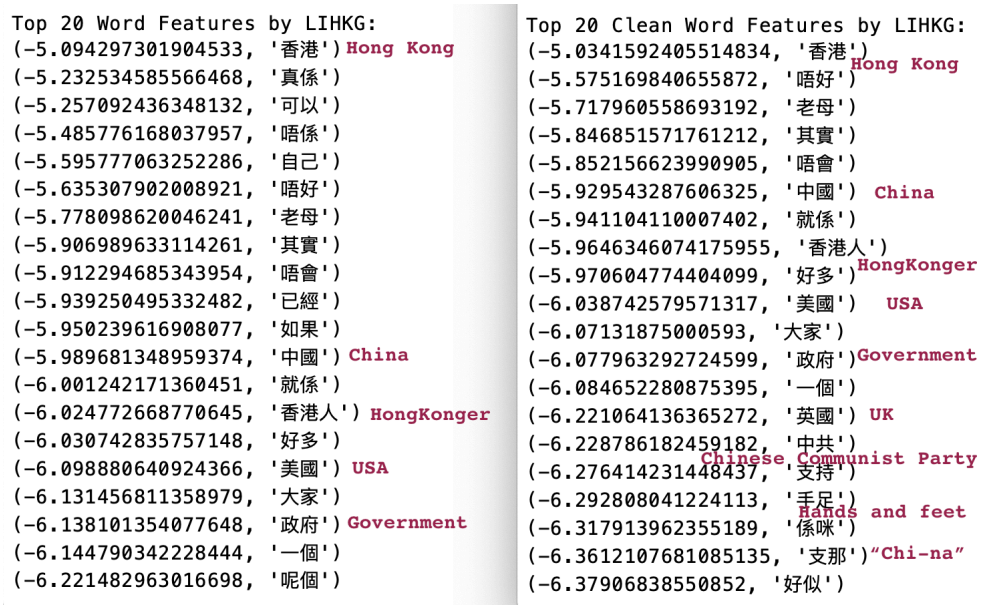


Figure 4.4: The most important uncleaned and cleaned word features from LIHKG

Last but not least, the most important cleaned word lists for LIHKG revealed more hints on the topics of discussion. With the provided translations, an anti-China sentiment was evident. Knowing that one of the tactics used by yellow-ribbon-protesters was to involve the international community, country names of the USA and the UK appeared on the lists. Other terms such as HongKonger and "hands and feet" (used to describe fellow protesters) were rather prominent in LIHKG comments too. Swear words and Cantonese expressions highlighted in the previous lists were still observable from the segmented word lists.

To sum up, despite of the fact that the Discuss list could not provide many insights on the distinctive characters or words being used by the forum, some valuable information on LIHKG writing style could be concluded from the above. To be exact, LIHKG writing is nearer to the Cantonese side on the Cantonese-standard Chinese

spectrum, with the frequent use of vulgar terms. Topics mainly revolve around Hong Kong and Chinese governments, with a negative sentiment especially when it comes to the latter. Still, the insights provided by the important features list were rather general. Hence, a deeper dive into the wrongly classified sentences would be required. The following section provides a more detailed analysis on possible critical differences on the writing style of the two forums.

4.3 Error Analysis

The figures on the next page show the confusion matrices for the baselines and the best-performing deep learning models. As a reminder, the confusion matrix for the baseline models is the average number of the results yielded from the Count Vectorizer models using unigrams, bigrams, uncleaned and cleaned word unigrams. The confusion matrix for the LSTM models displayed results from the unigram and bigram models (with the average numbers in brackets). Finally, the results from the CNN model using uncleaned word segments and ToastyNews vectors was presented. It should also be noted that the baseline models were trained with a train-test split while the deep learning models made use of a train-validation-test split. This results in the smaller numbers in the confusion matrices of the deep learning models.

Overall speaking, as the number of LIHKG comments collected was greater than that of Discuss comments, there was a bias of tending to classify an unseen comment as coming from LIHKG. In all the models, instances that were wrong labelled as from LIHKG were more than that of those of Discuss.

MNB (Count)	Predicted Discuss	Predicted LIHKG
Actual Discuss	1229.25	477.75
Actual LIHKG	319.75	1895.5

Figure 4.5: Averaged confusion matrix for the baseline Count Vectorizer models

LSTM (1+2-grams)	Predicted Discuss	Predicted LIHKG
Actual Discuss	500+388 (444)	197+309 (253)
Actual LIHKG	173+96 (134.5)	699+776 (737.5)

Figure 4.6: Combined confusion matrix for the LSTM unigram and bigram models (average number in brackets)

CNN segmented (HKfastText)	Predicted Discuss	Predicted LIHKG
Actual Discuss	400	297
Actual LIHKG	113	759

Figure 4.7: Confusion matrix for the CNN model using uncleaned word segments and ToastyNews vectors

4.3.1 Baseline Models

First and foremost, from the character-based classifiers, it was found that a lot of the comments which were wrongly identified as from Discuss were those written with standard Chinese. Those with normal or repeated punctuations, such as "???" or "。。。" were frequent in such comments. Simplified Chinese characters and Mandarin expressions used in an ironic manner results in classification errors by the model as well. Some of the examples, in spite of including class-important terms, were surprisingly mistakenly classified. This could entail the use of punctuation might override the use of class-distinctive terms when the model judged if the comment is from Discuss or LIHKG.

On the other hand, in the comments that were labelled as from LIHKG by error, the same ironic use of class-important items was noticed, despite repeated emojis were found in the same comment. Vulgar expressions, which were believed to be prominent in LIHKG comments were present in Discuss comments too. In the final example, even if simplified and standard Chinese characters were present, the model failed to label the comment correctly. From these observations, it could be learnt that general domain expressions, for example swear words and standard Chinese particles were yet to be accurate detectors of forum-specific language. What makes the situation worse is that sarcastic and intentional use of the opposite side' s language introduced serious distractions to the model.

Comment	Predicted	Actual
支那除咗騙子，乜都係假既，咁即係。。。。。	Discuss	LIHKG
武漢肺炎！！！！	Discuss	LIHKG
陳雲：香港阿叻，玩死美國特朗普。香港政客及香港人的虛偽與惡毒，連特朗普都嚇怕。是否簽署《香港人權法案》，特朗普不放在心上。	Discuss	LIHKG
沒選票，沒土地，沒政治權力的一群人，聚在一起高談民主的壞處。 我彷彿看到，一群太監在說性生活多傷身體，幸虧咱們閹了； 或者是一群乞丐在說錢是多麼骯臟的東西，還是咱要飯乾淨。	Discuss	LIHKG
五大訴求未爭取到，香港冇民主，去支共做咩呀	LIHKG	Discuss
講錯 手足先岩🤫🤫	LIHKG	Discuss
可能？我可能係你老豆，岩岩內射你老母	LIHKG	Discuss
创作力量和幻想	LIHKG	Discuss

Table 4.3.1.1: Comments wrongly labelled by the baseline character-based classifier

For the segmented comments, the same type of errors could be observed, especially those related to sarcastic language use. After segmentation and cleaning, only neutral everyday keywords were remained. Without any strong clues to rely on, the model, with training data biased towards LIHKG, could only make a wild guess and label the data as from the majority class. The third and forth examples in the following table demonstrates such kind of error. Code mixing is perceived to be more prominent among the educated, which a lot of them happen to be yellow ribbons. Yet, this was proved to be merely an assumption because blue ribbons in fact, code mix heavily as well, but in the quoted example, it just happens that the author was using a legitimate English word to transcribe a Cantonese syllable.

Comment	Predicted	Actual
對 啊 ， 中國 武漢 肺炎	Discuss	LIHKG
這裡 係 連登 你 懂 的	Discuss	LIHKG
須要 幫 開始 被 檢控 清算 受害者 包括 老師	Discuss	LIHKG
自求多福	Discuss	LIHKG
美國 鍾意 送 晒 比 地 law	LIHKG	Discuss
根本 班黃絲體 就係 癲狗 😏	LIHKG	Discuss
這是 我 唯一 覺得 有義意 的 投票	LIHKG	Discuss
全部 係 假 的 我 唔會 信 ！ 拍 唔 到 ， 我 唔 信 拍到 ， 我 唔 信 ！ 藍絲 如何 鍊 成 ！	LIHKG	Discuss

Table 4.3.1.2: Comments wrongly labelled by the baseline word-based classifier

4.3.2 Deep Learning Models

Proceeding to the deep learning models, again, stylistic features were proved to be useless in classification. Using the LSTM unigram model as an example, despite being split into separate lines, the first example was wrongly classified, possibly due to the presence of correct punctuation. Neutral terms forced the model to label the comment as from LIHKG yet another time. For Discuss comments that were labelled as from LIHKG, sarcasm was to be the blame once more. It was also demonstrated just by using the presence of Cantonese items to draw a line between the two groups of comments is just not feasible.

Comment	Predicted	Actual
無國界醫生上去救援未？		
香港警暴又無反應		
咁武漢滿足到佢啦呱？	Discuss	LIHKG
籃屍越禁越走得快過黃絲	Discuss	LIHKG
點只白票？直頭叫你投建制。	Discuss	LIHKG
又贏	Discuss	LIHKG
On99	LIHKG	Discuss
拖得就拖， 希望修例快啲完成。	LIHKG	Discuss
還有遣責再遣責	LIHKG	Discuss
兄弟如手足，有隆大家篤	LIHKG	Discuss

Table 4.3.2.1: Comments wrongly labelled by the LSTM unigram-based classifier

For the worst-performing deep learning models, the LSTM model using cleaned segments was the model that falsely labelled the most LIHKG comments as from Discuss, and the CNN model using cleaned segments with standard Chinese fastText embeddings mistook the most Discuss comments as from LIHKG.

The core reason behind such errors is believed to be the removal of stopwords. Without stopwords, the context of the sentence could go missing. In highly sarcastic political opinions, this could easily result in problematic classification. These errors range from trusting important keywords used for opposite meanings, to blindly labelling comments with the class with more samples due to the bias in training data.

Comment	Predicted	Actual
九龍 議會 聯合 港島 議會 聯合 新界 西 議會 聯合 新界 東 議會 聯合	Discuss	LIHKG
集中營 嚟 唔好 話藥 口罩 夠 叫 先進 國家 仲話 厲害 了 国 笑 撚 死人 一個 瘟疫 得出 支那 仲 存活 七十年代 衛生 意識	Discuss	LIHKG
投 建制	Discuss	LIHKG
中國 武漢 肺炎 五毛 拉	Discuss	LIHKG
若 義士 😷 睡醒 😴 😉 被 泛民 用 完 即棄 就算 錯過 參選 報名 時間 分分鐘 日日 做 家訪 😡 😡 😡 連儂 💩 💩 💩 泛民参 選人 攪炒 🤭 🤭 🤭	LIHKG	Discuss
無法 叫 共匪 退返 錢 無法 煞停	LIHKG	Discuss
天真	LIHKG	Discuss
叫班 紅衛兵 金鐘 行 兩行 就算 陳 同佳 做 特首 什么	LIHKG	Discuss

Table 4.3.2.2: Examples of wrongly labelled sentences from the models producing the most false negatives (labelled as Discuss) and false positives (labelled as LIHKG)

At last, despite for being the best CNN model, the model trained with uncleaned segmented data and ToastyNews Hong Kong Chinese vectors was unsuccessful in classifying 30% of the test set data. It seems that even without stopword removal, the model fails to tell if someone was being sarcastic. Especially in the comments that were miscalculated as from LIHKG, even if the presence of typical expressions that blue ribbons use (as in the second and third examples), the model did not recognise that they were from Discuss. On the other hand, from the interaction from the first and the second comment, it could be seen that not all users of the same forum hold the same political view. This is in fact, the most vital limitation of this dissertation.

Comment	Predicted	Actual
你 冇 睇 歷 史 唔好 亂講 咁 輕鬆 將呢 兩個 字 講出 你 都 好 冇 人性	Discuss	LIHKG
美國 行得 第一步 ， 其他 國家 就 跟住 做	Discuss	LIHKG
太子 站 冇 死人 油麻 地 唔見 月餅 要 ck cctv	Discuss	LIHKG
一屋 三姓 八 選民 ， 大陸 既 美孚 村 都 可以 投票 ， 點 夠 佢 地 玩 ？ ？ ？	Discuss	LIHKG
721 警黑 勾結 何君堯 操控 鄉 黑 無差別 打 市民 黑警 縱容	LIHKG	Discuss
樓主 係 知名 極黃 甲由	LIHKG	Discuss
連登 弱智 仔 真係 以為 共產黨 怕 個 法案 ， 都 列 左個 癡 線 仔 做 四人幫 又 點會 比 你 入閘 ？	LIHKG	Discuss
agree	LIHKG	Discuss

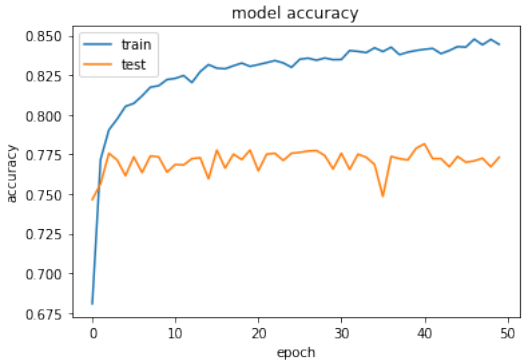
Table 4.3.2.3: Comments wrongly labelled by the CNN classifier using uncleaned word segments and ToastyNews vectors

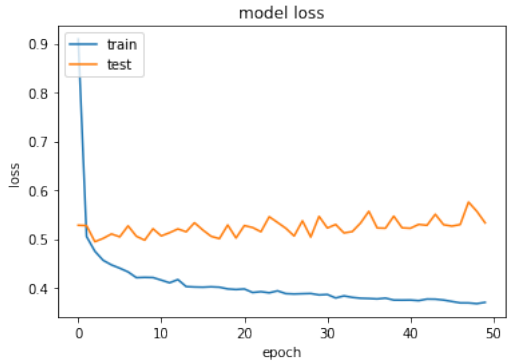
Wrapping up this section, it could be learnt that multiple major causes gave rise to inaccurate classification. First is presence of sarcastic use of words in online political discussion. Second is the faulty assumption that the degree of Cantonese/standard-Chinese and the use of certain writing styles are strongly correlated to political orientation or forum an individual uses.

4.4 Limitations and Suggestions

In fact, apart from the fundamental assumptions of the topic gave rise to the errors found in this dissertation, a number of improvements could be made to possibly enhance experimental performance.

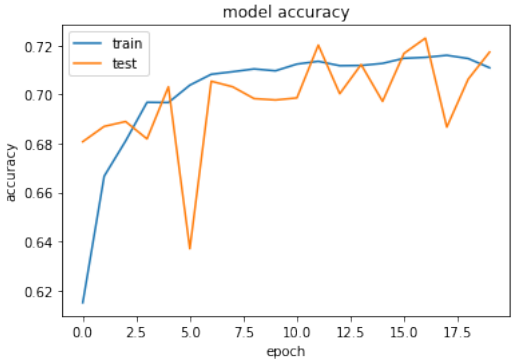
As repeatedly emphasized, Cantonese data was messy, which is also the reason why it is scarce. Even though data could be collected automatically, heavy manual effort is required to clean and normalize the data. In the experiments carried out, it was evident that the data prepared for this dissertation was insufficient. Below are the learning curves recording the learning accuracy and loss of the LSTM unigram model. With the validation data, the model was unable to improve its accuracy after the few initial epochs. The gradual increase in loss is yet another proof of overfitting, which could be resulted from not having abundant data. The possibility of future research would be to give up classifying online political discourse, but to focus on discourse in general on different forums, because this topic constraint minimized the data that could be used. By including a wider range of topics, more data could be exploited for machine training.

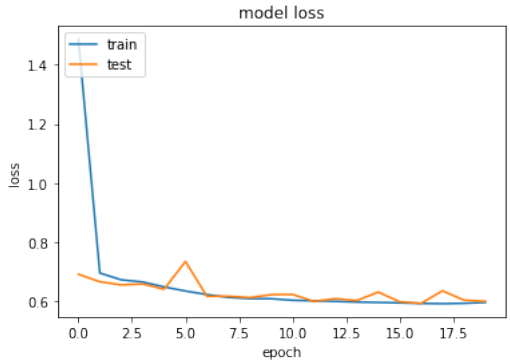




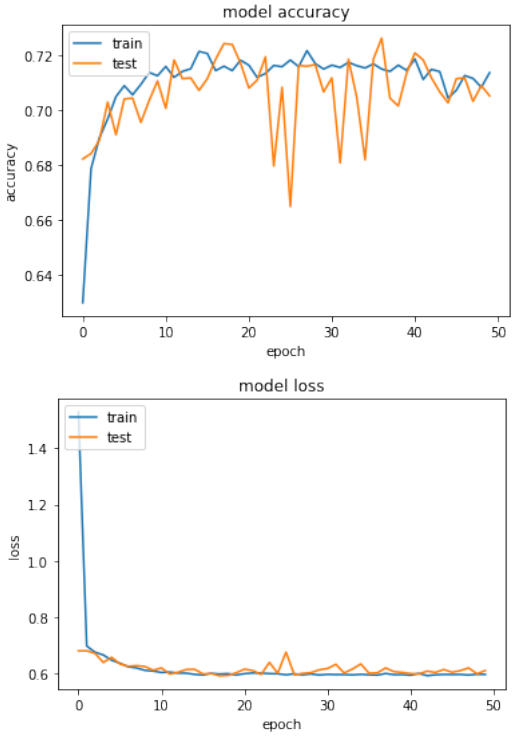
Figures 4.4.1 and 4.4.2: Learning curves of the LSTM unigram classifier

On the other hand, learning curves from the best CNN model and the Transformer models called for further training with more epochs. In the case of the CNN model using uncleaned segments and ToastyNews vectors, by increasing training epochs from 20 to 50, better-looking learning curves could be obtained. For the RoBERTa model, it is likely that 10 epochs were too few for any improvement. The XLNet model simply demonstrated model underfitting. However, due to the user quota of Google Colabotory, the model was not able to carry on training after eleven or twelve epochs.

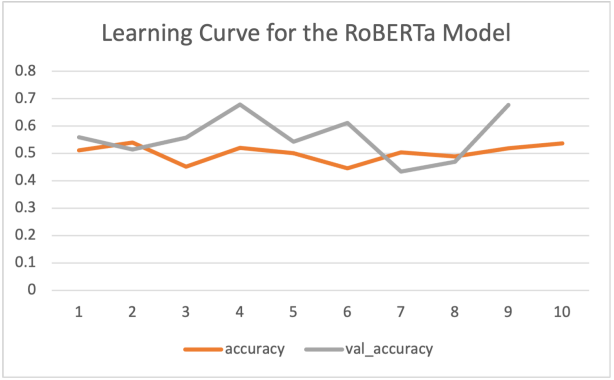




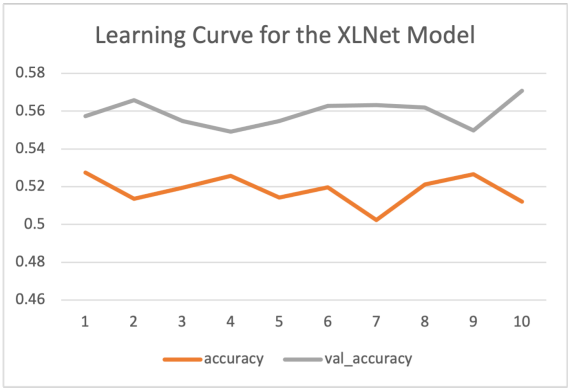
Figures 4.4.3 and 4.4.4: Learning curves of the best CNN classifier on 20 epochs



Figures 4.4.5 and 4.4.6: Learning curves of the best CNN classifier on 50 epochs



Figures 4.4.7: Learning curve of the RoBERTa model



Figures 4.4.8: Learning curve of the XLNet model

5 Conclusion

Concluding this dissertation, despite of the rather ambitious aim to distinguish people of different political orientation through classifying texts from different forums, valuable attempts were made to construct a text classifier for Hong Kong Cantonese. Knowing that Cantonese is a poorly standardized language, 12 models, combining traditional text classification techniques and innovative standard Chinese and Cantonese NLP materials were built. Although the baseline model was not challenged, most of the deep learning models, with roughly 70% accuracy, were actually not far behind the baseline. With more training data and time, better results could be attained.

Meanwhile, current technology allows the exchange of efforts between researchers. There are always more advanced resources for Cantonese NLP. For instance, ToastyNews actually offered ELECTRA models for Transformer model building as well [10]. It is said that ELECTRA is efficient by training all tokens simultaneously. Further attempts using other models could be valuable towards the path of better Cantonese NLP. It is also believe that with advanced Cantonese NLP technology, speakers of the variety could have a better understanding of their own language, as well as enhancing preservation of the language since it has been a concern that the variety is dying.

Last but not least, facing the challenge brought by echo chambers and stratification of the society, the hope for putting text classification in regulating content consumption is yet to be extinguished.

[10] Hong Kong Transformer Models And Other NLP Resources:https://medium.com/@kyubi$_fox/hong-kong-transformer-models-and-other-nlp-resources-42821ad5a700$

Bibliography

Bäck, E. A., Bäck, H., Gustafsson Sendén, M. & Sikström, S. (2018), 'From i to we: Group formation and linguistic adaption in an online xenophobic forum'.

Banisch, S. & Olbrich, E. (2019), 'Opinion polarization by learning from social feedback', *The Journal of Mathematical Sociology* **43**(2), 76–103.

Bauer, R. S. (1988), 'Written Cantonese of Hong Kong'.
URL: *https://www.persee.fr/doc/clao_0153-3320_1988_num_17_2_1272*

Bauer, R. S. (2018), 'Cantonese as written language in Hong Kong', *Global Chinese* **4**(1), 103–142.
URL: *https://www-degruyter-com.ezproxy.wlv.ac.uk/view/j/glochi.2018.4.issue-1/glochi-2018-0006/glochi-2018-0006.xml?format=INT*

Bauer, R. S. & Matthews, S. (2003), Cantonese, *in* G. Thurgood & R. J. LaPolla, eds, 'The Sino-Tibetan Languages', Routledge London and New York, pp. 146–155.

Cao, S., Lu, W., Zhou, J. & Li, X. (2018), cw2vec: Learning chinese word embeddings with stroke n-gram information, *in* 'Thirty-second AAAI conference on artificial intelligence'.

Cavnar, W. B., Trenkle, J. M. et al. (1994), N-gram-based text categorization, *in* 'Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval', Vol. 161175, Citeseer.

Census And Statistics Department (2019), Use of language, Technical report.
URL: *https://www.censtatd.gov.hk/fd.jsp?file=B11302662019XXXXB0100.pdfproduct*$_i d = C0000086 lang = 1$

Chan, C.-h. & Fu, K.-w. (2017), 'The relationship between cyberbalkanization and opinion polarization: Time-series analysis on facebook pages and opinion polls

during the hong kong occupy movement and the associated debate on political reform', *Journal of Computer-Mediated Communication* **22**(5), 266–283.

Chen, J., Hu, Y., Liu, J., Xiao, Y. & Jiang, H. (2019), Deep short text classification with knowledge powered attention, *in* 'Proceedings of the AAAI Conference on Artificial Intelligence', Vol. 33, pp. 6252–6259.

Chung, R. (2019).
URL: *https://static1.squarespace.com/static/5cfd1ba6a7117c000170d7aa/t/5d5a081034ee4800018f4*

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L. & Stoyanov, V. (2019), 'Unsupervised cross-lingual representation learning at scale', *CoRR* **abs/1911.02116**.
URL: *http://arxiv.org/abs/1911.02116*

Dalal, M. K. & Zaveri, M. A. (2011), 'Automatic text classification: a technical review', *International Journal of Computer Applications* **28**(2), 37–40.

Devlin, J., Chang, M., Lee, K. & Toutanova, K. (2018), 'BERT: pre-training of deep bidirectional transformers for language understanding', *CoRR* **abs/1810.04805**.
URL: *http://arxiv.org/abs/1810.04805*

Edwards, A. (2013), '(how) do participants in online discussion forums create 'echo chambers' ?: The inclusion and exclusion of dissenting voices in an online forum about climate change', *Journal of Argumentation in Context* **2**(1), 127–150.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A. & Mikolov, T. (2018), Learning word vectors for 157 languages, *in* 'Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)'.

Grimmer, J. & Stewart, B. M. (2013), 'Text as data: The promise and pitfalls of automatic content analysis methods for political texts', *Political analysis* **21**(3), 267–297.

Groves, J. M. (2010), 'Language or dialect, topolect or regiolect? A comparative study of language attitudes towards the status of Cantonese in Hong Kong', *Journal of Multilingual and Multicultural Development* **31**(6), 531–551.
URL: *https://doi.org/10.1080/01434632.2010.509507*

HaCohen-Kerner, Y., Ido, Z. & Ya' akobov, R. (2017), Stance classification of tweets using skip char ngrams, *in* 'Joint European Conference on Machine Learning and Knowledge Discovery in Databases', Springer, pp. 266–278.

Jaki, S., De Smedt, T., Gwóźdź, M., Panchal, R., Rossa, A. & De Pauw, G. (2019), 'Online hatred of women in the incels.me forum: Linguistic analysis and automatic detection', *Journal of Language Aggression and Conflict* **7**(2), 240–268.

Kato, Y., Kurohashi, S., Inui, K., Malouf, R. & Mullen, T. (2008), 'Taking sides: User classification for informal online political discourse', *Internet Research* .

Kewalramani, M. & Seth, R. (2020), 'Networked protests & state responses: The case of hong kong 2019-20', *Takshashila Discussion Document, Discussion Document* **3**, 12.

Kobayashi, T. (2012), 'Social media and fragmented social reality(<special issue>twitter and social media)', *Journal of Japanese Society for Artificial Intelligence* **27**(1), 51–58.
URL: *https://ci.nii.ac.jp/naid/110008898262/en/*

Kobayashi, T. & Ikeda, K. (2009), 'Selective exposure in political web browsing: Empirical verification of 'cyber-balkanization' in japan and the usa', *Information, Communication & Society* **12**(6), 929–953.

Lample, G. & Conneau, A. (2019), 'Cross-lingual language model pretraining', *arXiv preprint arXiv:1901.07291* .

Lee, F. L. F., Yuen, S., Tang, G. & Cheng, E. W. (2019), 'Hong kong' s summer of uprising: From anti-extradition to anti-authoritarian protests', *China Review*

19(4), 1–32.
URL: *https://www.jstor.org/stable/26838911*

Lee, F. L., Liang, H. & Tang, G. K. (2019), 'Online incivility, cyberbalkanization, and the dynamics of opinion polarization during and after a mass protest event', *International Journal of Communication* **13**, 20.

Lee, J. L. (2015), 'Pycantonese: Cantonese linguistic research in the age of big data', *Talk at the Childhood Bilingualism Research Centre, the Chinese University of Hong Kong* .

Lee, K. S. & Leung, W. M. (2012), 'The status of Cantonese in the education policy of Hong Kong', *Multilingual Education* **2**(1), 2.

Li, Y., Wang, X. & Xu, P. (2018), 'Chinese text classification model based on deep learning', *Future Internet* **10**(11), 113.

Liao, J., Sun, F. & Gu, J. (2019), Combining concept graph with improved neural networks for chinese short text classification, *in* 'Joint International Semantic Technology Conference', Springer, pp. 205–212.

Liu, B., Li, X., Lee, W. S. & Yu, P. S. (2004), Text classification by labeling words, *in* 'AAAI', Vol. 4, pp. 425–430.

Liu, X. (2018), 'A Comparative Study of Language Attitudes in Hong Kong: Towards English, Cantonese and Putonghua', *International Journal of English Linguistics* **8**(3), 195–209.

Liu, Y. (2015), 'Yellow or blue ribbons: analysing discourses in conflict in the televized government-student meeting during the occupy movement in hong kong', *Chinese Journal of Communication* **8**(4), 456–467.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. & Stoyanov, V. (2019), 'Roberta: A robustly optimized bert pretraining

approach', *arXiv preprint arXiv:1907.11692* .

Luo, X., Ohyama, W., Wakabayashi, T. & Kimura, F. (2011), A study on automatic chinese text classification, *in* '2011 International Conference on Document Analysis and Recognition', IEEE, pp. 920–924.

Luo, X., Ohyama, W., Wakabayashi, T. & Kimura, F. (2012), Impact of word segmentation errors on automatic chinese text classification, *in* '2012 10th IAPR International Workshop on Document Analysis Systems', IEEE, pp. 271–275.

Malte, A. & Ratadiya, P. (2019), 'Evolution of transfer learning in natural language processing', *arXiv preprint arXiv:1910.07370* .

Matakos, A., Terzi, E. & Tsaparas, P. (2017), 'Measuring and moderating opinion polarization in social networks', *Data Mining and Knowledge Discovery* **31**(5), 1480–1505.

Melville, P., Gryc, W. & Lawrence, R. D. (2009), Sentiment analysis of blogs by combining lexical knowledge with text classification, *in* 'Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining', pp. 1275–1284.

Mirończuk, M. M. & Protasiewicz, J. (2018), 'A recent overview of the state-of-the-art elements of text classification', *Expert Systems with Applications* **106**, 36–54.

Mou, L., Meng, Z., Yan, R., Li, G., Xu, Y., Zhang, L. & Jin, Z. (2016), 'How transferable are neural networks in nlp applications?', *arXiv preprint arXiv:1603.06111* .

Ng, K. Y., Feldman, A. & Leberknight, C. (2018), Detecting censorable content on sina weibo: A pilot study, *in* 'Proceedings of the 10th Hellenic Conference on Artificial Intelligence', pp. 1–5.

Norman, J. (2003), The Chinese dialects: Phonology, *in* G. Thurgood & R. J. LaPolla, eds, 'The Sino-Tibetan Languages', Routledge London and New York, pp. 72–83.

Pan, S. J. & Yang, Q. (2010), 'A survey on transfer learning ieee transactions on knowledge and data engineering', *22 (10): 1345* **1359**.

Poria, S., Cambria, E. & Gelbukh, A. (2015), Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis, *in* 'Proceedings of the 2015 conference on empirical methods in natural language processing', pp. 2539–2544.

Purbrick, M. (2019), 'A report of the 2019 hong kong protests', *Asian Affairs* **50**(4), 465–487.
URL: *https://doi.org/10.1080/03068374.2019.1672397*

Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. (2018), 'Improving language understanding by generative pre-training'.

Rao, A. & Spasojevic, N. (2016), 'Actionable and political text classification using word embeddings and lstm', *arXiv preprint arXiv:1607.02501* .

Rho, E. H. R., Mark, G. & Mazmanian, M. (2018), 'Fostering civil discourse online: Linguistic behavior in comments of# metoo articles across political perspectives', *Proceedings of the ACM on Human-Computer Interaction* **2**(CSCW), 1–28.

Shen, A. & Chow, K. P. (2020), Time and location topic model for analyzing lihkg forum data, *in* '2020 13th International Conference on Systematic Approaches to Digital Forensic Engineering (SADFE)', IEEE, pp. 32–37.

Snow, D. (2004), *Cantonese as written language: The growth of a written Chinese vernacular*, Vol. 1, Hong Kong University Press.

Snow, D. (2008), 'Cantonese as written standard?', *Journal of Asian Pacific Communication* **18**(2), 190–208.
URL: *https://www.jbe-platform.com/content/journals/10.1075/japc.18.2.05sno*

Snow, D. (2013), 'Towards a theory of vernacularisation: insights from written Chinese vernaculars', *Journal of Multilingual and Multicultural Development* **34**(6), 597–610.
URL: *https://doi.org/10.1080/01434632.2013.786082*

Tripathy, A., Agrawal, A. & Rath, S. K. (2016), 'Classification of sentiment reviews using n-gram machine learning approach', *Expert Systems with Applications* **57**, 117–126.

Turney, P. D. (2002), 'Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews', *arXiv preprint cs/0212032* .

Wang, Y., Zhou, Z., Jin, S., Liu, D. & Lu, M. (2017), Comparisons and selections of features and classifiers for short text classification, *in* 'IOP Conference Series: Materials Science and Engineering', Vol. 261, IOP Publishing, p. 012018.

Wei, Z., Miao, D., Chauchat, J.-H. & Zhong, C. (2008), Feature selection on chinese text classification using character n-grams, *in* 'International Conference on Rough Sets and Knowledge Technology', Springer, pp. 500–507.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. & Le, Q. V. (2019), 'Xlnet: Generalized autoregressive pretraining for language understanding', *arXiv preprint arXiv:1906.08237* .

Yue, A. O. (2003), Chinese Dialects: Grammar, *in* G. Thurgood & R. J. LaPolla, eds, 'The Sino-Tibetan Languages', Routledge London, pp. 84–125.

Zhang, Z., Ye, Q., Zhang, Z. & Li, Y. (2011), 'Sentiment classification of internet restaurant reviews written in cantonese', *Expert Systems with Applications*

38(6), 7674 – 7682.

URL: *http://www.sciencedirect.com/science/article/pii/S0957417410015101*

Declaration

UNIVERSITY OF WOLVERHAMPTON

SCHOOL OF HUMANITIES: MA in Computational Linguistics

Name: Hei Man, Haley Fong

Date: 8 February 2021

Title: Blue or Yellow? An Investigation on Hong Kong Political Short Texts

Module Code: Dissertation 7LN007/UM1

Presented in partial fulfilment of the assessment requirements for the above award.

Supervisor: Dr. Le An Ha

Declaration:

(i) This work or any part thereof has not previously been presented in any form to the University or to any other institutional body whether for assessment or for other purposes. Save for any express acknowledgements, references and/or bibliographies cited in the work, I confirm that the intellectual content of the work is the result of my own efforts and of no other person.

(ii) It is acknowledged that the author of any project work shall own the copyright. However, by submitting such copyright work for assessment, the author grants to the University a perpetual royalty-free license to do all or any of those things referred to in section 16(i) of the Copyright Designs and Patents Act 1988 (viz: to copy work; to issue copies to the public; to perform or show or play the work in public; to broadcast the work or to make adaptation of the work). (iii) This project did not involve direct contact with human subjects, and hence did not require approval from the LSSC Ethics Committee.

Date: 8 February 2021

Student's signature: