Hidden Discourses in Twitter Data:

Examining the Limitations of Data Collection and Information Dissemination in a Climate of Repression

ANNA BLEAKLEY

A thesis submitted in partial fulfilment

of the requirements for the degree

New Media and Digital Culture

Utrecht University

a.e.bleakley@students.uu.nl // bleakleyanna@gmail.com

5785456

June 6, 2017

First Reader Imar de Vries

Second Reader Indrid Hoofd

Illustrations

Figures

Tables

Acknowledgements

I'd like to extend my gratitude to Jake Henderson for assisting me in the development of the Demonanna data mining tool (referred to this thesis as the self-developed tool) and to Yana Podkosova, whom provided me with advice on the handling Twitter data. Without the exchange of information with a content creator, this thesis would also not have come in this direction, therefore I'd like to extend my thanks to my friends in Thailand in exposing me to the *hidden transcripts* referred to in this thesis. Furthermore, I'd to thank Imar de Vries for the constant feedback on the content and structure of my thesis. Lastly, I'd like to extend a special thanks to the Twitter social networking platform, without which this thesis would have been finished at a much earlier date. Your updates have not helped in the process.

Abstract

Twitter is an information dissemination tool that has been utilized by news organizations and individuals alike to share relevant links and responses to important current events. It is recognized as having an informational focus amongst users, wherein events subsequently become trending topics. However, distinct language based user groups respond differently due to cultural manners in which the users engage with information as well as environmental factors, including political climate and algorithmic biases. This study analyzes how information relating to lese majeste is being communicated. By resurrecting Scott's *hidden discourse*, this thesis displays that within social data passive resistance does exist through *media tactics* that produce creative methods for online activism in a chaotic language only understood by subgroups. Twitter's algorithm has created a barrier to understand how data is archived. In this thesis, methodological considerations are taken in account to understand the manner in which information can be approached in a comparative analysis of English and a less-resourced language, Thai. Using a wide range of literature from software studies, anthropology, and new media, this thesis presents findings on lese majeste information behavior of activists in Thai and English language groups and offers a critical perspective on alternative channels of resistance. In this thesis, three key challenges for comparing two language groups using specific keywords are presented. Firstly, less-resourced languages have significantly smaller databases. Next, the key terms used to extract information are sensitive in the Thai political environment. Furthermore, user behavior may direct information flow to one language opposed to another. Additionally, there are algorithmic limitations in detecting keywords in language, resulting in the inability to identify participatory culture in digital spaces. Finally, using a mixed approach to data collection to derive information from these limitations, this thesis presents a methodological approach to identify a culture of information dissemination in a manner where resistance can exist under draconian laws.

Background

Thailand's controversial lese majeste[1] law has been the cause of detainment for many political activists. Consequently, this has resulted in widely critical news coverage both in Thailand and abroad. In one of the more extreme cases, a factory worker that posted on social media to disparage King Bhumibol Adulyadej's beloved dog was subject to a 37-year prison sentence on these charges. The law has actively stifled any form of criticism of the royal family since the Chakri Dynasty was founded in 1781. Because of the spread of social media platforms such as Facebook and Twitter, new laws have been derived and enforced. Namely, the Computer Crime Act has escalated offenses based on online behaviour including comments and 'likes' on these platforms. In a tweet from the Thai PBS, a government owned public broadcasting network, the news agency expressed the need to collaborate in the reporting of offenses to the royal family. The tweet contained instructions to refrain from 'liking', 'sharing', 'commenting', or clicking on critical links and urged users to report the owners, creators, or sharers of the content. After the various waves of instability in Thailand, public forms of protest were under scrutiny, leaving social media as a channel for Thai dissidents to engage in online activism against the ruling regime. Additionally, this also made it the most effective sphere for the regime to enforce their political agenda and manipulate civil sentiment.[2] These measures have had the intended effect of creating a collaboration of royal supporters to track and report offenders (witch-hunting), but counterproductively for the regime, have resulted in increased online dissent by critics. On Twitter, groups and individual activists have created channels for the dissemination of links, comments, and retweets – filling the Twittersphere with a stream of resistance with sarcastic, informative, and resistant voices. In examining Twitter as a platform for voicing lese majeste discourse, differences in the Thai- and English-speaking user groups can be derived. This thesis aims to examine the lese majeste Twittersphere and draw parallels and differences on an English and Thai dataset. The nature of this thesis is exploratory, and therefore results will stand to elucidate restrictions and possible avenues for future research for Thai as a less-resourced languages along with alternative social discourses in in repressive climates.

[1] The term lese majeste has variations on spelling i.e. lèse-majesté, lese majesty, lese-majesty. The term is The term lese majeste has variations on spelling i.e. lèse-majesté, lese majesty, lese-majesty. The term is derived from French and means offense of the majeste. The most popular circulation of social media sites have utilized the chosen spelling 'lese majeste', therefore throughout this thesis the spelling 'lese majeste' will be used.

[2] Pinkaew Laungaramsri, "Mass Surveillance and the Militarization of Cyberspace in Post-Coup Thailand," *Austrian Journal of South-East Asian Studies* 9, 2 (2016): 197.

1. Introduction

Twitter holds a wealth of information including 'real world' and 'real time' responses to current events. During the Arab Spring and the Occupy Wall Street movement, Twitter played an integral role as both a means of communication, as well as for the dissemination of news.[3] Short messages that are generated on Twitter provide a collection of data that maps out the digital geography of social discourses.[4] The ability that Twitter has to display real time comments and opinions allows the platform to assimilate a kernel of information in 140 letters or less that sheds light on political sentiment and citizens' activism. Resulting in a growing interest among the research community to understand online discourses utilizing social media data in the fields of social science, humanities, and other related fields. However, the adoption of the use of Twitter data for research purposes has mainly taken place in the developed world. Therefore, the adoption of digital methods for Twitter data scraping and lexical resources has been concentrated within tools that aid linguistic data analyses using the roman alphabet. Since 2009, few English research papers have considered less-resourced languages in their datasets that have evaluated the Twitterspheres or investigated tools to analyse them.[5] Asian languages therefore suffer from the scarcity of resources like sentiment, lexicon, or corpus, resulting in a lack of research groups that focus on Asian- and Indo-Aryan languages.[6] Despite these limitations, there is methodological value to understand phenomena in these less-resourced languages to investigate alternative discourses relating to differences of information dissemination surrounding an event. Because an increasing number of tweets are written in languages other than English, it is becoming increasingly important to process tweets in these languages.[7] . As online conversations become important markers of ideological movements, it is important to acknowledge the language used to shape

[3] Shamanth Kumar., Fred Morstatter., and Huan Liu, *Twitter data analytics*, (New York: Springer Science and Business Media, 2013), 1.

[4] Alan Mislove., Sune Lehmann, Yong-Yeol Ahn., Jukka-Pekka Onnela., and J. Niels Rosenquist, "Understanding the Demographics of Twitter Users," (paper presented at International AAAI Conference on Web and Social Media, Barcelona, Spain, July 17 – July 21, 2011), 1.

[5] Ulrich Bügel and Andrea Zielinski, "Multilingual Analysis of Twitter News in Support of Mass Emergency Events," *International Journal of Information Systems for Crisis Response and Management* 5,1 (2013), accessed January 2, 2017, doi: 10.4018/jiscrm.2013010105.

Nikola Ljubešic., Darja Fišer., and Tomaz Erjavec, "TweetCaT: a tool for building Twitter corpora of smaller languages," (proceedings to the ninth international conference on language resources and evaluation, Reykjavik, Iceland, May 26 – 31, 2014).

Lichan Hong., Gregorio Convertino., and Ed H. Chi, "Language Matters In Twitter: A Large Scale Study," (paper presented at International AAAI Conference on Web and Social Media, Barcelona, Spain, July 17 – July 21, 2011).

Courtenay Honeycutt and Susan C. Herring, "Beyond microblogging: Conversation and Collaboration via Twitter," (proceedings of the 42nd International Conference on System Sciences, Hawaii, USA, January 5 – 8, 2009).

Choochart Haruechaiyasak and Alisa Kongthon, "LexToPlus: A Thai Lexeme Tokenization and Normalization Tool," (4th Workshop on South and Southeast Asian NLP (WSSANLP), International Joint Conference on Natural Language Processing, Nagoya, Japan, October 14-18, 2013).

[6] "Sentiment Analysis for Asian Languages," accessed February 10, 2017, http://www.amitavadas.com/SAAL.

[7] Frank Jacobs, "539-Vive le tweet! A map of Twitter's languages," accessed February 8, 2017, http://bigthink.com/strange-maps/539-vive-le-tweet-a-map-of-twitters-languages.

the political discourses that are extracted from Twitter – the platform that maintain these digital infrastructures.

1.1 Objectives and Aims

These facts leave an important question unanswered: How can the selection of language affect the perception of participatory culture on Twitter when examining lese majeste? This thesis will explore a case study of a language with a non-roman alphabet, Thai, via Twitter to understand a political event – in this case, current lese majeste detainments and restriction to freedom of speech on social media platforms in Thailand. The aim is to create an exploratory exercise to understand how information collection can be affected by keyword input in two languages and examine the discursive consequences of information perception along with reactions to draconic measures of repression within online spaces. In this thesis, insights will be taken from the data collection process to identify characteristics of technological reappropriation[8] of the Twitter platform based on information flow (or in some cases, lack of information flow) in the select languages. Through carefully documenting methodological choices and closely examining limitations, this thesis will assess information flow relating to lese majeste to illuminate the digital linguistic parochialism that undermine less-resourced languages in Twitter's digital social sphere.

The collection of data will be done using a self-developed[i] tool to scrape data based on the lese majeste term (to be extrapolated upon later). This method serves as a gateway to closely examine the different manners in which Twitter users in both language groups disseminate information and express sentiment in different political climates. And furthermore, examine Twitter as a platform that dictates the perceptions of lese majeste, which may differ depending on language. These questions aim to shape the future scope of research in discourses on Twitter in the manner in which they are archived and how this can affect the manner a phenomenon in perceived relating to 'findability' on Twitter.

This thesis is structured as follows: firstly, repression will be discussed in relation to existing theories, in cultural studies, anthropology, and software studies in order to dissect the limitations concerning the Twitter algorithm and communication practices that are an affordance of the Twitter platform. Different sub-sections within this literature review will discuss Twitter research in the Thai Twittersphere and how repression in this Twittersphere has resulted in self-censorship among users because of fear of scrutiny. Additionally, Twitter as a medium for data collection in the digital humanities will be discussed via a literature review that will examine research on Twitter relating to existing approaches to analysing Twitter data. By reviewing previous works that examine multi-language databases and power structures prevalent on social media platforms, the following section will outline the importance of this research methodology in the field of new media, social, and technological relevance. Thirdly, to draw out problems and limitations, a methodology of existing modes of analyses will be discussed in relation to the retrieved data. Fourthly, a methodological

[8] This thesis will address reappropriation, as the means of 'remixing' content forms of media that deviates from the intended use. The configurability of technology has afforded this form of reappropriation, where users have agency in the usage of a platform/technology.

proposal for analysing the datasets with a self- developed tool will be provided along with a description of data from the Thai and English datasets and features of the 2-step classification system. Then, the results will be discussed in relation to nominal differences in the databases – this section will compare reoccurring users in the databases and users' preferred language. In the qualitative section, insights will be taken from the deep data analysis to understand spaces of communications and digital activism using the background theories that will be discussed in section 2. In the last section, preliminary findings will be presented to provide insights and avenues for future research.

Figure 1 Overview of system for handling data.

2. Literature Review and Existing Theories

2.1 Manufacturing Consent and the Media: Explaining Theories Behind Event-Based Data

As Sankaranarayanan et al.[9] notes, Twitter has been a provider of news content and an arena for the expression of opinions on current news topics. This makes Twitter an excellent candidate for tracking news internationally and locally. In mainstream media, the micro-blogging platform has been considered a reliable source of gathering information used to legitimize existing claims – news organizations have utilized Twitter as a source for newsworthy information because of its awareness factor. It is then used to find evidence to support existing sources and to produce or reproduce news relevant content. The open source nature of the platform can simultaneously serve as both the democratization of

[9] Jagan Sankaranarayanan., Hanan Samet,., Benjamin E. Teitler., Michael Lieberman., and Jon Sperling,"TwitterStand: news in tweets," (proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Washington, USA November 04 - 06, 2009).

information or as a means of control by the state institutions or general populist sentiment. The model for the dissimilation of news content is closely tied to what Chomsky and Herman define as the *manufacturing of consent*: "institutions that carry out a system-supportive propaganda function, by reliance on market forces, internalized assumptions, and self-censorship, without overt coercion."[10] Twitter therefore makes a particularly interesting case in displaying existing perspectives under this system, because it is an arena that documents state sponsored propaganda, individual opinions, and news sources. The wide range of possibilities that Twitter offers provides various groups with differing incentives to approach information through its features and wide availability. Chomsky and Herman are not the only scholars that investigated the use of system supported propaganda in media spaces. In his investigation on alternative spaces in media, the media theorist David Garcia introduces the concept of *tactical media*: "media of crisis, criticism and opposition [..] tactical media are never perfect, always in becoming, performative and pragmatic, involved in a continual process of questioning the premises of the channels they work with."[11] Garcia and Lovink's definition describes an emergence of an ecology of new media activism that utilizes weak resistance and digital consumer technologies. In the early days of the web, electronic disobedience *tactical media* was used to bring together experiments by activist, artists, media practitioners, and theorist to create new styles of resistance through a collaborative effort in consolidated digital structures of power.[12] The initiatives apparent in *tactical media* can be seen as an effort in expressing political dissent with temporary reversals of power.[13] These tactics appropriate existing technology for a resistance oriented means to implement socio-political change. In a way, these tactics reappropriate technology to fit the means of the means of users, although it is not the intended use of the technology itself. Reappropriation tactics therefore serve as a manner in which digital platforms can be used to broaden the discursive sphere.

In examining sphere prevalent in social spaces scholars such as James C. Scott identify *hidden transcript*, spaces undetected by hierarchical forms of power. James C. Scott's vision of a *hidden transcript* describes a form of 'invisible power', or in his words "the conditions in which they do, or do not find public expression"[14] – beneath the surface of public accommodation to the existing distribution of power, wealth, and status presents.[15] In the digital age, class relations are deemed to less visible because of space offered by social platforms.[16] In these platforms, user representation in cyberspace allow users to impart on information and become involved in discussions and engagement, literature regarding this

[10]Edward S Herman, and Noam Chomsky, *Manufacturing Consent*. (New York: Pantheon Books), 306.
[11] "The Concept of Tactical Media", last modified March 7, 2017, http://www.tacticalmediafiles.net/articles/44999.
[12] "The Concept of Tactical Media", last modified March 7, 2017, http://www.tacticalmediafiles.net/articles/44999.
[13] Michael Dieter, "FCJ-126 The Becoming Environmental of Power: Tactical Media After Control," *The Fibreculture Journal* 18 2011: Trans (2011).
[14] James C. Scott, *Domination and the Arts of Resistance: Hidden Transcripts* (New Haven and London: Yale University Press, 1990), 14.
[15] James C. Scott, *Domination and the Arts of Resistance: Hidden Transcripts* (New Haven and London: Yale University Press, 1990), 15.
[16] Zizi Papacharissi, *A Private Sphere: Democracy in a Digital Age,* (Cambridge :Polity, 2010), 104.

interaction illustrates a visible effect of these technologies on society. The goal here is not to discuss the invisible, but the hidden. In examining *hidden transcripts* in Twitter, the interface and the manner in which information is archived is brought to light – keywords and linguistic tricks are deemed to be the façade that hide resistance. The public sphere does not always display all discourses in a manner viewable to the general audience. Scott's *hidden transcripts* describes this process of passive resistance hidden from the oppressors. His examples arise from peasant resistant in South East Asia. Through his observations, it is noted that there is a change in discourse in relation to power. Through technological change I argue that these behaviours are supported on social platforms, where users behave differently through modifying user behaviour to stay undetected by search classifiers. In the age of social media, hidden communication among sub-groups are integral to the survival of resistance movements. In these spaces, dissemination is archived and documented as social data, leaving a watermark of usernames that allow content to be traceable, compromising the identities of users. In these *hidden transcripts*, *tactical media* utilizes playful tricks and tactics to collectively create a response that fuses chaos and humour. The use of these tactics has been recently documented in protest movements in the Arab Spring and on Iranian social networking platforms. Using internet memes and casual internet jokes users provide an alternative narrative by redefining sociability into alternative media forms. The examples of content dissemination in discrete channels using artful and creative communication forms, display a space where dialogue transforms a nature of a public space in the digital form. Within this cycle, internet memes and casual internet jokes allow users to provide an alternative narrative, the user begins with redefining cultures of sociability into alternative media forms, such as those mentioned.

2.2 The Thai Twittersphere and Social Media as Liberation

In relation to studies of the Thai Twittersphere, there have been very few publications despite the wide spread use of Twitter as a tool for communication. A study carried out by Vongsoasup and Iijima[17] outlined how events affected the use of Twitter as a form of political communication, demonstrating that censorship affected the dissemination of information on Twitter. The authors focused on the political climate in Thailand considering limited internet freedom. Their results pointed to the nature of tweets as a form of information dissemination. Other studies by Thai scholars utilized Twitter to understand the social media climate in Thailand. In examining the interrelation of the state and social media that contributes to the creation of a *cyber-dystopia*, Laungaramsri introduces social media as a 'digital panopticon' that has transformed cyberspace into a cyber-dystopia as a result of the 2014 coup. In his analysis of the Thai cybersphere, he remarks that new laws such at the Computer Crime Act were one of the first measures used to penalize activists that utilized social media. The presence of lese majeste laws in Thailand has transformed civilian and government relations

[17]Naphatsorn Vongsoasup and Junichi Iijima, "What is a Role of Twitter in Thai Political Communication?," (proceedings to the 19th Pacific Asia Conference on Information Systems (PACIS 2016), Chiayi, Taiwan, 27 June- 1 July, 2016).

within the social sphere. Literature on social media political transformation has displayed social media engagement as a form of liberation, however in the case of Thailand these very platforms that liberate users are the ones in which make them liable for acts of dissent, dissemination, and activism. This very same consequence is reflected in Morozov (2012, 2013). He states that new media can enable the consolidation of authoritarian regimes as they are used by the state for mass surveillance, repression, and the dissemination of propaganda. The internet's political role is further examined in the Thai context with Laungaramsri's mention of cyber scouts, a government run program created to recruit online volunteers to protect the image of the monarchy in the online sphere by reporting 'disrespectful' websites.[18] The phenomena of witch-hunting, also fuelled by the *Garbage Collecting Organization (GCO)*, an active organization that assists people in *witch-hunting* activities and identifies those deemed to be 'garbage', to impose conformity to the ultra-royalist ideology. [19] Witch-hunting activities included methods to impose online scrutiny by harassing and punishing people with different views that disseminate anti-monarchy information online.[20] These activities have included, sending threatening messages to social media users that disseminated forms of anti-sentiment.

2.2 Online News Sources and Their Algorithms

Social dynamics alone are not the sole determiner of online power structures and the movements that emerge from them, algorithmic structures of online social networks can influence perspectives on user engagement and the appearance of online discourses. In David Beer's work on algorithms, he explains how the algorithm's predictive power has influenced expressions of power. Algorithms can express and enable authority, through their algorithmic capabilities, it's not data input per se that influence results but rather the algorithms themselves that provide purpose and direction to big data. The perception of the algorithm and its role in indexing and searching structures has allowed it to shape knowledge and produce outcomes by prioritising, sorting, filtering data relevant to the user.[21] Because algorithms assess 'authority' through search terms, they can sort and prioritize the media we encounter by having results appear in a certain order. In algorithmic studies, authors such as Kitchin have taken critical approaches in understanding algorithmic processes. In Kitchin and Dodge's 'Code/Space', they examine the production of code to create space. Algorithms can limit or delimit the creation of social spaces, the integration of these algorithms have given life to cultural activities in certain areas or transformed existing processes.[22] After all, by design algorithms are an automated system that make decisions, although they are not visible or 'black boxed' in most cases, their source code is the agency that provides one with data.

[18] Pinkaew Laungaramsri, "Mass Surveillance and the Militarization of Cyberspace in Post-Coup Thailand," *Austrian Journal of South-East Asian Studies* 9, 2 (2016): 204.
[19] Pinkaew Laungaramsri,, "Mass Surveillance and the Militarization of Cyberspace in Post-Coup Thailand," 205.
[20] Pinkaew Laungaramsri,, "Mass Surveillance and the Militarization of Cyberspace in Post-Coup Thailand," 206.
[21] David Beer, "The Social Power of Algorithms," *Information, Communication and Society 20* (2017): 1, accessed 23 May, 2017, doi: 10.1080/1369118X.2016.1216147.
[22] Rob Kitchin and Martin Dodge, *Code/Space* (Cambrdige: MIT Press, 2011), 3.

Twitter has become a popular source to keep up with news content (over 59% of Twitter users use the site to keep up with news), content produced on Twitter has the authority to act as a reliable news source.[23] This statistic outlines the agency Twitter's algorithm has on information seekers, that use the platform as a search engine to read responses and extract information from user generated content. In engaging with Twitter there are challenges depending on the languages used to extract information – the research methodology used along with approaches taken to tackle information are dependent on tools and the availability of resources. The following sub-section will discuss previous research that has utilized self-developed tools for mixed language data and research relating to information behaviour in various languages.

2.3 Multi-Languages Databases: Challenges, and Filtering Methods

Multi-language research has been proven to be a challenging field because of the lack of tools that are adjusted to languages with different alphabets and varying manners in which users create content. However, various authors have taken up this challenge by becoming involved in the process of creating the tools to examine these languages or studying the communication differences that occur based on language based user groups. Ljubešic., et al [24]presents a linguistic tool to analyse less-resourced languages in Croatian, Serbian, and Slovene. Through the creation of the TweetCaT tool the researchers created a tool that would retrieve a larger source of data for these languages. By using seed terms and simple language identification modules, the authors find new users as well as new tweets from already known users that tweet in the target language. In Hong et al's[25] study on the top ten popular languages used in tweets, they discovered that there were cross-language differences in the adoption of URLs, hashtags, mentions, replies, and retweets. Although there have been studies in the multiple languages used in tweets, most studies have analysed languages quantitatively through evaluating other languages that are used popularly in the international Twittersphere.[26] Few researchers have studied the cultural differences in multiple datasets

[23]" News Use Across Social Media Platforms 2016," Pew Research Center, last modified May 26, 2016, http://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016/.

[25] Lichan Hong., Gregorio Convertino., and Ed H. Chi, "Language Matters In Twitter: A Large Scale Study," (paper presented at International AAAI Conference on Web and Social Media, Barcelona, Spain, July 17 – July 21, 2011), 519.

[26] Choochart Haruechaiyasak and Alisa Kongthon, "LexToPlus: A Thai Lexeme Tokenization and Normalization Tool," (4th Workshop on South and Southeast Asian NLP (WSSANLP), International Joint Conference on Natural Language Processing, Nagoya, Japan, October 14-18, 2013).

"Twitter reaches half a billion accounts more than 140 millions in the US," Last modified Febuary 24, 2010, http://semiocast.com/downloads/Semiocast_Half_of_messages_on_Twitter_are_not_in_English_20100224.pdf.

despite the internationality of Twitter.[27] In the words of Hong et al.,[28] researchers have mainly studied how the users themselves are affected by cultural differences and the manner in which they use online tools.[29] This signifies the concentration of research on language and the association with cultural differences. Researchers that have taken the path of exploring these differences have not properly considered the restrictions that come with comparing data from languages with different language corpora. Their focus on Twitter has been as a cross-cultural communication tool, ignoring the methodological implications of language and the influence it has on keywords relating to phenomena, particularly content that can alter perceptions of user engagement.[30] Few studies have mentioned the input methods and their influence on results. Exceptionally, in Chakma and Das[31] evaluation of Hindi and English content. In their study, they outline the difficulties of analysing roman transliterations of Hindi on social media. They found that multiple languages could be integrated into user-generated content (UGC) due to improper input methods for their own language creating difficulties in filtering methods. Having discussed the limitations that relate to language research on Twitter, the literature review will outline the experimental approach of this thesis.

2.4 Relevance

The second wave of digital humanities has been primarily concerned with creating the environment and the tools to produce, generate, and interact with knowledge that is born digital. [32] In the case of Twitter data, this has led to research of the tools used for data retrieval, the biases of datasets, and the influence of algorithms in black boxing data. As the

[27] Akshay Java., Xiadan Song., Tim Finin., and Belle Tseng, "Why We Twitter: Understanding Microblogging Usage and Communities," (proceedings to the 9th WebKDD and 1st SNA-KDD 2001 workshop on web-mining and social network analysis, California, USA, August 12, 2007.)
Balachander Krishnamurthy., Phillipa Gill., and Martin F. Arlitt,"A Few Chirps About Twitter," (proceedings to the first workshop on Online Social Networks, Washington, USA, August 17-22, 2008).
[28]Lichan Hong., Gregorio Convertino., and Ed H. Chi, "Language Matters In Twitter: A Large Scale Study," (paper presented at International AAAI Conference on Web and Social Media, Barcelona, Spain, July 17 – July 21, 2011).
[29] Shipra Kayan., Susan R Fussell,. and Leslie D. Setlock, "Cultural Differences in the Use of Instant Messaging in Asia and North America," (proceedings of the 2006 20th anniversary conference on computer supported cooperative work, Alberta, Canada, November 4 – 8, 2006).
Susan Herring., John C. Paolillo., Irene Ramos-Vielba., Inna Kouper., Elijah Wright., Sharon Stoerger., Lois Ann Scheidt., and Benjamin Clark, "Language Networks on LiveJournal," (proceedings of the 40th Hawaii International Conference on System Sciences, Hawaii, USA, January 3 - 6, 2007).
Brent Hecht and Darren Gergle, "The Tower of Babel Meets Web 2.0: User-Generated Content and Its Applications in a Multilingual Context," (proceedings to the SIGCHI Conference on Human Factors in Computing Systems, Georgia, USA, April 10-15, 2010).
[30] Lichan Hong., Gregorio Convertino., and Ed H. Chi, "Language Matters In Twitter: A Large Scale Study," (paper presented at International AAAI Conference on Web and Social Media, Barcelona, Spain, July 17 – July 21, 2011).
[31] Chakma Kunal and Amitava Das, "CMIR: A Corpus for Evaluation of Code Mixed Information Retrieval of Hindi-English Tweets," *Computación y Sistemas* 20, 3 (2016), accessed 17 January, 2017, http://www.cys.cic.ipn.mx/ojs/index.php/CyS/article/view/2459/2178.
[32] Todd Presner, "Digital Humanities 2.0: A Report on Knowledge", accessed January 3, 2017 http://cnx.org/content/m34246/1.6/?format=pdf, 6.

use of datasets increases in the field of humanities[33], the need to evaluate the gaps in knowledge for creating a coherent analysis of the completeness of datasets has become more integral. In the words of Rogers,:

> Follow the methods of the medium as they evolve, learn from how dominant devices treat natively digital objects, and think along with those object treatments and devices so as to recombine or build on top of them. Strive to repurpose the methods of the medium for research that is not primarily or solely about online culture.[34]

Here Roger proposes that researchers should build upon new methods in analysing dominant devices to generate new methods for experimentation. Taking Roger's statement into account this thesis will integrate the use of a self-developed tool for data mining and to create a necessary examination of alternative discourses that can co-exist within Twitter along with the influential factors that can affect our perception of information engagement among users. Because Twitter data is used to infer public interest pertaining to a topic – it is increasingly important to structure research with consideration to user behaviour surrounding a topic of scrutiny. The social relevance of this examination would be the general perception of engagement based on language – the purpose here is to understand information behaviour with alternative avenues used to spread dissent. The previous sections have examined the multiple challenges and perspectives in relation to research design in language with consideration to dominant power structures, highlighting the need for experimentation through tools for data mining as well as considerations for data scarcity. Tweets are not as telling as they are and there are methodological and technological considerations that need to be considered to create a balanced analysis that considers environmental circumstances and scarcity in language.

This thesis will use the key terms defined in this section to dissect information derived from a deep data analysis of tweets. To create a qualitative analysis these key terms will be integrated in order to understand the manner in which activists circumvent detection in virtual public spaces.

---

[33] Christine Borgman, "The Digital Future is Now: A Call to Action for the Humanities," *Digital Humanities Quarterly* 3,4 (2009), accessed February 20, 2017, http://www.digitalhumanities.org/dhq/vol/3/4/000077/000077.html.
Danah Boyd and Kate Crawford, "Critical Questions for Big Data," *Information, Communication and Society* 1,5 (2012): 662–679, accessed February 21, 2017, http://doi.org/10.1080/1369118X.2012.678878.
Anne Burdick., Johanna Drucker., Peter Lunenfeld., Todd Presner., and Jeffrey Schnapp, *Digital Humanities* (Cambridge, Massachusetts: The MIT Press, 2012), accessed February 21, 2017, http://doi.org/10.1108/S20449968(2013)0000007006.
Joost Potting, "Completeness in Twitter Datasets A critical review on Twitter research methodologies," (Master's thesis, University Utrecht, 2016).
Lev Manovich, "Trending: The Promises and the Challenges of Big Social Data, " in *Debates in the Digital Humanities*," ed. by Matthew K. Gold (Minneapolis: University of Minnesota Press, 2011).

[34] Richard Rogers, *Digital Methods* (Cambridge, Massachusetts: The MIT Press, 2013), 5. Originally found in, Joost Potting, "Completeness in Twitter datasets A critical review on Twitter research methodologies," Master's thesis, University Utrecht, 2016, 5.

3. Methodology of Research and Implementation

The case study aims to understand the ecological system of tweets among English- and Thai- users based on the lese majeste law, along with the shortcomings of creating a comparative analysis of two different language datasets. By using a mixed methods research design, this methodological approach utilizes quantitative as well as qualitative methods to analyse the data collected. In this section data collection will be discussed in regard to the implementation of the self-developed tool and filtering methods used to create the final databases for both languages.

3.1 Data Collection

In order to retrieve the tweets, a self-developed tweet mining tool was used to scrape data through Twitter's search option. The tool was developed by Python modules that imported the tweets found on mobile.twitter.com, and therefore results appear in a chronological order based on Twitter's algorithm selection of tweets pertaining to the keyword. After the datasets were retrieved, the tweets were placed into CSV format. The file was then imported into an Excel Workbook with the UTF-8 (Unicode) encoding that would recognize Thai and English characters as well as emoticons. The time frame for the collection of data took place between January 23rd to March 3rd, 2017. This time frame was chosen because of ongoing discussions about the succession of the throne and the rising lese majeste cases since the death of the reigning monarch, King Bhumibol Adulyadej. Since the death of the king, more than 20 people have been charged for their online anti-royalty statements. During the process of data cleaning, duplicate tweets and those that didn't not contain written content (only links) were filtered out. The keyword classifiers were chosen as a measure to create an encompassing analysis around lese majeste tweets that included the term in single and mixed databases. Certain tweets encompassed a mix of English and Thai (the keyword would appear in English and the context of the tweet would be in Thai). The use of non-roman alphabets often integrated the use of roman letters in their online messages. These improper input measures created difficulties in extracting user content. Thai users were found to use keywords such as ม 112 or มาตรา 112, meaning article 112 or lese majeste along with lese majeste and parts of tweets in English. To objectively sort out tweets with mixed languages, Twitter's query operators and labelling the search language specific search option with e.g. lese majeste lang:th was used.[ii] The process is illustrated in Figure 2.

While compiling the data, it was noted that the English database contained a larger amount of data in comparison to the Thai database, the reason for this is that different languages on Twitter produce differing amounts of tweets on a daily basis. In a study done by Hong et al.[35] it was displayed that 51% of tweets collected on a 4-week basis were found to be in English, while other languages displayed varying percentages of the Twittercorpora collected.

[35] Lichan Hong., Gregorio Convertino., and Ed H. Chi, "Language Matters In Twitter: A Large Scale Study," (paper presented at International AAAI Conference on Web and Social Media, Barcelona, Spain, July 17 – July 21, 2011).

The meta data retrieved from the tweets in each dataset:

- Text: Extracted from the tweets itself.
- User Id: The Twitter user name of the person that posted the tweet.
- UTC Time: The date and time of the tweets (the self-develop tool standardizes the time of different time zones).

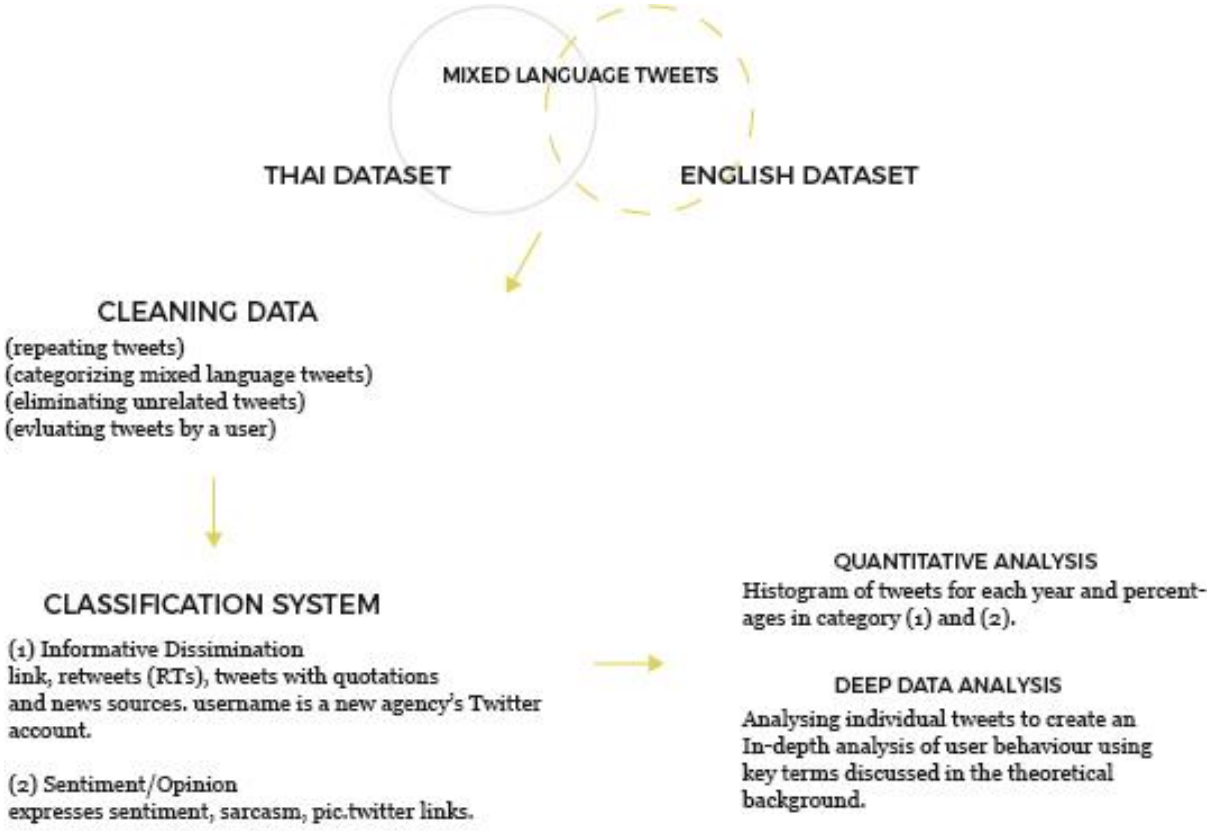

Figure 2 Modes of analyses and justification model.

4.3 Methodology of Case Study

The case study represents a political event and its social implications online. The decision to use the Thai and English Twittersphere stems from the researcher's expertise in the language. Lese majeste was used as a case study because it is an under researched area despite the rise of arrests for online content in social media platforms that were considered a breach of the law. Previous research on Thai media have also avoided this topic because of crackdowns on academics that have critically engaged with the topic.[36] In section 2.2, Laungaramsri (2016) has outlined key government sponsored initiatives to supress online dissent, but there is yet to be an examination of online spaces where dissent exists despite these repressive measures. Furthermore, by choosing Thai content as a case study, responses to a controversial topic can be considered along with methodological possibilities in

[36] Sam Micheal, "Thailand's Big Step Backwards," *The Diplomat*, March 5, 2015, accessed May 1, 2017, http://thediplomat.com/2015/03/thailands-big-step-backwards/.

comparing a less-resourced language to one that is larger resourced. Furthermore, there has been a scarce amount of research conducted on less – resourced languages. One of the reasons being the lack of tools and scare data. In this study, alternative channels that engage with content creators are proposed in order to understand alternative channels that users use to engage with information on Twitter. In this section, the foundation of the deep data analysis will be laid by introducing the 2-step classification method to find reoccurring patterns of information dissemination in both language groups to understand the appropriation of the platform for English and Thai language users. The difference in engagement with information flow will set the pillars for the deep data analysis (analysis of individual tweets) to reveal alternative avenues users have taken to engage with content. Additionally, ethical considerations are taken into account to preserve the anonymity of users in the databases.

4.5 Two Step Analysis or 2 Categorization

In this thesis, a two-step analysis method is utilized for analysing tweets based on information dissemination and sentiment/opinion in the year 2011 (the year with the largest amount of data for both datasets), to create the closest comparison of keyword results. In order to understand the way in which user groups engage with the topic of lese majeste, a manual examination of sentiment/opinion vs. information dissemination was implemented. Although it may be self-evident that English tweets are more likely to include information dissemination because English is the *lingua franca* of the world – this categorization method was introduced to understand the manner in which users engaged with lese majeste considering the existing censorship laws. Using this analysis, information behaviour can be examined statistically using the two predominant categories (1) information dissemination, and (2) sentiment/opinion. Information dissemination is a natural consequence to the engagement of news sources in online platform. In tweets with sentiment/opinion, attitudes that respond to lese majeste are outlined that involved sarcasm, questions, or links to personal media pic.twitter (often leads to other platforms such as Instagram). In the Thai database, sentiment involved personal questions and user engagement with the topic however, the content was in all cases neutral. Information dissemination: categorized by Twitter's various features links, mentions (@) ,and retweets (RTs). See examples below.

Example 1 Information dissemination tweet with link.

Thai Court Reduces Jail Time for Editor Convicted of Insulting Monarchy https://t.co/9cKuQeuFww

Example 2 Information dissemination tweet with a link to a news source.

GOOGLE NEWS - Australian tried for lese majeste - BBC News: AFPAustralian tried for lese majesteBBC. http://tinyurl.com/75ucc6

Example 3 English lese majeste tweet with a link to a news source

rt @user3: 'Police say they are investigating a total of 32 cases of lèse-majesté, the highest number in deca.. http://tinyurl.com/825or6

Example 4 Thai information dissemination tweet

Good morning! คำว่า lese majeste แปลว่ากฎหมายที่เกี่ยวกับการดูหมิ่นพระมหากษัตริย์ ที่คนไทยเรียกว่า มาตรา 112 (Article 112 ในภาษาอังกฤษ)

Translation of Example 4 Thai Information dissemination tweet

Good morning! Lese majeste means insulting the royal families Thai people call it Article 112 or Article 112 in English

Sentiment/opinion tweets involve positive, neutral, or negative sentiment. They are often responses with emotion and can sometimes be satirical. Thai tweets are noted to be neutral when utilizing sentiment/opinion. See examples below.

Example 5 Sentiment/Opinion Tweet (mixed language)

Jan has ended, These are the most boring words I've always seen Donald Trump Immigrants Inauguration Lèse-majesté WALL วนๆอยู่แค่นี้

Translation of Example 5 Sentiment/Opinion Tweet (mixed language)

Jan has ended, These are the most boring words I've always seen Donald Trump Immigrants Inauguration Lèse-majesté WALL we're not going to get further than this

Example 6 Thai Sentiment/Opinion Tweet

แท็ก #LM นี่บุหรี่หรือ lese majeste 55555

Translation of Thai sentiment/opinion tweet. LM here is referred to as a cigarette brand.

Is the #LM tax for the cigarette or lese majeste 55555

Example 7 English tweet with Sentiment/Opinion

@user3 I ban this brand name as the company fired the leader of union who wore a campaign t-shirt challenging lèse majesté law.

4.6 Ethics

In order to protect active activist and individuals that have not consented to this study, all names, retweets (RT) and references (with @) are anonymous. The examples given throughout this thesis will be provided without user names and end in order to order to protect the privacy of individuals that produce tweets on the sensitive topics of lese majeste. Tweets that have mentioned of other users will be provided using token username such as 'user1' in the place of a username within a tweet. The uploaded Twittercorpora for each respective language will also employ this method in order to provide the possibilities of these datasets being used in further research.

4. Results

The results are displayed in the following manner. First dates that are analysed will be discussed. For example, only certain quantitative numbers in this thesis examines the entirety of data collection. Naturally, there are discrepancies in dates in which data is provided by each language dataset .Then, the details of each database will be displayed in separate categories along with personal observations that were taken during the manual two-step categorization

process. The results that refer to reoccurring users will be discussed in relation to activists and communication and then the theoretical framework will be referred to in order to understand information behaviour among different user groups. In this section, the analyses of Twitter data collected from the 2011 databases are presented along with results and observations based on the two categorization systems: information dissemination and personal opinion. Data was mined from the years 2007-2017, however the self-developed tool presented results that differed based on the language of the search. For the Thai database, results went back to 18/06/2008 while the results in the English database went back to 05/04/2007. Dates for data analyses were then chosen in order to ensure that tweets would be consistent based on topics and current news events (as mentioned in the previous section). The following dates were chosen based on the availability of tweets in the Thai dataset (18/06/2016-15/07/2008). The histograms below display the frequencies of tweets in each year.

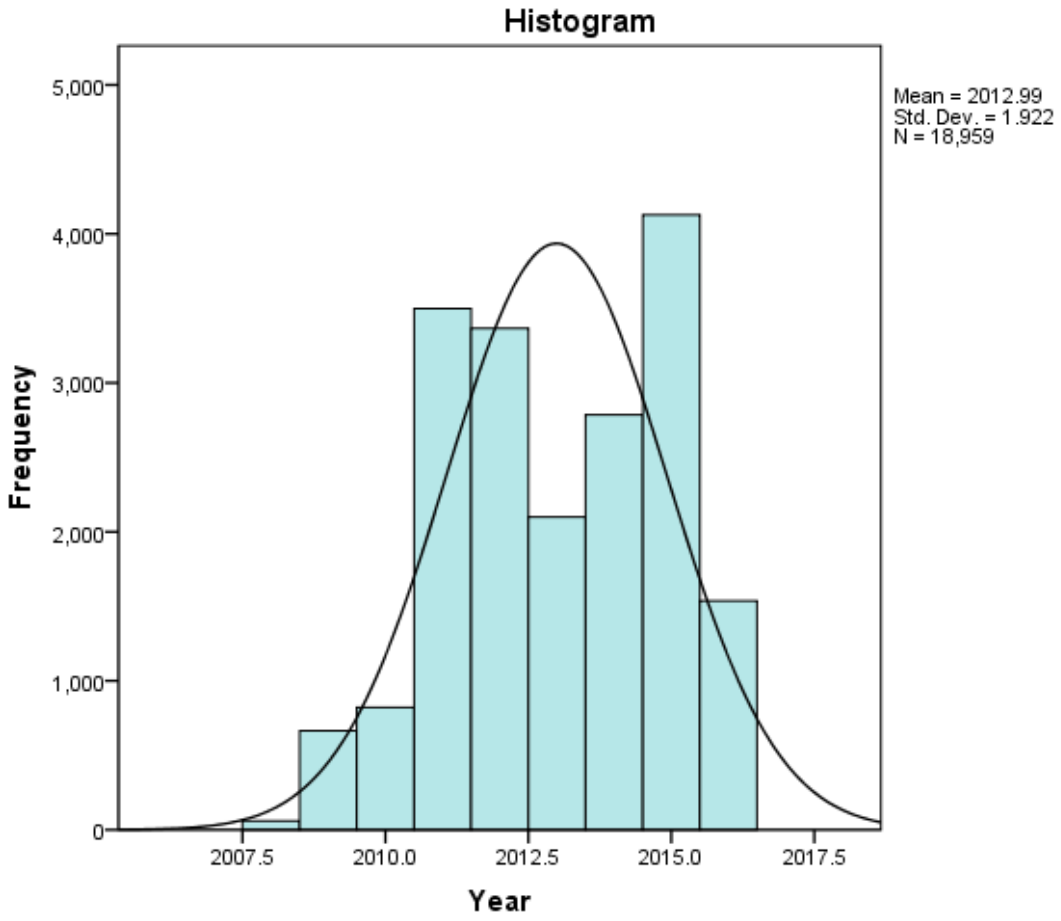

Figure 3 Histogram of English Database

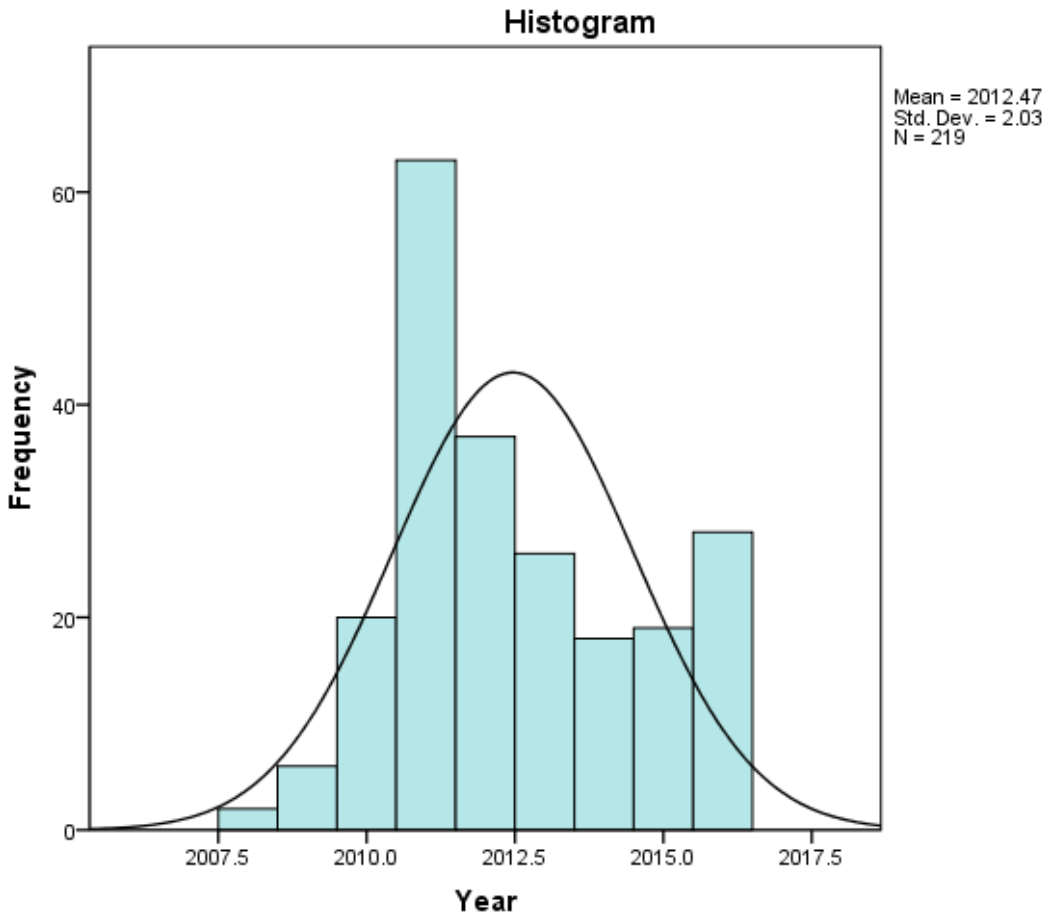

Figure 4 Histogram of Thai database

4.1 English Database

The English dataset consists of 6079 users and 18959 tweets cumulatively. The usernames that tweeted most were identified to be activists or news sources. This was identified by manually studying their Twitter profiles. These details are identified in the profile of the users by looking at previous tweets. In the process of examining the profile, some users were found to reoccur in both databases (further explored in section 4.4). In the English dataset, the largest categorization was the number of informative tweets, of which 80% of the database consisted, while 20% of tweets were categorized as personal sentiment. This displays that information dissemination behaviour was predominant in English, meaning that political participation in this language category were significantly higher than in the Thai database. The increased dissemination of information by users in the English database could suggest that there are some liberating values in disseminating information in English. Because of existing lese majeste law, the information in the Thai language is under increased scrutiny. Users that participated in information dissemination therefore could have disseminated information in English in order to actively engage with lese majeste by sharing news links as

well as using Twitter's other features to communicate on the platform without fear of detection.

	English Dataset	Thai Dataset
Informative Tweets	80%	39%
Sentiment/Opinion	20%	61%

Table 1 Percentages of Tweets that Correspond to the 2 Step Categorization (see section 4.5 for the criteria of this categorization)

4.2 Thai Database

The Thai database contains 151 users for 255 tweets in total. The user names that tweeted the most were identified to be activists. News sources that tweeted on the topic were found to be absent from the 2011 database, this is mostly likely due to censorship laws that have driven attention away from lese majeste convictions (particularly in Thai media). In the Thai lese majeste database (only 1.65% the size of the English database), it was noticeable that Twitter users utilized neutral sentiment in expressing their tweets: informative tweets that were categorized due to a tweet's ability to disseminate information with links to news articles or retweets, only made up 39% of the database. The most prominent category was option/sentiment which composed 61% of the Thai lese majeste databases. However, within this opinion/sentiment category definitions and references to foreign lese majeste laws were made, with 19% of these tweets referring to foreign lese majeste laws and 6.25% with definitions of the lese majeste word itself along with the pronunciation of lese majesties (e.g. เลส-มา-เจส-ท หรือ เลส-มา-เจส-เท, English equivalent les-e-ma-jes-te or les-ma-jes-ty). These responses were neutral and did not display positive and negative sentiment, but included laughter '555', which is a common cultural reoccurring used for laughter and the ';)' icon. These findings, demonstrate that users tend to openly engage with the content using the lese majeste classifier to draw comparisons on the lack of foreign laws mirroring lese majeste, however they avoid direct engagement with local lese majeste law because possible arrests and scrutiny by other users online.

4.3 Comparing User Behaviour in Different Languages: Activism and Communication

Activities that represent forms of resistance were found in the English dataset. These examples can be seen in the tweets below. The meta-information found on Twitter represents forms of passive and active engagement. For example, the category of information dissemination can be displayed as a manner in which users can engage with information. Although users disseminate information, this does not necessarily mean that they take an active role. In contrary, there are also active examples of information engagement through references to other Twitter users, through features such as 'RT@user' or in response to users in '@user'. These features provided by Twitter allow users to connect with other users within users group and disseminate information. Communication activities in the English database were found to utilize tweets that engaged users within groups, for instance a *call for action* tweet as well as questions and answers can be found prominently within the English database because of the larger number of tweets per user (average of 3.83 tweets per a user). Below

there are individual tweets that display communication initiatives for *call for action*, and efforts for coordination amongst users that are found in the English database.

Example 8 Communication among activist

@user1 Thai is running out of excuses for playing Abhisit's clone on lese majeste abuses. R. Amsterdam's silence is also deafening.

Example 9 A call for action retweet for a jailed lese majeste activist, Somyot Pruksakasemsuk

RT @user2: We repeat call for release of #Thailand activist & editor #Somyot Pruksakasemsuk who is in jail for lese-majeste https://t.câ€¦

Example 10 Tweet calling upon supporters to gather in support of lese majeste

Tomorrow supporters of lese majeste law to gather at Royal Plaza at 1.30pm, move to UN at 2pm and to US Embassy at 3pm /TANN

Users therefore refer to one another in a larger frequency in English. The same users can often be found in the English and Thai database; however, their use of information dissemination tools differ depending on the language utilized. For example, 39% percent of the users in the Thai databases could be found on the English database, the repeated users found in the databases exhibit a preference for English to engage with the global Twittercorpora. Looking at the frequencies in engagement among language groups, the large percentage of users that disseminate information in English suggests that the reoccurring users in the Thai database use English as an information broadcasting tool.

4.4 Alternative Communication Channels

The different sizes of the Thai and English database can be described as intentional. To further understand the large difference in number of tweets in the English and Thai database, an activist was asked why users preferred to disseminate information in English. In her response, in her words “when you use a hashtag, you want the world to see what you’re doing, that’s not the case with Thailand”, she stated that hashtags and keywords were for the “outside world”, and “informal language” was used in order to express sentiment discretely.[37] One example she gave was references to Pizza Company, which were used to engage with the lese majeste because of one common element. Lese majeste in Thai is often referred to as ‘112’, the references in to Pizza company were made because of the dial code (with slight similarities) of Pizza Company being 1112. One of these examples can be seen in a tweet below.

Example 11 Tweet clarifying Pizza Company reference.

Ordering **Pizza** Thailand can get you arrested. The **Pizza company** uses hotline **112**. Lese Majeste use the word "**Pizza**" to refer to article **112**.

Example 12 Media attached to the 112 references.

[37] Anonymous Employee of Amnesty International Thailand, Facebook call, March 25th, 2017.

The image attached to the tweet above display a form of user generated content or an inside joke. The finding demonstrates that although there are limitations in retrieving tweets for researchers attempting to scrape data on Twitter, investigations in these restrictions can reveal the reappropriation and use of *media tactics* within these social spaces by understanding new ways of communication outside of those classified by the search topics. By asking content creators directly, one can gain access to the *hidden transcripts* that are created in response to a topic that is under scrutiny. In this process, the examination of the lack of data in the Thai dataset sparked the drive to search for *hidden transcripts*. The autonomous fashion in which users can operate in a space that is determined by algorithmic rules that can process information and make decisions demonstrate that algorithmic rules are not all encompassing. Users still have agency in these online virtual environments by appropriating language and choosing to create content that will not be used as search queries under the topic of scrutiny. The method used to display these differences has revealed how *noise* in databases can interfere with the sampling of datasets, and projected interest in topics. However, the same *noise* can also serve as a *media tactic* to bypass digital censorship and give birth to artful creative forms of resistance that are disguised to create new media channels where dissent can be expressed in what can be described as a *hidden transcript*.

4.5 Vernacular Language in Noise

The tweets that are within these *hidden transcripts* are not random, they are culture specific and refer to common terms found in vernacular language. For examples references to 'olieng': dark coffee mixed with sugar and syrup – refer to jail visits in the context of tweets in examples 13-16. Along with 'olieng', 'crab fried rice', a common street dish is mentioned, which is also associated food visitors bring to inmates during visits. The examples below are coupled with English translations.

Example 13 Tweet disseminating media in example 12.

#พิซซ่า ไหมครัช อยากทานแบบไหน เป็นรู หรือเป็นชิ้นๆ **#pizza** #1112 **#112 #pizzacompany**
http://instagram.com/p/l_7XJfiPRB/

Translation of example 13.

#pizza? Which would you like to try the one with wholes of pieces
#pizza#1112#112#pizzacompanyhttp://instagram.com/p/l_7XJfiPRB/

Example 14 Tweet referencing to bringing food to jail visits to those detained because of lese majeste.

*หมายเหตุ คดี **112** ไม่มีการเข้าเยี่ยม โปรดเก็บโอเลี้ยงข้าวผัดปูของท่านไว้*

Translation of example 14.

For case 112, no one is allowed to visit convicts, keep your olieng and crab fried rice.

Example 15 Tweet referencing to food for jail visits of those detained because of lese majeste

*ใครโดน **112** ช่วงนี้อดโอเลี้ยงข้าวผัดนะครับ ทหารเขาไม่ให้เยี่ยม*

Translation example 15.

Anyone that is convicted for 112 now will miss out on the olieng and fried rice. The army doesn't allow visits.

Example 16 A Twitter user jokingly tweeting about lese majeste prisoners missing out on iced coffee.

@user1 ถ้าได้ **112** เดี๋ยวเราหิ้วโอเลี้ยงไปเยี่ยมนะ #ผิด 55

Translation example 16.

@user3 if you are detained under I'll visit you with ice coffee #wrong 55.

These tweets are not stand-alone cases, in Thailand, comments, or sarcastic remarks on social media could lead up to 15 years in jail, some examples of this include the conviction of a factory working for insulting the king's dog.[38] This can explain one of the reasons in which these tweets have referred to topics far-fetched from lese majeste to express sentiment. As mentioned by an activist source, a direct rationale in material cannot be identified. The tweets offer a manner to engage with the material, but offer no insight into personal opinion. Engagement, or interest are not always indicated with frequencies of keywords and can exist within *noise* among material where discourse is expressed through forms of media – memes, comments, retweets, and videos that become the language in which sentiment occurs. The affordance of Twitter as a space allows social discourses that exist in the real world to be transformed into an augmented space that affords both information dissemination, censorship, and activism. Twitter as a social media platform has a limited word count of 140 characters – this limitation reduces forms of sentiment and dissemination to the simplest form where keywords, hashtags, and external media links are the means in which tweets are disseminated. Because of Twitter's limitation on words, media that is produced on the platform is restrained. Therefore, those that attempt to bypass the design and appropriation of the platform is forced to reduce content to the simplest form. In the case study of lese majeste, these sentiments embody known icons and discrete cultural references specific to the country referred to in the examples above. In these forms of media, users integrate a discrete language only known to subgroups in digital terms through the Twitter platform. The limitations of expression on Twitter has created the unattended side-effect of an unstructured chaotic social sphere that sparks ritualized interactions that exist within a *hidden transcript*. Because the tweets are not informative nor meaningful to the public, they are a disguised as *noise*. The examples of tweets displayed in examples 13-16 display the mannerism used to diffuse information disguised as *noise*. The satirical nature of these tweets

[38] Oliver Holmes, "Thai man faces jail for insulting king's dog with 'sarcastic' internet post," *The Guardian*, December 15, 2015, accessed February 6, 2017, https://www.theguardian.com/world/2015/dec/15/thai-man-faces-jail-insulting-kings-dog-sarcastic-internet-post.

on a platform that is a vehicle of detection for dissent under a repressive regime show that there are no spacial confines to the existence of dissent under a platform where hegemonic power is at play. Twitter's hybrid space that allows for vernacular language to be undetected is the very example of street movements that utilize digital communication.

In the *hidden discourses*, authority is delegitimized. As famously quoted by Scott, "when the great lord passes the wise peasants bows deeply and silently farts".[39] The concept of resistance here is not direct and confrontational, but rather humorous in a way that those repressed can keep up appearances with those in power – public appearance will always conform to the correct behaviours expected by those in power. These profane forms of defiance highlight dissent and exist within communities to conceal and misrepresent aspects of social relations.[40] With the Thai lese majeste corpus, the dominating public tweets of '112' or '1112' as a reference to Pizza Company provides a manner in which resisting tweets could be disguised under a popular discourse. In the digital sphere, social control is depicted in a dystopic vision, where state control uses social media channels such as Twitter for arbitrary detention of activists and those that disseminate inflammatory or insulting comments. Countries with tough censorship laws such as Thailand push users to disseminate content in another language or using linguistic manoeuvres in order avoid detection. These tweets serve as a form of everyday resistance. In contrasting informational behaviour in English and Thai tweets, one can see that the English tweets are a direct form of user-generated behaviour with the intent of distributing facts. As with the Thai user group, the lack of tweets or the manner in which users generate personal opinions relating to foreign lese majeste law rather than local lese majeste law, or tweets that do not particularly address Thai lese majeste indicate false compliance, pilfering, and feigned compliance.[41] The linguistic tricks that the activist use in referring to culture specific references through food (olieng and khao pad) transform the tweet from a strategy of resistance to randomness or *noise* amongst the digital public sphere. The linguistic tricks that allow dissent to be disseminated through chaos outlines that beneath the surface of power, one can still respond with resistance passively.

The tweets that are disguised as *noise* are a form of quiescence that does not highlight participation or in this case, engagement with lese majeste content. Quiescence in this case is much like lack of political participation despite ongoing equalities in an open political system.[42] Each of these tactics I argue begin with the understanding of how dominant ideology works its magic by manipulating social reality via the dissemination of propaganda and values to explain subordination such as *witch-hunting,* the enforcement of (as mentioned in section 2.2) ultra-royalist content online and finger-pointing to impose conformity. The tactics used to justify subordination are circulated, in a form that manufactures consent among the general public, this drives the grievances of those subordinated to what Scott term

[39] James C. Scott, *Domination and the Arts of Resistance: Hidden Transcripts* (New Haven and London: Yale University Press, 1990), v.
[40] James C. Scott, *Domination and the Arts of Resistance: Hidden Transcripts* (New Haven and London: Yale University Press, 1990), 71.
[41] James C. Scott, *Domination and the Arts of Resistance: Hidden Transcripts* (New Haven and London: Yale University Press, 1990), 188.
[42] James C. Scott, *Domination and the Arts of Resistance: Hidden Transcripts* (New Haven and London: Yale University Press, 1990), 71.

'realm of the possible'.[43] Here, subordinates or users realize what is realistic and unrealistic to drive aspirations to the realm of the possible.[44]

4.6 Weapons of the Weak in Twitter's Ustopic Space

Looking at the two user groups, the English lese majeste Twitter sphere has been largely influenced by information dissemination as a form of empowerment. The *call to action* tweets are a manner in which users engage with their global audience. In section 2.1, *manufacturing consent* is discussed in Chomsky's description of a supportive propaganda system through the use of information dissemination to propagate information through self-censorship on Twitter. In light of the climate of repression and *witch-hunting* on Twitter, the social platform can facilitate repression and the counter-movement against it. The power-structure of Twitter describes the affordances of digital media technology for the dissemination of opinions and idea. With the current political situation in Thailand, the very use of Twitter's filters which is dominated by keyword archiving display Chomsky's and Herman's hypothesis of dominant media hierarchies even within the liberating digital sphere. The digital landscape has seen the rise of self-reported news with *digital journalism* and *collective action* through social media platforms, however despite these liberating promises social data is still experience the repression found in the social sphere.

Although the digital age provides researchers, analyst, and everyday users with access to social media platforms to extract information on the social, personal, and political level. Through the use of keywords, social data becomes readily available through categories that are archived via Twitter. The manner in which tweets are registered via hashtags and keywords becomes an in-built system in which users and researchers extract information for marketing, political insight, and a behavioural understanding Twitter's corpus. In identifying, unachievable data, or in this case *noise*, data can be interpreted as significant to political resistance. This brings a *hidden transcript* in the virtual realm under the lens of an academic analysis that can make sense of the nonsensical and humorous response specific to a user group. Twitter's affordance of facilitating digital coercion and dissidence in a single space is an example of a dual utopic and dystopic visions of big data. Although repression is prominent in Twitter's social space, online freedom still has a space; by camouflaging dissident tweets amongst popular brands and keywords that are only recognizable amongst individuals in culturally specific groups, users can express a form of freedom. These *hidden transcripts* are the hidden discourses that can blend opinion of dissidence within the common popular discourse.

[43] James C. Scott, *Domination and the Arts of Resistance: Hidden Transcripts* (New Haven and London: Yale University Press, 1990), 74.
[44] James C. Scott, *Domination and the Arts of Resistance: Hidden Transcripts* (New Haven and London: Yale University Press, 1990), 74.

5. Conclusion

5.1 Dissemination of Tweets

The results in this study confirm that Twitter is a platform for information dissemination. However, the manner in which information dissemination is undertaken is more prominent in the English language in comparison to the Thai database. The large amount of data in the English database displays the uneven amount of meta-data that is extracted from Twitter in more popular languages in comparison to less-resourced ones through the use of conventional keyword searches. Although this observation is in some way self-evident, the following analyses have explored other possibilities of viewing alternative channels of information dissemination and content creation through surveying a new approach. By surveying a quantitative and qualitative approach for data analysis that accounts for alternative channels, this thesis presents differences based on user groups in their information behaviour using the lese majeste keyword. The rich information found in the English database suggests that activists and citizens that follow lese majeste topics disseminate information relating to awareness in English rather than Thai. With 80% of tweets in the English database relating to the category and only 39% in the Thai database. These results could be due to the large English audience, which raises questions about local engagement with the topic. Some factors that affect these quantitative significances in information behaviour may be self-censorship in the Thai language due to existing government sponsored programs that target dissent online and the preference of English as the *lingua franca* of the world. As demonstrated by the alternative channels of information dissemination through the use of *media tactics – noise* under other keywords is also a form of lese majeste discourse that does not directly utilize the keyword; but rather disguises it under popular icons and cultural references to food. The usage of cultural references to refer to lese majeste challenge the very appropriation of Twitter as an information dissemination tool that can detect sentiment and engagement. Users in this case have utilized the platform in an unconventional way to engage with lese majeste indirectly, without tagging the tweets under lese majeste, to avoid detection and choosing to not utilize Twitter's functions to bring to light their dissent. This tactic has changed the nature of Twitter as a digital space that can cater to movements that facilitate dissent through unconventional resistance in opposed to a platform that acts as a digital panopticon. The nature in which protest or resistance is made through chaos in the form of *noise* in data, is in many ways the utilization of *media tactics* within Twitter.

5.2 Limitations that Come with the Re-appropriation of Twitter

The examples in this thesis presents the restrictions in methodology for studying lese majeste data along with counter movements created to reappropriate Twitter's virtual spaces. When examining dissemination, methodological considerations in evaluating Twitter as an information dissemination platform need to consider how Twitter operates as a social media platform through the archiving of keyword based meta-data and the implications this can have, especially in a repressive political climate. The methodological corpus of this thesis

demonstrates the limitations of quantified data when examining social data. By using a qualitative lens to approach a less-resourced language on a topic where censorship and propaganda is prevalent, an examination of the culture of participation on Twitter relating to an lese majeste keyword can be made, revealing alternative spaces of activism and a participatory culture that is created in response to system supported propaganda and censorship in digital social spaces. In examining the lese majeste, the examples suggest that despite statistical methods used, sample sizes from Twitter can be influenced heavily by *noise* that create difficulties for data collection pertaining to a topic. Keywords are an integral function of Twitter's search algorithm, the manner in which tweets appear on one's newsfeed is influenced by previous clicks and online behaviour through preferences of topics, which are determined by keywords. Because Twitter's search algorithm is black boxed, it is difficult to determine the selection process behind the results to keyword searches. In order to create a methodological framework that can account for cultural differences in response to events and different cultural climates, the user needs to become integrated into the analysis of data.

Limitations in the unequal retrieval of data based on the keywords serve as a drive to survey alternative methods for data retrieval. In this case study, it is demonstrated that tweets relating to the same event can be disguised under other tags or topics. For example, in the use of the Pizza Company or "pizza" in general as a response to the harsh lese majeste sentences and convictions, individuals are able to respond collaboratively in a hidden manner. Unfortunately, this process is difficult to document because of niche parsing software for extracting the context of these tweets; even with the ability to teach modules to parse topic specific tweets, there are limitations in existing Python modules, new modules may also struggle with the nature of these tweets being unclear – one cannot tell if the user is engaging with a topic when language is unclear and understood solely among members belong to these sub-groups. The collection of these tweets can be done in-depth interviews with content creators and manually by a researcher.

5.3 Problems of Access and Twitter Spaces

Twitter's massive growth and the large amount of Twitter users present challenges for accessing Twitter data and interpreting the stream of data effectively. The research represents a space in between the *cyber-dystopia* that has led to the crackdown of activist and sparked a ustopic (utopic and dystopic) space where repression and resistance exist alongside of each other, never meeting eye to eye. There is a public discourse that is prevalent were tweets display submission and acceptance (within Thai user groups) and a *hidden discourse* that is neither directly labelled nor visible to those that do not belong to sub-groups disseminating the content. The manner in which tweets become visible and invisible in entirely up to its algorithmic makeup. Twitter's algorithm determines the tweets that are displayed within the popular feed, therefore creating an automated selection bias for the completeness of tweets that fall under the category of a topic. If a user chooses to purposely disguise a tweet under another category, this information would only be available to sub-groups, the user themselves, or those that understand references to unrelated keyword.

In creating a program that scrapes tweets using Twitter's search function, the focal point of the study was to understand the perception of data – in this case, it was done by retrieving results based on key terms to understand perceptions relating to a topic under scrutiny. In the 'computational turn' in the humanities, research need to assess the tools that retrieve the data as well as the algorithmic processes that influence data output. Because Twitter has become a virtual space, various discourses exist and can be influenced by methodological decisions in research. The surveying methods of this thesis represents the exploration of the tools that scrape that data along with the social platforms that host the information – this exploration has led to results that make-up a *hidden discourse* that would not come into view without inside knowledge of cultural references and individuals in this circle of information.

5.4 Further Research

Exploring lese majeste through Twitter's search functions presents both possibilities and restrictions. Firstly, this thesis's use of a self-developed tool to mine tweets presents results in a different manner to existing Twitter research that has relied mainly on the Twitter open API that retrieves data 18 days back. In presenting Twitter's search results to selected keywords, this study provides an overview of how data can be scraped with alternative tools that offer an experimental method to retrieve sample data from various years. A benefit of this mode of analysis is that the collection of data is focused on what the general perception of tweets on a keyword throughout various years; the self-develop tool in this thesis can be used to further research that explores tweets throughout years. The methods used in this thesis to detect *noise*, has also brought a new approach to data filtering methods in databases that question the relevancy of tweets pertaining to a topic. Engagement with content makers to examine keywords in repressive environments has also been hi-lighted as an effective channel to lay the groundwork to examine *hidden discourses* in digital spaces.

The use of Thai to create a database of tweets can also further research in natural language processing to train future Python modules (NLTK library for sentiment) in Thai language processing to detect sub-Unicode libraries for Thai. The manual process of categorization in the 2-step method also exposes a system of categorization that can be modified into a module to create a general filter to use in comparative language studies of Twitter data. The starting point of this thesis in uncovering the lack of engagement on the basis of topic – in studying information behaviour relating to lese majeste, restrictions have become the lens that have driven investigations to *media tactics* and *hidden transcripts* to uncover information behaviour in these alternative channels that exists as *noise* under unrelated topics. The presentation for the discourses of *media tactics* in the form of *noise* and alternative channels for the dissemination of lese majeste related sentiment provides a manner in which hidden discourses exist within Twitter data. Furthermore, this study also lays the groundwork for studies relating to the reappropriation of content creation and digital passive resistance through identifying *noise* in data as a manner of participation with content. In this thesis, problems of access have been bypassed because of technological expertise, but there are few alternatives that exist outside retrieving data via the API. Although data extraction via Firehose is an option, it is rather costly. The technical restrictions of this paper present an overview of possible obstacles of information access that Twitter provides to users and

researchers. In identifying the barriers of information access via Twitter, tools can be built to enable researchers with less technological expertise to access Twitter data in less-resourced languages. Although Twitter's search engine algorithm remains unknown, the findings of this study present new channels in which research repressive climates and technological limitations can examine data.

6. References

Anonymous Employee of Amnesty International Thailand. Facebook call. March 25th, 2017.

Beer, David."The Social Power of Algorithms." *Information, Communication and Society 20* (2017): 1-13. Accessed 23 May, 2017. doi: 10.1080/1369118X.2016.1216147.

Borgman, Christine. "The Digital Future is Now: A Call to Action for the Humanities." *Digital Humanities Quarterly* 3,4 (2009): 1–30. Accessed February 20, 2017. http://www.digitalhumanities.org/dhq/vol/3/4/000077/000077.html.

Boyd, Danah and Kate Crawford. "Critical Questions for Big Data." *Information, Communication and Society* 1,5 (2012): 662–679. Accessed February 21, 2017. http://doi.org/10.1080/1369118X.2012.678878.

Bügel, Ulrich and Andrea Zielinski. 2013. "Multilingual Analysis of Twitter News in Support of Mass Emergency Events." *International Journal of Information Systems for Crisis Response and Management* 5 (1): 77-85. doi: 10.4018/jiscrm.2013010105.

Burdick, Anne., Drucker, Johanna., Lunenfeld, Peter., Presner, Todd, and Jeffrey Schnapp. *Digital Humanities*. Cambridge, Massachusetts: The MIT Press, 2012. Accessed February 21, 2017. http://doi.org/10.1108/S20449968(2013)0000007006.

Chakma Kunal and Amitava Das. "CMIR: A Corpus for Evaluation of Code Mixed Information Retrieval of Hindi-English Tweets." *Computación y Sistemas* 20, 3 (2016): 425-434. http://www.cys.cic.ipn.mx/ojs/index.php/CyS/article/view/2459/2178.

Chomsky, Noam and Herman, Edward S. *Manufacturing Consent*. New York: Pantheon Books, 1988.

Dieter, Michael. "FCJ-126 The Becoming Environmental of Power: Tactical Media After Control." *The Fibreculture Journal* 18 (2011). http://eighteen.fibreculturejournal.org/2011/10/09/fcj-126-the-becoming-environmental-of-power-tactical-media-after-control/

Frank, Jacobs. "539-Vive le tweet! A map of Twitter's languages." Accessed February 8, 2017. http://bigthink.com/strange-maps/539-vive-le-tweet-a-map-of-twitters-languages.

Haruechaiyasak, Choochart and Alisa Kongthon. "LexToPlus: A Thai Lexeme Tokenization and Normalization Tool." In the 4th Workshop on South and Southeast Asian NLP (WSSANLP), International Joint Conference on Natural Language Processing, Nagoya, Japan, October 14-18, 2013.

Hecht, Brent and Darren Gergle. "The Tower of Babel Meets Web 2.0: User-Generated Content and Its Applications in a Multilingual Context." Proceedings to the SIGCHI Conference on Human Factors in Computing Systems, Georgia, USA, April 10 – 15, 2010.

Herring, Susan., Paolillo, John C., Ramos-Vielba, Irene., Kouper, Inna., Wright, Elijah., Stoerger, Sharon., Scheidt, Lois Ann., and Benjamin Clark. "Language Networks on

LiveJournal." Proceedings of the 40th Hawaii International Conference on System Sciences, Hawaii, USA, January 3 - 6, 2007.

Holmes, Oliver. "Thai man faces jail for insulting king's dog with 'sarcastic' internet post." *The Guardian*, December 15, 2015. Accessed February 6, 2017, https://www.theguardian.com/world/2015/dec/15/thai-man-faces-jail-insulting-kings-dog-sarcastic-internet-post.

Honeycutt, Courtenay, and Susan C. Herring. "Beyond microblogging: Conversation and Collaboration via Twitter." Proceedings of the 42nd International Conference on System Sciences, Hawaii, USA, January 5 – 8, 2009.

Hong, Lichan., Convertino, Gregorio, and Ed H, Chi. "Language Matters In Twitter: A Large Scale Study." Paper presented at International AAAI Conference on Web and Social Media, Barcelona, Spain, July 17 – July 21, 2011.

Java, Akshay., Song, Xiadan., Finin, Tim., and Belle Tseng. "Why We Twitter: Understanding Microblogging Usage and Communities." Proceedings to the 9th WebKDD and 1st SNA-KDD 2001 workshop on web-mining and social network analysis, California, USA, August 12, 2007.

Kayan, Shipra., Fussell, Susan R., and Leslie D. , Setlock. "Cultural Differences in the Use of Instant Messaging in Asia and North America." Proceedings of the 2006 20th anniversary conference on computer supported cooperative work, Alberta, Canada, November 4 – 8, 2006.

Krishnamurthy, Balachander., Gill, Phillipa., and Martin F. Arlitt ."A Few Chirps About Twitter." Proceedings to the first Workshop on Online Social Networks, Washington, USA, August 17-22, 2008.

Shamanth, Kumar., Morstatter, Fred., and Huan Liu. *Twitter Data Analytics*. New York: Springer Science and Business Media, 2014.

Ljubešic, Nikola., Fišer, Darja., and Tomaz Erjavec. "TweetCaT: a tool for building Twitter corpora of smaller languages." Proceedings to the ninth international conference on language resources and evaluation, Reykjavik, Iceland, May 26 – 31, 2014.

Laungaramsri, Pinkaew. "Mass Surveillance and the Militarization of Cyberspace in Post-Coup Thailand." *Austrian Journal of South-East Asian Studies* 9, 2 (2016): 195-214. https://aseas.univie.ac.at/index.php/aseas/article/download/1356/1473.

Manovich, Lev. "Trending: The Promises and the Challenges of Big Social Data. " In *Debates in the Digital Humanities*," edited by Matthew K. Gold, 452-459. Minneapolis: University of Minnesota Press, 2011.

Michael, Sam. "Thailand's Big Step Backwards." *The Diplomat*, March 5, 2015. Accessed May 1, 2017. http://thediplomat.com/2015/03/thailands-big-step-backwards/

Morozov, Evgeny. *The Net Delusion: The Dark Side of Internet Freedom*. New York: Public Affairs, 2012.

Morozov, Evgeny. *To Save Everything, Click Here: The Folly of Technological Solutionism*. New York: PublicAffairs, 2013.

Mislove Alan., Jørgensen Sune Lehmann., Ahn Yong-Yeol., Onnela Jukka-Pekka., and J. Niels Rosenquist. "Understanding the Demographics of Twitter Users." In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, Barcelona, Spain, July 17 – 21 2011.

Papacharissi, Zizi A. *A Private Sphere: Democracy in a Digital Age*. Cambridge: Polity, 2010.

Pew Research Center. "News Use Across Social Media Platforms 2016." Last modified May 26, 2016. http://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016/

Potting, Joost. "Completeness in Twitter Datasets A critical review on Twitter research methodologies." Master's thesis, University Utrecht, 2016.

Rob Kitchin and Martin Dodge. *Code/Space*. Cambridge: MIT Press, 2011.

Rogers, Richard. *Digital Methods*. Cambridge, Massachusetts: The MIT Press, 2013.

Sankaranarayanan, Jagan., Samet, Hanan., Teitler, Benjamin E., Lieberman, Michael., and Jon Sperling."TwitterStand: news in tweets." Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. Seattle, Washington, November 04 - 06, 2009.

Scott, James C. *Domination and the Arts of Resistance: Hidden Transcripts*. New Haven and London: Yale University Press, 1990.

Siam Voices. "Thai journalist under investigation amid lese majeste complaint." *Asian Correspondent*, 27th May 2012. Accessed January 4, 2017. https://asiancorrespondent.com/2012/05/without-fear-or-favor-journalist-pravit-rojanaphruk-under-investigation-for-charges-of-lese-majeste/.

Semiocast ."Twitter reaches half a billion accounts more than 140 millions in the US." Last modified Febuary 24, 2010. http://semiocast.com/downloads/Semiocast_Half_of_messages_on_Twitter_are_not_in_English_20100224.pdf.

Tactical Media Files. "The Concept of Tactical Media." Last modified March 7, 2017, http://www.tacticalmediafiles.net/articles/44999.

The 25th International Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems. "Sentiment Analysis for Asian Languages." http://www.amitavadas.com/SAAL/.

Vongsoasup, Naphatsorn and Junichi Iijima. "What is a Role of Twitter in Thai Political Communication?." Proceedings to the 19th Pacific Asia Conference on Information Systems (PACIS 2016), Chiayi, Taiwan, 27 June- 1 July, 2016.

8. Endnotes

[i] To request the self-developed tool either contact the author or a description can be found here http://jake.henderson3.net/.
[ii] Link for Twitter's query operators https://dev.twitter.com/rest/public/search

9. Appendix

9.1 Glossary of Terms

API/ Application Programing Interface

A protocol and tool integrated into the process of building software that allow different counterparts to interact with one another.

Black boxed

Term used to refer to software programs and machines where the process is unknown to the public. Source code and details on their construction are not often revealed to the public.

Hashtag

Keywords preceded on Twitter with the '#' sign found in tweets.

Module

A file consisting of Python code, a module can define functions, classes and variables.

Tweet

A 140-character message on the Twitter social media website.

Twitter

A 140-character found on the Twitter micro-blogging website.

Twittersphere

All Twitter users belonging to a specific country or region.

Twittercorpora

The plural of a Twitter corpus that is composed to tweets and meta-data.

9.2 Code for the Self-Developed Tool

The code for the self-developed tool was developed in the Pycharm code editor and features 511 lines of code. The API keys here (consumer keys) here are changed for safety precautions of the authors and those involved. The tool can be requested via http://henderson3.net/ or http://annableakley.com/work/

```
dom
import re
import os
import sys
import urllib.request
import urllib.parse
import urllib3
import queue
import datetime
from timeit import default_timer as timer
from selenium import webdriver
from TwitterSearch import *
from collections import Counter
from nltk.corpus import stopwords
from textblob import TextBlob


# The DemonWordCount class creates total count of all the various words (minus
commonly used so-called "stop words") used in all the mined tweets.
class DemonWordCount:
    __count = Counter()
    __isWritten = False
    __stop_words = None
    __search_name = None
    __isAdded = False
    __punct = [".", ",", "!", "?", ":", ";", "&", "'", "\"", "-", "_", "`", "…"]

    def __init__(self, s):  # The class is initialized by getting the search name
for formatting the output file, as well as placing NTTK's stop-words in an array.
        self.__stop_words = set(stopwords.words("english"))
        self.__search_name = s

    def add_to_counter(self, s): #For each tweet, the individual words are first
extracted, then each word, assuming it's not a stop word or a puncuation mark, it's
added to the counter,
        s = s.split(" ")
        for ii in s:
            k = str(ii)
            try:
                if k[-1] in self.__punct:
                    k = k[:-1]
                if k[0] in self.__punct:
                    k = k[1:]
            except:
                pass
            if k not in self.__stop_words and k not in self.__punct and k != "":
                self.__count[k] += 1
                self.__isAdded = True

    def write_count_file(self):  # After all the tweets have been processed,
everything in the counter is formatted and written to a text file.
        global fname
        if self.__isAdded is False:
            return
        try:
            dwc_fname = str(fname) + "_DemonWordCount"
            if os.path.isfile(dwc_fname + ".txt"):
                dwc_fname += "_" + str(datetime.datetime.now().strftime("%Y-%m-
```

```
%d_%H.%M.%S"))
            demon_word_count_file = open(dwc_fname + ".txt", "wb")
            demon_word_count_file.write(bytes("Demonanna Word Count File for: " +
str(self.__search_name) + "\r\n\r\n", "utf-8"))
            for ii in self.__count.most_common():
               demon_word_count_file.write(bytes(str(ii[0]) + ": " + str(ii[1]) +
"\r\n", "utf-8"))
            self.__isWritten = True
            self.__count = None
            demon_word_count_file.close()
            print("\nWrote Demonanna Word Count file to \"" + dwc_fname + ".txt\"")
         except Exception as e:
            print("\n\nOh dear! There's a problem writing the Demonanna Word Count
file...")
            print(str(e))
            tb = sys.exc_info()[2]
            print("Error at line: " + str(tb.tb_lineno))

      def __del__(self):  # In case there's a problem, anything that's already in the
counter will be written.
         if self.__isWritten is False:
            self.write_count_file()


def get_sent(p):  # This function is used for converting sentiment and polarity to
a more user-friendly ranking.
      r = p
      if p < -0.75:
         r = "Extremly negitive"
      if p < -0.5:
         r = "Very negative"
      if p < -0.25:
         r = "Negative"
      if p < 0:
         r = "Slightly negative"
      if p == 0:
         r = "Equal"
      if p > 0:
         r = "Slightly positive"
      if p > 0.25:
         r = "Positive"
      if p > 0.5:
         r = "Very positive"
      if p > 0.75:
         r = "Extremly positive"
      return r


htmlremove = re.compile(r"<.*?>")  # Pre-compile the regular expression for
identifying HTML tags in order to improve speed.


def formatcsv(s):  # Make sure all text is suitible for CSV format
      s = str(s).replace('\n', '')
      s = str(s).replace('\r', '')
      s = htmlremove.sub("", s)
      s = str(s).replace("&amp;", "&")
      s = str(s).replace("&#39;", "'")
      s = str(s).replace("&gt;", ">")
      s = str(s).replace("&lt;", "<")
      s = str(s).replace('"', '""')
      s = str(s).replace('&quot;','""')
      s = str(s).replace("&amp;", "&")
      return s


# The following 3 functions convert Twitter's month and time naming scheme to a
numerical value,
```

```
# so that Excel, etc. can recognize the timestamp.
def getM(m):
   if m=="Jan":
      m = "1"
   elif m=="Feb":
      m = "2"
   elif m=="Mar":
      m = "3"
   elif m=="Apr":
      m = "4"
   elif m=="May":
      m = "5"
   elif m=="Jun":
      m = "6"
   elif m=="Jul":
      m = "7"
   elif m=="Aug":
      m = "8"
   elif m=="Sep":
      m = "9"
   elif m=="Oct":
      m = "10"
   elif m=="Nov":
      m = "11"
   elif m=="Dec":
      m = "12"
   return m


def formatapitime(s):
   # print(s + "\n")
   s = str(s).replace('\n','')
   s = str(s).replace('\r','')
   s = str(s).replace('"','""')
   s = s.split()
   m = s[1]

   m = getM(m)

   y = s[5]
   d = s[2]
   t = s[3]

   return str(y + "-" + m + "-" + d + " " + t)


def formatsitetime(s):
   s = str(s).replace('\n', '')
   s = str(s).replace('\r', '')
   if s[-1:] == "m":
      tm = datetime.datetime.now() - datetime.timedelta(minutes=int(s[:-1]))
      s = tm.strftime("%Y-%m-%d %H:%M:%S")
   elif s[-1:] == "h":
      tm = datetime.datetime.now() - datetime.timedelta(hours=int(s[:-1]))
      s = tm.strftime("%Y-%m-%d %H:%M:%S")
   else:
      ss = s.split(" ")
      if len(ss) == 3:
         s = "20" + str(ss[2]) + "-" + str(getM(ss[1])) + "-"
         s += str(ss[0])
      elif len(ss) == 2:
         m = int(getM(ss[0]))
         d = int(ss[1])
         y = datetime.datetime.now().year
         if datetime.datetime(y, m, d) > datetime.datetime.now():
            y -= 1
         s = str(y) + "-" + str(m) + "-"
         s += str(d)
```

```
      else:
         s = str(s).replace('"', '""')
   return s

#This function is used for retrieving tweets from Twitter's website.
def usessite(p, q):
   global i
   try:
      # The search URL is formulated corresponding to the user's search query
      url = "https://mobile.twitter.com/i/nojs_router?path=%2Fsearch%3Fq%3D"
      qq = urllib.parse.quote_plus(q)
      qq = str(qq).replace("+", "%2B")
      url += qq
      # Chromedriver (https://sites.google.com/a/chromium.org/chromedriver/) is
started for interacting with the Twitter website
      chrome_path = "chromedriver.exe"
      chrome_options = webdriver.ChromeOptions()
      chrome_options.add_argument('--disable-extensions')
      driver = webdriver.Chrome(chrome_path, chrome_options=chrome_options)
      driver.set_window_position(-10000, 0)
      driver.get(url)
      dwc = DemonWordCount(q)
   except Exception as e:
      print("\n\nOh dear! There's a problem...")
      print(str(e))
      tb = sys.exc_info()[2]
      print("Error at line: " + str(tb.tb_lineno))
      return

   try:
      print("Page:")
      sys.stdout.write("1...")
      sys.stdout.flush()
      # Loop through all the pages according to the user input
      for ii in range(0, p):
         # Check if there are any search results on this page, otherwise exit
         if len(driver.find_elements_by_class_name("noresults")) > 0:
            sys.stdout.write("nothing!\n")
            sys.stdout.flush()
            break
         # Get all the tweets, users, and timestamps on the page.
         tweets = driver.find_elements_by_class_name("tweet-text")
         users = driver.find_elements_by_class_name("user-info")
         timestamps = driver.find_elements_by_class_name("timestamp")
         # There should be an equal amount of tweets, users, and timestamps...if
not there's a problem.
         if len(tweets) != len(users) != len(timestamps):
            raise Exception("Didn't find the the same number of tweets, users, or
timestamps. Twitter must've changed their site. :(")
         elif len(tweets) == 0:
            sys.stdout.write("nothing!\n")
            sys.stdout.flush()
            break
         # At this point, there are definitely some tweets to parse.
         # It loops through all the tweets and while formulating the line for the
CSV file
         # Then, it adds to the counter (see the class above), and writes to the
CSV file
         for t in range(0, len(tweets)):
            i += 1
            txt = '"' +
formatcsv(users[t].find_element_by_class_name("username").text) + '",'
            txt += '"' + formatcsv(tweets[t].text) + '",'
            txt += '"' + formatsitetime(timestamps[t].text) + '",'
            tb = TextBlob(tweets[t].text)
            txt += '"' + str(get_sent(tb.sentiment.polarity)) + '",'
            txt += '"' + str(get_sent(tb.sentiment.subjectivity)) + '"\n'
            tf.write(bytes(txt, "UTF-8"))
```

```
            dwc.add_to_counter(tweets[t].text)

         sys.stdout.write("complete.\n")
         sys.stdout.flush()
         # If there are still pages left to parse, it loads the next page and
continues the process.
         if (p-ii) > 1:
            sys.stdout.write(str(ii + 2) + "...")
            sys.stdout.flush()

driver.find_element_by_xpath("""//*[@id="main_content"]/div/div[4]/a""").click()

      driver.quit()
      # write the word count file, if there's anything to be written.
      if i > 0:
         dwc.write_count_file()
   except Exception as e:
      # error handling...
      sys.stdout.write("error!")
      sys.stdout.flush()
      print("\n\nOh dear! There's a problem parsing the tweets...")
      print(str(e))
      tb = sys.exc_info()[2]
      efn = "Demonerror_" + str(datetime.datetime.now().strftime("%Y-%m-
%d_%H.%M.%S")) + ".txt"
      ef = open(efn, "wb")
      em = "DEMONANNA v. 0.6.3b ERROR DUMP\n\n"
      em += "Time: " + str(datetime.datetime.now()) + "\n"
      em += "Error at line: " + str(tb.tb_lineno) + "\n"
      em += "Message: " + str(e) + "\n"
      em += "Query: " + str(q) + "\n"
      em += "Pages: " + str(p) + "\n"
      try:
         em += "Current page: " + str(ii+1) + "\n"
      except:
         pass
      try:
         em += "Current URL: " + str(driver.current_url) + "\n"
      except:
         pass
      try:
         em += "---HTML Output---\n\n"
         em += str(driver.page_source)
      except:
         pass
      ef.write(bytes(em, "UTF-8"))
      ef.close()
      print("Error dump wrote to: " + str(efn))
      try:
         driver.quit()
      except:
         pass

   return

# This function is for getting tweets by using Twitter's API
# It uses to TwitterSearch Python module, available on GitHub and PIP
def useapi(ln, q):
   global i
   global tf
   qs = urllib.parse.urlencode({"q": q})
   try:
      # Create the TwitterSearch object and set the query parameters
      tso = TwitterSearchOrder()
      tso.set_search_url(qs)
      tso.set_language(ln)
      tso.set_include_entities(False)
```

```
      ts = TwitterSearch(
         consumer_key = 'xxxxxxxxxx',  # API keys have been masked for privacy.
         consumer_secret = 'xxxxxxxxxx',
         access_token = 'xxxxxxxxxx',
         access_token_secret = 'xxxxxxxxxx'
      )

      dwc = DemonWordCount(q)
      # Format the tweets associated metadata to CSV format, then write to the CSV
and word count files.
      for tweet in ts.search_tweets_iterable(tso):
         i += 1
         if str(tweet['place']) == "None":
            place = ""
         else:
            place = formatcsv(tweet['place']['full_name'] + ", " +
tweet['place']['country'])
         txt = '"@' + formatcsv(tweet['user']['screen_name']) + '",'
         txt += '"' + formatcsv(tweet['text']) + '",'
         txt += '"' + formatapitime(tweet['created_at']) + '",'
         tb = TextBlob(tweet['text'])
         txt += '"' + str(get_sent(tb.sentiment.polarity)) + '",'
         txt += '"' + str(get_sent(tb.sentiment.subjectivity)) + '",'
         txt += str(tweet['retweet_count']) + ','
         txt += str(tweet['favorite_count']) + ','
         txt += '"' + formatcsv(tweet['source']) + '",'
         txt += '"' + place + '"\n'
         tf.write(bytes(txt, "UTF-8"))
         dwc.add_to_counter(tweet['text'])

      if i > 0:
         dwc.write_count_file()

   except TwitterSearchException as e:
      print("\nOh dear! There's a problem...\n")
      print(str(e))
      tb = sys.exc_info()[2]
      print("Error at line: " + str(tb.tb_lineno))

   return

# Beginning of the program is here...
# Get the desired file name, search query, determine whether or not to do a site or
api search,
# and other necessary parameters to mine the tweets, and preform error checking.
# Finally, start the search.
print('''
\t\t*********************************
\t\t*  THE DEMON ANNA TWITTER TOOL  *
\t\t*            v.0.6.3b           *
\t\t*       By Jake & Anna          *
\t\t*********************************
''')
print("Search operator guide available at
https://dev.twitter.com/rest/public/search")
while True:
   i=0
   while True:
      qry = input("Enter search query: ")
      if str(qry).strip():
         break
      else:
         print("Umm. Maybe enter something to search...?")
   while True:
      fname = input("Enter file name (.csv): ")
      if str(fname).split():
         fname += ".csv"
         try:
```

```
            if os.path.isfile(fname):
               print("\n\t!!!WARNING!!!\nThe file " + fname + " already exists!")
               rw = input("1 = choose new file name, 2 = overwrite file, 3 = add
new data to file: ")
               if rw == "2":
                  tf = open(fname, "wb")
                  tf.write(bytes("Handle,Tweet,UTC-
Timestamp,Sentiment,Subjectivity,RTs,Likes,Source,Place\n", "UTF-8"))
                  break
               elif rw == "3":
                  tf = open(fname, "ab")
                  break
            else:
               tf = open(fname, "wb")
               tf.write(bytes("Handle,Tweet,UTC-
Timestamp,Sentiment,Subjectivity,RTs,Likes,Source,Place\n", "UTF-8"))
               break
         except PermissionError:
            print('Error! You don\'t have permission to write to "' + fname + '".
Try again.')
      else:
         print("Umm. Maybe enter a file name...?")

   while True:
      c = input("Choose search source (1 = api, 2 = website): ")
      if c == "1":
         print("Full list available at
https://dev.twitter.com/web/overview/languages")
         print("Examples - th = Thai, en = English, nl = Dutch, de = German")
         while True:
            lang = input("Enter language code: ")
            if str(lang).split() is False:
               print("You didn't enter a language code.")
            else:
               break
         tmr = timer()
         useapi(lang, qry)
         tmr = timer() - tmr
         break
      elif c=="2":
         while True:
            print('''
ATTENTION! By pressing \"ENTER\", you agree to NOT even think of touching the
opening Google Chrome window,
regardless of how tempting it may be. Doing so may create a supermassive black hole
that
destroys the universe, and/or you may not get your tweets.\n''')
            p = input("Enter max number of search result pages: ")
            if str(p).strip():
               if p.isdigit():
                  p = int(p)
                  if p>0:
                     break
                  else:
                     print("Umm. Maybe you'd like to search for 1 or more pages?")
               else:
                  print("Umm. Last time I checked, " + p + " isn't a positive
integer. Try again.")
            else:
               print("Umm. Maybe you'd like to search for 1 or more pages?")
         tmr = timer()
         usessite(p, qry)
         tmr = timer() - tmr
         break
      else:
         if str(c).split():
            print("Last time I checked, " + c + " isn't 1 or 2. Try again.")
         else:
```

```
            print("Nothing isn't a valid search source. Try again.")
   tf.close()
   if i==0:
      print('\nDONE! But no tweets found for "' + qry + '" :(\n')
   else:
      print("\nDONE! Wrote " + str(i) + " tweet(s) to \"" + fname + "\" in " +
str(round(tmr,2)) + " seconds.\n")
   if input("Run a new search? (y/n): ").lower() != "y":
      break
try:
   exit()
except:
   pass
```