INTELLIGENT SOCIAL MEDIA TEXT MINING FOR POLITICAL INFLUENCE ANALYSIS: A CASE STUDY ON MAKASSAR MAYOR ELECTION

JUFRI

ASIA e UNIVERSITY

2024

INTELLIGENT SOCIAL MEDIA TEXT MINING FOR POLITICAL INFLUENCE ANALYSIS: A CASE STUDY ON MAKASSAR MAYOR ELECTION

JUFRI

A Thesis Submitted to Asia e University in
Fulfilment of the Requirements for the
Degree of Doctor of Philosophy

March 2024

ABSTRACT

Social media has become a common means of communication in society and plays a big part in many community activities in areas such as politics, economy, education, and information sharing. Important platforms such as Instagram, Twitter, Facebook and others can drive increased participation in elections. In this context, this research uses the Twitter application to influence political conversations through social media platforms. In the context of political communication during the Makassar mayor election, this study examines the application of a hybrid text mining approach that combines Support Vector Machine (SVM), Naïve Bayes, and K-means clustering for sentiment analysis. Traditional methods of sentiment analysis often fail to capture the nuanced sentiments of the electorate because of the complex nature of political discourse. This study aims to address these limitations by leveraging the use of SVM, Naïve Bayes, and K-Means. This research conducts data preparation involves cleansing and organizing textual data, segmenting it using K-means clustering, classifying it into sentiment classes using Naïve Bayes classifier, and enhancing classifications with the SVM. The results demonstrate that the hybrid model has superior performance compared to traditional methods, attaining an accuracy rate of 85.43% in contrast to the 64.96% accuracy rate achieved by traditional approaches. The hybrid approach demonstrates superior performance in sentiment accuracy and thematic analysis compared to traditional methods, highlighting its potential to extract meaningful insights from complex textual data. The findings reveal significant sentiment trends and discourse themes that influenced public opinion during the election. Furthermore, the research showcases the adaptability of the hybrid approach to diverse data sources and its applicability to other domains requiring detailed sentiment and thematic analysis. These findings constitute a valuable contribution to the fields of political science and computational linguistics by presenting a novel framework for sentiment analysis. This framework improves the analytical abilities of political analysts, campaign strategists, and policymakers.

Keywords: Social media, twitter, political communication, K-means, Naive Bayes and SVM

APPROVAL

This is to certify that this thesis conforms to acceptable standards of scholarly presentation and is fully adequate, in quality and scope, for the fulfilment of the requirements for the degree of Doctor of Philosophy.

The student has been supervised by: **Prof Ts Dr Aedah binti Abd Rahman, Professor, SST & Head, ADT, ICTS & ACE and Dr Suarga, STMIK Dipanegara Makassar**

The thesis has been examined and endorsed by:

Assoc Prof Dr Nurazzah Abdul Rahman,
Universiti Teknologi Mara (UiTM)
Examiner 1

Prof Ts Dr Mokhairi Makhtar,
Universiti Sultan Zainal Abidin (UniSZA)
Examiner 2

This thesis was submitted to Asia e University and is accepted as fulfilment of the requirements for the degree of Doctor of Philosophy.

Professor Dr. Siow Heng Loke
Asia e University
Chairman, Examination Committee
(7 March 2024)

DECLARATION

I hereby declare that the thesis submitted in fulfilment of the PhD degree is my own work and that all contributions from any other persons or sources are properly and duly cited. I further declare that the material has not been submitted either in whole or in part, for a degree at this or any other university. In making this declaration, I understand and acknowledge any breaches in this declaration constitute academic misconduct, which may result in my expulsion from the programme and/or exclusion from the award of the degree.

Name: Jufri

Signature of Candidate: **Date**: 7 March 2024

Copyright by Asia e University

ACKNOWLEDGEMENTS

I would like to express my deep thanks to Prof. Ts. Dr. Aedah Binti Abd. Rahman and Drs. Suarga, M. Sc., M. Math., PhD., as Principal Supervisor and Co-Supervisor, for guidance, encouragement, and criticism that are very useful for my research. Special thanks to Rector of Dipa Makassar University Dr. Y. Johny W. Soetikno, S.E., MM. and Mr. Huisdjono and Dr. Paulus Djohan Lolo, S.E., MM., chairman of Dipanegara foundation for their support and kindness in furthering my Doctoral studies at AeU. Finally, beloved to my wife Musdalifa Thamrin dan family, thank you for your support and understanding.

TABLE OF CONTENTS

ABSTRACT ii
APPROVAL iii
DECLARATION iv
ACKNOWLEDGEMENTS vi
TABLE OF CONTENTS vii
LIST OF TABLES x
LIST OF FIGURES xi
LIST OF ABBREVIATION xiii

CHAPTER 1 INTRODUCTION 1

1.0 Background of the Study 1
1.1 Problem Statement 8
1.2 Research Objectives 11
1.3 Research Questions 13
1.4 Scope and Limitation of the Research 17
1.5 Justification and Significance of the Research 17
1.6 Chapter Summary 18

CHAPTER 2 REVIEW OF LITERATURE 20

2.0 Introduction 20
2.1 Literature Review 21
2.1.1 Political Communication 21
2.1.2 Social Media 25
2.1.3 Data Mining 35
2.1.4 Text Mining 39
2.2 Sentiment Analysis and Text Mining 43
2.3 Natural Language Processing and Machine Learning 44
2.4 Supervised Learning Method and Support Vector Machine 47
2.4.1 Support Vector Machine (SVM) 51
2.4.2 Naïve Bayes, K-means, and Other Methods 53
2.4.3 Naïve Bayes Classification 54
2.4.4 K-Means 57
2.4.5 Machine Learning Method 59
2.4.6 Twitter as Social Media Platform for Political Communication & Text Mining Analysis 61
2.4.7 Text Mining Techniques 63
2.4.8 Algorithm Term Frequency-Inverse Document Frequency (TF-IDF) 67
2.5 Recent Related Studies 69
2.6 Gap Analysis in this Research 78
2.7 Chapter Summary 79

CHAPTER 3 METHODOLOGY 82

3.0 Introduction 82
3.1 Research Framework 85

3.2 Propose Technique for Analyzing Political Communication on Social Media 89
3.3 Makassar City 91
3.3.1 Levels of Indonesian Government 95
3.3.2 Province in the Sulawesi Island 98
3.4 Data Processing Tools 99
3.5 Conceptual Research 100
3.6 Determining the Recent Problems, Issues, Status and Trends in Political Sentiment Analysis at Makassar City, Indonesia 108
3.6.1 Population and Sampling 108
3.6.2 Tracking Data Documents Collection Using Crawling Data Technique 109
3.6.3 Develop a New Improved Approach to Support Social Media Text Mining in Political Communication 111
3.6.4 Text Mining Using Hybrid Algorithm Naïve Bayes and K-means Technique 113
3.6.5 Validation of the New Approach to Support the Improved Text Mining Methods for Political Communication 116
3.6.6 Classification and Clustering Data 126
3.6.7 Validation Using Confusion Matrix 133
3.7 Chapter Summary 135

CHAPTER 4 RESULTS AND DISCUSSION 137

4.0 Introduction 137
4.1 Data Collection 137
4.1.1 Crawling Data 137
4.1.2 Data Pre-Processing 145
4.2 Data Processing 158
4.2.1 Tweet Data Sentiment Analysis Process Using Naïve Bayes 158
4.2.2 Tweet Data Clustering Process Using K-means 172
4.2.3 Classification Process with Support Vector Machine (SVM) 190
4.3 Validation 195
4.3.1 Design K-Fold Cross Validation Naïve Bayes 200
4.3.2 Design K-Fold Cross Validation SVM 201
4.4 Chapter Summary 202

CHAPTER 5 CONCLUSION, IMPLICATION AND RECOMMENDATIONS 206

5.0 Introduction 206
5.1 Overall Conclusion 206
5.2 Research Contribution 207
5.3 Objectives Fulfilled in this Research 210
5.4 Research Challenge 211
5.5 Research Limitation 214
5.6 Recommendation for Future Research 215
5.7 Chapter Summary 217
REFERENCES 219
APPENDICES 228

Appendix A 228
Tweet Data from Crawling Results 228
Pre-processing Data Result 240
Results of Data Classification with Naive Bayes 261
Results of Data Classification with Support Vector Machine 279
Data Clustering Results with K-means 294
Data Result of Confusion Matrix using Naïve Bayes and SVM 335
Appendix B: Research Gantt Planner 351
Appendix C: Letter of Permission to Conduct Research 354

LIST OF TABLES

Table		Page
1.1	Mapping relation among of PS, RQ, and RO	15
2.1	Reasearch gap summary	75
3.1	Sub-district city of Makassar	84
3.2	Village in the Sulawesi Island	98
3.3	Dataset	126
3.4	Dataset label	127
3.5	Data pre-processing	127
3.6	Sentiment analysis prediction results	129
3.7	Example of opinion sentiment data for the Makassar mayor election that has been changed in numerical form	133
3.8	Initial centroid data example	133
3.9	Results of calculation of the distance between data to the centroid	133
4.1	Example of crawling data on twitter	139
4.2	Example of preprocessing results	147
4.3	Examples of classification models data	159
4.4	Examples of clustering using K-means data	173
4.5	Example of data classification results with SVM	190
4.6	Example of source data confusion matrix using Naïve Bayes and SVM process	196

LIST OF FIGURES

Figure		Page
2.1	Types of hoax social media receives	31
2.2	Hoax news channel	32
2.3	The knowledge discovery process	38
2.4	Text mining process	40
2.5	Machine learning method	61
2.6	Text mining techniques	64
3.1	Research framework	87
3.2	Map of Indonesia	91
3.3	Map of Makassar city	91
3.4	Structure of Indonesian Government	95
3.5	Map of Sulawesi Island	98
3.6	Conceptual research	104
3.7	Text collection process	104
3.8	Collection using crawling data technique	110
3.9	Technique-improved approach to support social media text mining	112
3.10	Data pre-processing	117
3.11	Data conversion	119
3.12	Cleaning	121
3.13	Normalization	123
3.14	Tokenizing	124
3.15	Stop-words removal	125
4.1	Classification model process	167
4.2	TF-IDF result	168

4.3	Sentiment analysis model process	169
4.4	Data non label	170
4.5	Results of Naïve Bayes sentiment analysis	171
4.6	Nominal value	171
4.7	Sentiment analysis results graph	172
4.8	Process of model K-means	184
4.9	Cluster result	185
4.10	Centroid value final results	185
4.11	Clustering results K-means	186
4.12	Results of visualizer model clustering	186
4.13	Clustering model visualizer result	187
4.14	Transform cluster process	187
4.15	Wordcloud cluster 1	188
4.16	Wordcloud cluster 2	189
4.17	Wordcloud cluster 3	189
4.18	Wordcloud cluster 4	189
4.19	SVM process model	194
4.20	SVM validation process model	195
4.21	Data testing result	195
4.22	Cross validation Naïve Bayes stages	200
4.23	Confusion matrix Naïve Bayes	201
4.24	Cross validation SVM stage	201
4.25	Confusion matrix Naïve Bayes	202

LIST OF ABBREVIATION

AEU	Asia e University
JEAS	Joint Entity-Aspect-Sentiment
JSIS	Joint Sentiment-Issue-Stance
KPU	General Election Commission
ML	Machine Learning
NLP	Natural Language Processing
SVM	Support Vector Machine

CHAPTER 1

INTRODUCTION

1.0 Background of the Study

The term social media implies the use of web and mobile-based technologies to communicate and interact based on user-generated content and dialogue. Enabled by multiple platforms and accessible characteristics, social media is significantly changing the way individuals, organizations, and communities communicate (Wong et al., 2021). The advent of social media platforms has provided a platform for engaging in political discourse and communication. Despite the waning popularity of the direct dialogic communication model, there is an optimistic outlook for the establishment of social media platforms in a virtual format. Social media platforms provide users with practicality and facilitate a diverse range of communication opportunities (Meiserli, 2021).

Social media, which can be considered a novel type of media, operates on the basis of an interactive and personalized communication framework. The necessity of engaging with conventional media or doing in-person meetings to effectively communicate with the public has diminished. This advantage has numerous prospects for generating personalized media content and circumventing traditional media platforms. However, it also poses a difficulty as it necessitates the acquisition of specific technical and non-technical competencies. Government elites endeavor to delineate certain communicative techniques on social media platforms. The individuals in question exhibit a tendency to adhere to conventional media practices, a phenomenon commonly referred to as normalization. Simultaneously, they also display a desire to engage with altered and innovative approaches associated with new media practices. The process of professionalizing political communications and

consolidating government communications has resulted in press officers assuming a more significant role in facilitating interactions among politicians, journalists, and the general public, including on social media platforms (Johansson, 2019).

The space that exists between actual reality and derivative reality serves as a safeguard for social, cultural, and technological processes, hence giving rise to implications for the creation of a novel reality and its subsequent impact on the audience. Media messages are a product of social construction. The political ramifications of the media do not stem solely from the inherent nature of the press, but rather arise from the manipulative nature of the media (Meiserli, 2021).

The main aim of this research is that social media helps mediate political discourse between society, candidates and political parties. Social media offers a more direct and engaging political communication channel than traditional media (Gilardi et al., 2022). People can now more easily approach politicians and public officials to convey their thoughts, criticism and aspirations regarding political issues and agendas via social media (Gilardi et al., 2022). Although social media is used for political communication, society as a whole benefit from its use. Therefore, the use of social media for political communication will be examined critically in this research. Thus, this research will offer findings regarding the use of social media in political communication (Sakti & Nainggolan, 2023).

The concept of political influence analysis pertains to the methodical scrutiny and assessment of the methods by which different entities, such as people, interest groups, political parties, or the media, exert their impact on political procedures, policies, determinations, and public sentiment (Michalski et al., 2023). The examination of political influence holds significant importance in comprehending the operation of democracy, the efficacy of governance, and the dynamics of political

communication (Michalski et al., 2023). This analysis offers valuable perspectives that can assist many stakeholders, such as political parties, candidates, policymakers, activists, and the general public, in making informed decisions and enhancing their participation in the political sphere. With the advancement of technology and data analytic tools, the discipline is constantly evolving, offering more profound and intricate understandings of the intricate network of factors that affect the political environment (Myeong et al., 2021).

Within the given document, the term "intelligent" pertains to the utilization of sophisticated techniques in the field of text mining. This entails utilizing methodologies from natural language processing (NLP), machine learning, artificial intelligence (AI), and data mining to scrutinize and extract significant insights from textual data (Mah et al., 2022). The objective is to reveal pertinent information, extract important observations, and develop helpful understanding by effectively handling and evaluating the intricate characteristics of human language included in texts.

This form of intelligent text analysis is considered to be a component of a wider strategy that utilizes the most recent technology breakthroughs to improve the extraction, interpretation, and usage of valuable information from written data. These advanced technologies are especially valuable in fields such as political communication analysis, where the intricate and nuanced nature of language can make it difficult to identify underlying feelings and patterns without their assistance.

The empirical findings indicate that social media exerts a significant impact on the political engagement of voters, whereas its effect on political knowledge is comparatively limited. The use of social media enhances the element of political engagement, although it does not ensure that voters possess sufficient political knowledge. Social media, as a form of new media, exhibits a significant level of

engagement due to its ability to enable voters to track the activities and online news of political candidates or parties (Jandevi, 2019). The connectivity aspect of social media is significantly pronounced since it enables direct communication between the general public and political candidates via their respective social media profiles. Members of the general public have the opportunity to engage in direct interaction with candidates, offering comments, sharing messages, or expressing approval through the act of like content (Ritonga & Syahputra, 2019). This fosters a dynamic type of communication that allows for reciprocal exchange. Similarly, social media plays a significant role in facilitating mobilization efforts, which are typically undertaken by political candidates or parties. These mobilization activities include the utilization of campaign stickers, campaign attire, participation in gatherings, and other related actions (Rahman & Prihatini, 2019). Nevertheless, social media has proven to be an inadequate platform for campaign purposes due to its inability to deliver comprehensive information, particularly within the Indonesian setting where the general public's reading and listening skills remain relatively low (Maharani & Nurafifah, 2020).

Social media serves as a communication platform that politicians utilize to engage in campaign-related endeavors. The issue is strategically utilized to elicit compassion from the urban people (Subekti et al., 2023). These political groups primarily engage in the collection of opinions to explicitly assign blame to individuals who are not favored, while consistently examining communications within the context of competition, or at least that is the prevailing perception. Social media, as the primary platform utilized by political groups, does not operate independently in the pursuit of information dissemination for the collective advantage of the group. This is due to the presence of numerous social media users who support and sympathize with the group,

yet propagate misleading information that deviates from the official stance of the institution or political group (Grover et al., 2022).

Social media is not only a means of communication and expression, but also an opportunity for political actors during the election season. Social media can be used as a means of political participation. This participation is in the form of electoral activity that can be carried out by the community using social media. The candidates, through social media, try to attract and convince the public to vote for them on the voting date. Candidates can deliver their vision and mission and orations without having to go outside. On the other hand, prospective voters also only need a gadget to know many things about the election candidates. That type of campaign is certainly more efficient in terms of cost and time than conventional campaign (Maharani & Nurafifah, 2020). Several things make social media a campaign tool (Purbolaksono, 2018):

1. The contemporary electoral contest is significantly shaped by the impact of social media platforms in augmenting the level of rivalry among candidates and political parties. In certain instances, minor political parties with constrained resources may lack significant influence, particularly inside electoral processes. Nevertheless, the internet, being a cost-effective and readily available medium compared to traditional communication technologies, enables smaller entities to effectively compete against larger counterparts with greater resources. Social media platforms provide minor political parties with the opportunity to effectively engage with potential supporters, akin to their larger counterparts.
2. The contemporary electoral contest is significantly shaped by the impact of social media platforms in augmenting the level of rivalry among candidates

and political parties. In certain instances, minor political parties with constrained resources may lack significant influence, particularly inside electoral processes. Nevertheless, the internet, being a cost-effective and readily available medium compared to traditional communication technologies, enables smaller entities to effectively compete against larger counterparts with greater resources. Social media platforms provide minor political parties with the opportunity to effectively engage with potential supporters, akin to their larger counterpartsb (Andriyendi et al., 2023).

3. Social media as an institutional adaptation. The utilization of social media platforms as a means of institutional adaptation. Institutional adaptation refers to the transition from traditional offline modes of political engagement to digital platforms for political participation. Policy makers, including general election commissions, election supervisory bodies, political parties, and candidates, have the ability to utilize social media platforms for the purpose of distributing various policies, election rules, and other relevant information. Additionally, these entities can also engage in offline methods to distribute election-related material. Social media platforms offer a means by which policy makers, political parties, and candidates can enhance the efficacy of their communication efforts (Wardhani, 2018).

Social media platforms generate a substantial volume of digital data, comprising diverse information, including political communication derived from user-generated textual data on these platforms social media platforms generate a substantial volume of digital data, comprising diverse information, including political communication derived from user-generated textual data on this platform (Purbolaksono, 2018).

Intelligent text mining encompasses the utilization of sophisticated and intelligent methodologies in the examination of textual material with the aim of uncovering pertinent information, extracting valuable insights, and generating informative knowledge (Asritha et al., 2021). The field encompasses various academic domains, including natural language processing (NLP), machine learning, artificial intelligence (AI), and data mining, to effectively manage and analyses the intricate nature of human language inside textual data (Mah et al., 2022). The subject of intelligent text mining is expanding. It utilizes advanced technology to tackle these difficulties, to enhance the extraction, interpretation, and utilization of insights from textual data. The advancement of AI and machine learning technology will expand intelligent text mining capabilities and applications, increasing its potential across various sectors.

This research presents a new approach to text mining that integrates three potent analytical techniques K-means clustering, Naïve Bayes, and Support Vector Machine (SVM) (Chaudhry et al., 2023). Every technique contributes distinct advantages to the examination; for example, SVM resilient classification skills, Naïve Bayes' effective probabilistic modelling, and K-means aptitude for identifying hidden patterns in the data. Our method seeks to offer a thorough sentiment analysis of political communication during the Makassar mayor election a significant event that has drawn interest from both academics and practitioners by combining these approaches.

Due to its importance in the political environment and the abundance of rich textual material it has produced, the Makassar mayor election is a perfect case study for our research. In addition to providing a technical explanation of the hybrid technique, this research aims to show how it may be used in practice to analyse

sentiment in political communication (Li et al., 2021). Hope to clarify public perceptions of the candidates, the main topics influencing the election discourse, and the general dynamics of political communication during the election season with this analysis (Yang et al., 2023).

Contribution this research combines SVM, naïve bayes, and K-means clustering are combined in this study to provide a hybrid technique for sentiment analysis in election processes (W. Chen et al., 2020). This method offers a more nuanced knowledge of popular sentiment in political communication by enhancing topic analysis and sentiment categorization. The work has ramifications for political science, computational linguistics, and data science, among other disciplines. It advances the study of election communication, strengthens text mining techniques, and makes it possible to use hybrid analytical approaches in real-world problem-solving scenarios.

1.1 Problem Statement

The initial step is providing a definition of social media as a technologically-driven platform that facilitates interactive and user-generated content, hence revolutionizing communication among individuals, organizations, and communities (Stephens & Powers, 2020). The present study subsequently examines the influence of social media on political communication, highlighting its inherent benefit in facilitating a direct and interactive platform connecting society, politicians, and political parties. This direct channel facilitates a more individualized and prompt mode of political communication, enabling the public to readily engage with politicians and public officials to express their opinions, critiques, and ambitions concerning political matters.

The document elucidates the notion of political influence analysis, which scrutinizes the manner in which diverse entities exert influence on political processes,

policies, decisions, and public mood. The examination of this study holds significant importance in comprehending the functioning of democracy, the effectiveness of governance, and the dynamics of political communication.

The document presents empirical evidence that demonstrate the substantial impact of social media on voter political involvement, while acknowledging its modest influence on political understanding. The statement highlights the role of social media in enabling direct communication between the general public and political candidates, hence increasing involvement. However, it acknowledges that this connection may not inevitably lead to an enhanced comprehension of intricate political matters by the public.

Moreover, the study examines the deliberate utilization of social media by political figures, particularly during election periods, as a medium for campaign-related endeavors, mobilization initiatives, and as a method of political engagement by the public.

This research also examines the technological dimension of assessing political communication via social media platforms through the application of text mining methodologies (Antons et al., 2020). This research presents a new method that combines K-means clustering, Naïve Bayes, and Support Vector Machine (SVM) to analyses the sentiment of political communication. The method is especially applied to the Makassar mayor election case study. The primary objective of this hybrid methodology is to provide a comprehensive sentiment analysis, thereby offering valuable insights into the public's impressions of candidates, the key subjects that shape election conversations, and the broader dynamics of political communication during the election process. Based on the research background presented in this document, the problem statement for this research can be articulated as follows:

1. Social media has become a crucial forum for political debate in the political sphere, dramatically altering the dynamics of communication between politicians, political parties, and the general public. The amount of content created by users, how quickly false information may spread, and how difficult it is to gauge how social media affects public opinion and political participation (Taufiq et al., 2022).
2. Understanding the impact, attitude, and involvement of voters and candidates depends on one's capacity to evaluate political messages on social media (Alodat et al., 2023). Sarcasm, innuendo, and intricate thematic aspects are common in political speech, which makes it difficult for traditional text mining techniques to discern its subtleties. Furthermore, the normalization of social media use by society and political organizations has produced enormous volumes of data, calling for analytical methods that may extract relevant insights, effects, patterns, and feelings from the deluge of textual data.
3. This research explores the effectiveness of conventional and hybrid text mining techniques for analyzing political mood during the Makassar mayoral election. Conventional methods, such as keyword-based approaches and basic machine learning models, struggle to capture the intricacies of political discourse due to their inability to consider contextual subtleties and evolving language usage (Lakshmi & Murugesh, 2023). Hybrid methods, which combine SVM, Naïve Bayes, and K-means clustering, aim to overcome these constraints by leveraging the strengths of multiple algorithms. SVM robustness in dealing with high-dimensional data complements Naïve Bayes' probabilistic modeling, efficiently managing uncertainty and variance in sentiment data (Korovkinas