OPINION MINING USING HYBRID LEXICON-BASED APPROACH AND MACHINE LEARNING TECHNIQUES FOR POLITICAL SECURITY THREAT

NUR ATIQAH BINTI MALIZAN

Thesis submitted to the Centre for Graduate Studies, Universiti Pertahanan Nasional Malaysia, in fulfilment of the requirements for the Degree of Master of Science (Computer Science)

2023

ABSTRACT

One of the elements of the national security domain is political security that associated with risks and threats such as riots and civil war that could potentially jeopardize the social stability of a nation. The complexity of political security surged with the advancement of personal devices and internet connectivity contribute to massive online information sharing. The information may consist of sentiments that may pose risk to political security. Thus, maintaining political security currently require an enhanced mechanism explicitly designed to monitor sentiments or opinions, and detecting any extreme emotions triggered online that can lead to negative effect as well as threats. To fill that gap, this thesis presents a theoretical framework in predicting political security threats in cyberspace by combining a lexicon-based approach and machine learning. For machine learning threat classifier, three (3) machine learning techniques which are Naïve Bayes, Support Vector Machine, and Decision Tree were tested. The result was evaluated using performance assessment. The comparison of accuracy, precision, and recall was presented to validate the hybrid approach: lexicon-based approach and machine learning to be applied in the proposed framework. The evaluation result demonstrated the accuracy value of classification and prediction using the combination of the lexicon-based approach and Decision Tree was higher than other approaches and provided the optimum result for this framework. This framework offers valuable insights in opinion mining realm to predict threats focusing on political security elements and demonstrate the definite relation of emotion, sentiment with political security threats.

ABSTRAK

Salah satu elemen dalam keselamatan negara ialah keselamatan politik. Terdapat risiko dan ancaman yang berkaitan dengan keselamatan politik seperti rusuhan dan perang saudara yang berpotensi menjejaskan kestabilan sosial sesebuah negara. Risiko keselamatan politik meningkat dengan kemajuan peranti peribadi dan sambungan internet yang menyumbang kepada perkongsian maklumat dalam talian secara meluas. Maklumat tersebut mungkin terdiri daripada sentimen yang mungkin menimbulkan ancaman kepada keselamatan politik. Oleh itu, mengekalkan keselamatan politik pada masa kini memerlukan mekanisme yang direka secara ekskulsif untuk memantau sentimen atau pendapat, dan mengesan sebarang emosi berlebihan yang dicetuskan dalam talian yang boleh membawa kepada ancaman yang tidak diingini. Untuk menangani jurang ini tesis ini mencadangan untuk mengembangkan rangka kerja yang dapat meramalkan ancaman keselamatan politik dalam talian menggunakan teknik hibrid, yaitu pendekatan berasaskan leksikon dan pembelajaran mesin. Ini adalah pendekatan yang inovatif dan berguna dalam bidang perlombongan pendapat dan keselamatan politik. Dalam tesis ini, tiga teknik pembelajaran mesin iaitu Naïve Bayes, Mesin Vektor Sokongan dan Pohon Keputusan telah diuji dan dibandingkan untuk memilih teknik yang paling sesuai untuk pengelasan ancaman keselamatan politik. Hasil pengujian menunjukkan bahawa pendekatan hibrid yang menggunakan pendekatan berasaskan leksikon dan Pohon Keputusan memberikan hasil yang lebih baik daripada pendekatan lain dalam mengklasifikasikan dan meramalkan ancaman keselamatan politik.Rangka kerja yang dicadangkan dapat membantu pihak berkuasa untuk memantau dan mengesan ancaman keselamatan politik dalam talian, serta membantu mereka dalam mengambil tindakan yang sesuai untuk mencegah ancaman tersebut. Ini adalah langkah penting dalam menjaga keselamatan politik dan kestabilan sosial sesebuah negara.Secara keseluruhan, tesis ini memberikan sumbangan yang signifikan dalam bidang keselamatan politik dan perlombongan pendapat. Rangka kerja yang dicadangkan boleh membantu meningkatkan keselamatan politik dan mencegah ancaman keselamatan yang tidak diingini.

ACKNOWLEDGEMENTS

I would like to express my most profound appreciation to my main supervisor, Prof Madya Ts. Dr. Afiza, for her genuine passion, guidance, wisdom, and most importantly, her utmost patience in guiding me throughout my study journey. Her dedication and kindness always keeping me on track and motivated me in finishing this study. Thank you to my co-supervisors Ts. Dr. Asiakin for sharing their best knowledge and valuable advice to improve my study. Special thanks to Dr. Muslihah and Mister Khairul Khalil for the precious insight on opinion mining/sentiment analysis mechanism. Not forgetting to the examiners for their comments and suggestion which means a lot to me to improve this thesis.

Several organizations and individuals have offered support during my study. I sincerely wish to extend my thanks to the CGS and FSTP UPNM for assistance during my study process. I also have had the privilege of supportive and understanding colleagues and friends during my study. I wish to thank you all, especially Miss Nuraini and Maizeerah, for showing me the path to my main supervisor.

To my family for their sincerest prayer, love, and greatest inspiration, for giving the best motivation for my Master study. My parents' dedications and sacrifices gave me the strength and inspiration in whatever I work for. Thank you.

I look forward to more adventures and exploration of knowledge that benefit us and people around us, thus brings us closer to Allah. Insya-Allah.

APPROVAL

The Examination Committee has met on **8th February 2023** to conduct the final examination of **Nur Atiqah binti Malizan** on his degree thesis entitled '**Opinion Mining Using Hybrid Lexicon-Based Approach and Machine Learning Techniques for Political Security Threat'**.

The committee recommends that the student be awarded the Degree of Master of Science (Computer Science)

Members of the Examination Committee were as follows.

Prof. Madya Dr. Syahaneim binti Marzukhi
Faculty of Defence Science and Technology
Universiti Pertahanan Nasional Malaysia
(Chairman)

Prof Madya Dr. Mohd Rizal bin Mohd Isa
Faculty of Defence Science and Technology
Universiti Pertahanan Nasional Malaysia
(Internal Examiner)

Prof. Madya Dr. Izwan Nizal bin Mohd Shaharanee
Department of Decision Science, School of Quantitative Sciences
Universiti Utara Malaysia
(External Examiner)

APPROVAL

This thesis was submitted to the Senate of Universiti Pertahanan Nasional Malaysia and has been accepted as fulfilment of the requirements for the degree of **Master of Science (Computer Science)**. The members of the Supervisory Committee were as follows.

Prof Madya Ts. Dr. Noor Afiza binti Mat Razali
Faculty of Defence Science and Technology
Universiti Pertahanan Nasional Malaysia
(Main Supervisor)

Ts. Dr. Nor Asiakin binti Hasbullah
Faculty of Defence Science and Technology
Universiti Pertahanan Nasional Malaysia
(Co-Supervisor)

Dr. Muslihah binti Wook
Faculty of Defence Science and Technology
Universiti Pertahanan Nasional Malaysia
(Co-Supervisor)

UNIVERSITI PERTAHANAN NASIONAL MALAYSIA

DECLARATION OF THESIS

Student's full name : Nur Atiqah Binti Malizan

Date of Birth : 5 October 1996

Title : Opinion Mining Using Hybrid Lexicon-Based Approach and Machine Learning Techniques for Political Security Threat

Academic Session : 2021/2022

I hereby declare that the work in this thesis is my own except for quotations and summaries which have been duly acknowledged.

I further declare that this thesis is classified as:

☐ **CONFIDENTIAL** (Contains confidential information under the official Secret Act 1972) *

☐ **RESTRICTED** (Contains restricted information as specified by the organisation where research was done)*

■ **OPEN ACCESS** I agree that my thesis to be published as online open access (full text)

I acknowledge that Universiti Pertahanan Nasional Malaysia reserves the right as follows.

1. The thesis is the property of Universiti Pertahanan Nasional Malaysia.
2. The library of Universiti Pertahanan Nasional Malaysia has the right to make copies for the purpose of research only.
3. The library has the right to make copies of the thesis for academic exchange.

Signature

961005-146766

IC/Passport No.

Date:

**Signature of Supervisor/Dean of CGS/ Chief Librarian

**Name of Supervisor/Dean of CGS/ Chief Librarian

Date:

Note: *If the thesis is CONFIDENTIAL OR RESTRICTED, please attach the letter from the organisation stating the period and reasons for confidentiality and restriction.

** Witness

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

FIGURE NO. **TITLE** **PAGE**

LIST OF ABBREVIATIONS

AdaBoost	Adaptive Boosting
ANN	Artificial Neural Network
BERT	Bidirectional Encoder Representations from Transformers
BFTree	Best-First Decision Tree
BPNN	Back-Propagation Neural Networks
CNN	Convolutional Neural Network
COVID-19	Coronavirus disease 2019
CRF	Conditional Random Fields
DBN	Deep Belief Network
DNN	Deep Neural Network
DT	Decision Tree
EO	Expected Outcome
FMT	Free Malaysia Today
FN	False Negative
FP	False Positive
GRU	Gated Recurrent Unit
KNN	K-Nearest Neighbour
LDA	Latent Dirichlet Allocation
LR	Logistic Regression
LSTM	Long short-term memory
ME	Maximum Entropy
ML	Machine Learning
ML-KNN	Multilabel K-Nearest Neighbours
MPQA	Multi-Perspective Question Answering
NB	Naïve Bayes
NLP	Natural Language Processing
NNs	Neural Network
NSC	National Security Council
NST	New Straits Time
OneR	One Rule
RF	Random Forest
RNN	Recurrent Neural Network
RO	Research Objective
RQ	Research Questions
SGD	Stochastic Gradient Descent
SVM	Support Vector Machines
SWN	SentiWordNet
TF-IDF	Term Frequency–Inverse Document Frequency
TN	True Negative
TP	True Positive
URL	Uniform Resource Locator
Vader	Valence Aware Dictionary and Sentiment Reasoner.

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

The Worldwide Threat Assessment of the US Intelligence Community (2016) mentioned that 'Cyber and Technology' is the top priority ahead of other prominent threats, including terrorism and counterintelligence (Clapper, 2015). Currently, the cyberspace has become a new outstanding issue in national security. A nation's security requires the implementation of more sophisticated measures, and all traditional measures for protecting a nation need to be upgraded by considering massive information sharing due to the establishment of big data analytics phenomena such as in the oil and gas, utilities, financial services, manufacturing, transportation, and government. In the oil and gas industry, big data analytics can help companies identify potential drilling locations by analysing geological and seismic data, as well as historical drilling data. Financial services firms use big data to manage risk by analysing vast amounts of financial data in real-time. They can also use it to analyse market trends and make investment decisions. Manufacturers and transportation companies use big data to manage their supply chains and optimize delivery routes. By analysing data from various sources, they can identify bottlenecks and inefficiencies, and make changes to improve efficiency and reduce costs. Government agencies use big data for a variety of purposes, including emergency response, crime

prevention, and smart city initiatives. For example, police departments can use big data to analyse crime patterns and deploy resources more effectively, while city planners can use it to optimize traffic flow and reduce congestion. The platform created in cyberspace promotes various data sharing contains various emotional expression of the public. The emotion can pose threats that contribute to the development of risk to national security.

People with specific objectives can just create fake news and rumours online to trigger negative emotions that can be escalated to disruptive behaviours in the society, resulting in chaos that may jeopardise national security. This scenario was observed in the Arab Spring. Detection of disruptive emotions/ opinions and potential threats to national security in real time is essential to assist authorities manage the situation early.

Opinions or emotional sentences have the potential to provoking negative emotions or threatening national security such as rage and fear. However, there is a lack of comprehensive and reliable methods to assess and measure emotions in terms of national security. Emotions play a critical role in the decision-making process of individuals, including those in the national security domain. Emotions such as fear, anger, and anxiety can influence the behaviour of individuals and, in turn, affect the outcome of critical situations in national security. Therefore, limited research is being produce to relate this emotion with national security threats and how the measurement can be conducted. A mechanism to mine the opinions extracted from text using any opinion mining technique that complements other methods to measure emotion in the national security domain is significantly needed to extract both opinions and emotions.

Thus, based on related literature, this study discovered that there is no opinion mining research being conducted linked to the political security element within the national security domain.

The current methodology in measuring opinion is not capable of including the weight of emotions as the instrument for national security threat classification. The opinion mining method can determine the words' polarity (Barkavi & Vimali, 2016). Yet, the research on predicting threats is crucial for ensuring the nation's safety and the employment of lexicon-based approach and machine learning which is opinion mining method were less explored by the researcher (Razali et al., 2021).

In this study, fundamental research initiated by establishing the prediction of threat framework for national security focusing on the political security element is identified as the research scope. The scope was determined based on the thorough analysis on the national security domain based on literature reviews that suggest the formation of an analytic model with a strong fundamental background equipped with optimum capability in assessing human emotion, and the relationship with national security threat. Based on the established theoretical framework, the dataset was established by compiling text published in an online platform as a sample frame to extract text data.

The new framework shall create new research fields that incorporate multi-research domains which are opinion mining and national security. A new framework that uses a hybrid technique in measuring emotion and prediction of national security threats in cyberspace is proposed.

1.1 Chapter Overview

This chapter explains the background of this study, focusing on political security element, opinion mining, the problem statement, research objectives, research questions, scope and limitation, definition of terms, and thesis structure.

1.2 Background of Study

1.2.1 National Security

National security is defined as a concept that encompasses the protection and preservation of the nation and citizens' safety (Cosby, 2009). The definition of national security varies due to the complexity of the national security realm itself that is complicated but focuses on a very clear target in protecting the nation and its people in the society against determined threats that target economic coercion, armed invasion, and political repression (Stoykov Stoyko & G.S.Rakovski, Sofia, 2011). The highest priority of national security is safeguarding the state to defend the nation and its people from dangers and threats through the protection of state secrets and maintaining military forces.

National security acts as a vital feature of protection control for any country (Balzacq, 2015). The framework of the economic activity, economic growth and development, variation of weather and climate, welfare of citizens, political stability, and preservation of natural resources were included as dimensions of national security (Scheffran, 2011). It was also discovered that national security had eight elements which are military security, political security, human security, homeland security,

cybersecurity, environmental security, energy and natural resources security, and economic security (Balzacq 2015). Some studies have identified non-military and military factors as the two fundamental elements of national security. Military security was defined as the capability of a country to shield itself. Non-military elements include economic security, food security, human security, political security, health security, border security, energy and natural resources security, and cybersecurity. (Bahadur & Thakuri, 2018; Kshetri, 2016).

This study has highlighted politics as one of the four main aspects of national security which is also explained as one of the important concepts in national security (Vxyrnen, 2014). Figure 1.1 illustrates the relationship between national security domain and its elements.

Figure 1.1 Relationship between national security domain and its elements (Balzacq, 2015)

1.2.1.1 Political Security

Political security emphasises the steadiness of nations' governments and institutions. The political security stated by Robert Mandel was deliberated extensively, focusing

on preserving the stability of the organisational state's systems holdup by the underlying principles to guarantee fairness to the right people (Buzan, 1991). In the national security dimension, the role of political security is significant since these elements safeguard the state against any kinds of political repression that would jeopardise national security. Political oppression may result in an unwanted situation that include riots or civil war, which may disrupt the harmony of the populace (Ronken et al., 2020) and it is important in ensuring political stability and security.

Political stability has become a significant concern for nation political security. This is due to threats arising from a political perspective can bring damage to the nation political security and undermine the stability of the nation, including violations of the law. Research in India has highlighted that political security must be considered as a term of national security to safeguard and accomplish national security objectives and goals (Enclave et al., 2019).

In addition, the political matter can be spread out to cyberspace for the reason of the world-wise usage of the Internet that can give rise to threats to the nation's political environment (Sandoval-almazan & Gil-garcia, 2014). Threats that can disrupt the political security stability of a country are technology, political upheaval, and political violence, of which may threaten peace. Moreover, these threats are easily triggered in cyberspace because of emotions or opinions that are excessively expressed by cyber citizens in that area especially through online platforms. Thus, a certain approach is needed to analyse and detect people's opinions that can threaten the nation’s political security.

1.2.2 Opinion Mining

Opinion mining (sentiment analysis) can be a valuable tool for understanding public opinion on a specific target, such as a product, service, or political candidate. By analysing textual data documents, such as social media posts, reviews, and news articles, sentiment analysis can provide insights into the attitudes, opinions, and emotions of individuals towards a particular topic. Opinion mining is also known as a subfield in Natural Language Processing (NLP) applications and it is also knowns as sentiment analysis which can be defined as doing the task of mining opinions expressed in text and analysing the entailed sentiments and emotions.

Web 2.0 sensations boosted this approach and extensive research were performed within the opinion mining context. The newly created text forms in "Social Networks" include online news, blogs, e-commerce reviews, and e-forums contain texts that include people's perspectives and foster the expansion of research in this field. Moreover, various topics from politics, economics, environment, and technology are shared around the globe within these network space. This enormous volume of subjective data can be compared to determine how individuals felt about the product, brand, popularity tracking, consumer reviews, and social media analysis, as well as detecting general moods and "hot" news (Balahur et al., 2009). There are two sequential tasks in opinion mining namely identifying opinions, their polarity, and the strength of those viewpoints in text portions (Seerat, 2012). Opinion mining’s important feature is evaluating a written piece to figure out what it means.

Typically, opinion mining focuses on data extraction and the collection of sample data and author-expressed opinions. This method assigns positive or negative score values to the sentiment or opinion s of a text, often known as polarity. Always referred to as positive, neutral, or negative, the polarity score is indicative of the overall emotion. The marking is then utilised in summarising the content of texts embedded with emotion. Studies show that there are numerous criteria that can be used to determine viewpoints and polarity. (Barnaghi et al., 2016). In addition, researchers in text analysis commonly adopt an opinion mining approach in their work because this approach is rapidly evolving. New areas of study are constantly being developed, with more advanced and up-to-date methods as well as calculations are being introduced. This swift improvement supports the programmed information process (Feldman, 2013).

1.3 Problem Statements

Evolution of internet connectivity supported by advancements of devices offer a valuable medium and platform for everyone to express emotions, feelings, ideas, and opinions. Social media also plays a vital role in boosting interactions between people that live in different societies and cultures. It is expected that social media users will rise to 5 billion in 2022 (Hannahcurrey, 2022). This scenario will enable every single individual with devices and connectivity to share opinions, idea, and knowledge in the online platforms. The information available on social media platforms will provide a large pool of data that can be utilised for decision-making by the government or corporations. The information can also be used for provocation and executing cyberwarfare. The information can be on various forms including textual. Textual data can be manipulated in order to mine emotions underlying the text. When publishing

in digital media platforms, people are unaware that text with excessive emotions can threaten society's peace. Unwanted consequences like riots and civil war could result from overemotional citizens. It is important to address any threats to society's peace and take measures to overcome them. By doing that, governments can help to ensure the stability of their societies and, by extension, their national security. Mining opinions and emotions in relation to national security is an important research topic that can provide valuable insights into public perception and attitudes towards national security issues.

For research related to opinion mining, most research applied to the domain includes business, entertainment, medicine, and the government. But the researcher has not fully explored the national security domain. Moreover, political issues such as political elections can indeed be a tense time for a nation, as they often involve significant disagreements and competing interests among political parties and their supporters. In some cases, this can lead to negative consequences such as violence, instability, and even civil war. Also, fake information or "fake news" can have a significant impact on the outcome of a general election, as it can misinform and mislead voters. With the rise of social media, it has become easier for false information to spread quickly and widely, making it more difficult for voters to discern what is true and what is not. Thus, it creates a research gap that addressed in this research. The current opinion mining technique mostly focuses on finding opinions based on the text by showing positive and negative opinion polarity that does not include emotions and the relationship with national security threats and how the measurement can be conducted.

Opinions or emotional text for provoking negative emotions such as rage and fear is a threat to national security. However, the framework that includes assessment and analysis methods for emotions and its measurement in the national security domain is still lacking. there has been less attention given to measuring and analysing emotions related to national security threats using opinion mining techniques. While opinion mining has been widely used in other areas, such as marketing and customer service, its application to national security has been limited. A mechanism that mines opinions extracted from text using various opinion mining techniques complemented with other methods which includes emotion measurement in the national security domain is crucially needed. Thus, this research will propose this framework.

1.4 Research Objective

This research aims to explore the potential and propose a threat prediction framework based on an opinion mining mechanism involving the national security domain. Political security is an element of the national security domain and was one of the elements chosen for this study's scope. Thus, this study is focused in creating a framework focusing on political security. The implementation of this framework is intended to provide insightful information to ongoing research on opinion mining in forecasting risks with a political security element and demonstrate the close connection between emotions and political security threats. Thus, the acceptance of such an approach needs to be evaluated prior to further development and investment. In order to achieve this aim, this research must achieve three outlined main objectives as follows:

i. To explore the adoption of Lexicon-based Approach and Machine Learning technique in mining people's opinions and emotional expressions, as well as the relationship with political security element (in national security domain).

ii. To propose a threat prediction framework for political security element (in national security domain) using hybrid Lexicon-based Approach and Machine Learning technique to determine people's opinions and emotions in text.

iii. To evaluate the proposed framework by experimenting using hybrid Lexicon-based Approach and Machine Learning technique in mining and predicting people's opinions/emotions and threats according to the political security element (in national security domain).

1.5 Research Questions

Three research questions (RQ) must be addressed to meet this study's outlined research objectives (RO). The research questions are:

i. What approach and technique are suitable to mine people's opinions and emotions according to the political security element (in national security domain) in cyberspace?

ii. What are the components required for framework development to mine people's sentiments, emotions and prediction of threats according to the political security element (in national security domain)?

iii. How can the validation be done for the proposed framework in mining the people's sentiments, emotions and prediction of threats according to the political security element (in national security domain)?

Alignment between RQ, RO and the expected outcome (EO) of this research is shown in Table 1.1.

Table 1.1 Alignment between RQ, RO, and EO of this research

Research Questions	Research Objectives	Expected Outcomes
RQ1: What approach and techniques are suitable to mine people's opinions and emotions according to the political security element (in national security domain) in cyberspace?	**RO1:** To explore the adoption of Lexicon-based Approach and Machine Learning technique in mining people's opinion and emotional expressions, as well as the relationship with political security element (in national security domain)	Determination of the techniques that can be utilised to mine people's opinions and emotions according to the political security element in cyberspace
RQ2: What are the components required for framework development to mine people's sentiments, emotions and prediction of threats according to political security element?	**RO2:** To propose a threat prediction framework for political security element (in national security domain) using hybrid Lexicon-based Approach and Machine Learning technique to determine people's opinions and emotions in text	Establishment of threat prediction framework for mining people's sentiment, emotions, and prediction of the threat according to political security in cyberspace
RQ3: How can the validation be done for the proposed framework in mining the people's sentiments, emotions and prediction of threats according to the political security element (in national security domain)?	**RO3:** To evaluate the proposed framework by experimenting using hybrid Lexicon-based Approach and Machine Learning technique in mining and predicting people's opinions/emotions and threats according to the political security element (in national security domain)	Evaluated proposed framework by employing the hybrid approach in mining the sentiment and emotions also threat prediction according to political security aspect in cyberspace