PREDICTING PRESIDENTIAL ELECTIONS IN INDONESIA USING MACHINE LEARNING BASED ON CANDIDATE SOCIAL MEDIA PERFORMANCE

THESIS

Writing as one of the requirements
for obtaining a Masters degree
from Nusa Putra University

by
PUTU SUKMA DHARMALAKSANA DHADA
NIM: 20210130038
(*Master's Program in Informatics*)

UNIVERSITAS NUSA PUTRA
February 2024

ABSTRAK

MEMPREDIKSI PEMILU PRESIDEN DI INDONESIA MENGGUNAKAN MACHINE LEARNING BERDASARKAN PERFORMA MEDIA SOSIAL KANDIDAT

Oleh
Putu Sukma Dharmalaksana Dhada
NIM: 20210130038
(Program Studi Magister Informatika)

Indonesia memasuki tahun politik 2024 yang mana akan melaksanakan kompetisi politik berskala besar untuk pertama kalinya karena Pemilu 2024 adalah pertama kalinya Pemilu serentak dilaksanakan di Indonesia. Oleh karena itu, penting untuk memprediksi Pemilu agar negara dapat mengantisipasi kondisi keamanan dan ekonomi pasca Pemilu. Sampai saat ini di Indonesia, alat prediksi pemilu yang banyak digunakan adalah *polling*/survei. Padahal, *polling* memiliki kekurangan seperti membutuhkan waktu yang panjang, tidak mencakup seluruh demografi dan memerlukan biaya yang besar. Untuk menjawab kekurangan dari *polling* tersebut, penulis mengusulkan penggunaan *machine learning* untuk memproses data-data media sosial untuk memprediksi Pemilu Presiden di Indonesia tahun 2024. Algoritma *machine learning* yang digunakan yaitu SVR, dimana pada beberapa penelitian berhasil mendapat nilai MAE, MSE dan MAPE yang lebih baik dari pada Linear Regresion. Lebih dari 6.000 posting dikumpulkan langsung dari *official profile* media sosial (Facebook, Instagram dan Twitter) para kandidat dan diproses bersama dengan 28 hasil *polling* elektabilitas dari lembaga survei yang telah ditentukan sebagai variabel target. Hasilnya menunjukkan bahwa tingkat akurasi yang tinggi dapat dicapai dalam memprediksi perolehan elektabilitas para kandidat, ditunjukkan dengan nilai MAE, MSE dan MAPE yang kompetitif atau lebih baik dibandingkan *polling* yang dilakukan lembaga survei. Penggunaan Principal Component Analysis atau PCA pada tahap *data preparation* menjadi pembeda dalam penelitian ini karena memberikan akurasi yang jauh lebih baik dibandingkan pemodelan *machine learning* tanpa PCA.

Kata kunci: pemilu, prediksi, regresi, machine learning.

ABSTRACT

PREDICTING PRESIDENTIAL ELECTIONS IN INDONESIA USING MACHINE LEARNING BASED ON CANDIDATE SOCIAL MEDIA PERFORMANCE

By
Putu Sukma Dharmalaksana Dhada
NIM: 20210130038
(Master's Program in Informatics)

Entering the political year of 2024, Indonesia will conduct its first large-scale political competition because the 2024 Election is the inaugural simultaneous election in the country. Hence, predicting the election is crucial for anticipating post-election security and economic conditions. Currently, polls/surveys are widely used in Indonesia for election prediction, despite their drawbacks, such as lengthy timelines, incomplete demographic coverage, and high costs. To address these limitations, the author proposes the use of machine learning to process social media data to predict the 2024 Indonesian Presidential Election. The machine learning algorithm employed is Support Vector Regression (SVR), which, in several studies, has achieved better Mean Absolute Error (MAE), Mean Squared Error (MSE), and Mean Absolute Percentage Error (MAPE) values than Linear Regression. Over 6,000 posts were collected directly from the official social media profiles (Facebook, Instagram, and Twitter) of the candidates and processed along with the electability poll results from 28 survey institutions as target variables. The findings indicate that high accuracy in predicting the candidates' electability can be achieved, demonstrated by competitive or better MAE, MSE, and MAPE values compared to those of the survey institutions' polls. The use of Principal Component Analysis (PCA) in the data preparation stage marks a distinction in this research, as it provides significantly better accuracy than machine learning modeling without PCA.

Keywords: election, prediction, regression, machine learning.

1. AUTHOR'S STATEMENT

THESIS TITLE	: PREDICTING PRESIDENTIAL ELECTIONS IN INDONESIA USING MACHINE LEARNING BASED ON CANDIDATE SOCIAL MEDIA PERFORMANCE
NAME	: PUTU SUKMA DHARMALAKSANA DHADA
NIM	: 20210130038

"I declare and be responsible for the fact that this Thesis is my own work except for the snippets and summaries each of which I have explained the source of. If in the future there is another party who claims that this thesis is his work, which is accompanied by sufficient evidence, then I am willing to cancel my Master's degree in Computer Science along with all the rights and obligations attached to the title"

Sukabumi, August 2024

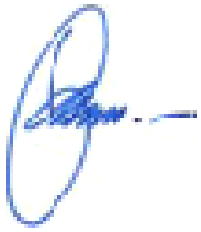
PUTU SUKMA DHARMALAKSANA DHADA

2. THESIS APPROVAL

THESIS TITLE	: PREDICTING PRESIDENTIAL ELECTIONS IN INDONESIA USING MACHINE LEARNING BASED ON CANDIDATE SOCIAL MEDIA PERFORMANCE
NAME	: PUTU SUKMA DHARMALAKSANA DHADA
NIM	: 20210130038

This script has been checked and approved

Sukabumi, August 2024

Head of Study Program,	Supervisor,
	
Prof. Ir. Teddy Mantoro, M.Sc., Ph.D.	Jelita Asian, M.Sc., PhD
NIDN. 0323096491	NIDN. 0406097702

3. THESIS VALIDATION

THESIS TITLE	: PREDICTING PRESIDENTIAL ELECTIONS IN INDONESIA USING MACHINE LEARNING BASED ON CANDIDATE SOCIAL MEDIA PERFORMANCE
NAME	: PUTU SUKMA DHARMALAKSANA DHADA
NIM	: 20210130038

This Thesis has been tested and defended in front of the Board of Examiners in Thesis session on 4 Agustus 2024. In our review, this Thesis adequate in terms of quality for the purpose of awarding the Master of Computer Degree (M.Sc.).

Sukabumi, August 2024

Supervisor I

Jelita Asian, M.Sc., PhD

NIP. 0406097702

Examiner I

Haris Al qodri Maarif, M.Sc., PhD

NIP. 0418068505

Supervisor II

Deshinta Arrova Dewi, S.Kom., M.Si., PhD

NIP. 0416127102

Examiner II

Prof. Ir. Media Anugerah Ayu, M.Sc., PhD

NIP.

Dipersembahkan kepada orang tua, istri, anak, mertua serta keluarga besarku tercinta yang senantiasa mendukung.

FOREWORD

My deepest gratitude goes to the presence of God Almighty for all the blessings and abundance of His grace so that the writer can complete the thesis research with the title "Predicting Election in Indonesia using Machine Learning based on Candidate Social Media Performance".

The purpose of writing this thesis is to fulfill the requirements to achieve a Masters degree in Informatics at the Postgraduate Study Program at Nusa Putra University.

TABLE OF CONTENTS

APPENDIX LIST

LIST OF IMAGES AND ILLUSTRATIONS

LIST OF TABLES

CHAPTER I INTRODUCTION

1.1. Background Research

Predicting presidential elections has been a topic of interest for many researchers and political analysts. The advance of machine learning and the rise of social media have opened up new avenues for predicting election outcomes. This research aims to explore the potential of using machine learning to predict presidential elections in Indonesia based on the social media performance of the candidates.

Indonesia as the world's third-largest democracy country has a vibrant political landscape (Sukma, 2022). The country transitioned to a democratic system in 1998, and since then, it has held six presidential elections. The first direct presidential election was held in 2004, and the most recent one took place in 2019. Each election has been marked by intense competition and high voter turnout, reflecting the robustness of Indonesian democratic system.

Social media has played an increasingly significant role in these elections. Indonesia has one of the highest rates of social media usage in the world, and politicians have leveraged this to reach out to voters. Candidates use social media platforms to communicate their policies, engage with voters, and respond to criticisms. The performance of candidates on social media, including the number of followers, likes, shares, and comments they received, can provide valuable insights into their popularity and their public perception of them.

There have been several attempts to predict election outcomes using machine learning. These studies have used various data sources, including opinion polls, economic indicators, and social media data. For instance, a study conducted on the 2016 U.S. presidential election used machine learning algorithms to analyze Twitter data and successfully predicted the election outcome (Srinivasan et al., n.d.), (Joyce & Deng, 2017). However, there has been limited research on predicting Indonesian presidential elections using machine learning and social media data. In addition, research on election predictions in Indonesia is still focuses on using data only from Twitter and using the sentiment analysis method. Several

researchers revealed that using only Twitter data is not enough in predicting elections (K. Brito et al., 2022; Gayo-Avello, 2012; Hinch, 2017).

This research could have significant implications. First, it could provide a more accurate and timely prediction of election outcomes, which could be useful for political analysts, policymakers, and voters. Second, it could shed light on the role of social media in Indonesian politics, which could inform strategies for political communication and engagement. Lastly, it could contribute to the growing body of literature on the use of machine learning in political science.

However, this research also poses several challenges. The accuracy of predictions could be affected by various factors, including the quality of social media data, the appropriateness of the machine learning algorithms used, and the volatility of public opinion. Moreover, ethical considerations, such as privacy and fairness, need to be taken into account when using social media data for research purposes.

In conclusion, predicting presidential elections in Indonesia using machine learning based on candidate social media performance is a promising but challenging research area. This study introduces an innovative method of utilizing Machine Learning models to forecast election results. This method is centered around the creation and application of social media data in a unique way, with a focus on the impact of posts from official candidate profiles across the three main social media platforms (Facebook, Twitter, and Instagram). This data was then merged with conventional polls and used to individually train Machine Learning models for each candidate to estimate their vote percentage. The modeling process employed a Support Vector Regression (SVR) and a traditional linear regression technique for baseline. Experiments were carried out in the context of the 2024 Indonesian presidential election and the results were statistically evaluated. Finally, our proposed method was compared with the final polls before the elections and the most recent cutting-edge research.

1.2. Problem Statement

The primary goal of this research is to establish a procedure and develop a machine learning model that leverages the social media performance of political

candidates to predict election outcomes. These predictions aim to compete with the results of conventional polling methods. Thus, the following research questions were defined:

RQ1. Is it possible to determine the process and develop a machine learning model that can forecast election outcomes by analyzing the social media performance of the candidates?

RQ2. In machine learning modeling with the SVR algorithm and candidate social media performance data, which social media features are most significant in predicting election results?

1.3. Research Objectives

The main objectives of this research are:

4. To provide more precise and imediate potential election prediction, which in turn could be beneficial for those involved in political analysis, policy creation, and the voting public. This is particularly important for the stakeholders who need to anticipate potential societal disruptions when election results are announced.
5. To shed light on the roles of social media in the political landscape of Indonesia, which could provide insights for the development of political communication and engagement strategies. This is particularly relevant for political entities in Indonesia, such as political parties.
6. To contribute the development of literature on the application of machine learning in the field of political science.

1.4. Significance of Research

This research determines the capability of machine learning algorithms to predict the outcomes of presidential elections by analyzing data from the social media performance of the candidates. It underscores the potential of social media as a tool for political figures to conduct their campaigns, enhance their chances of winning, and establish a closer connection with their electorate. These aspects can be quantitatively assessed through the application of machine learning.

1.5. Limitation of Problems and Assumptions

This research was conducted to predict the Indonesian presidential election in 2024. Data was collected from January 1st, 2023 to February 5th, 2024, which is 9 days before the voting date (February 14th, 2024). The social media studied were the candidates' personal official accounts, including Instagram, Facebook and Twitter. The candidates studied were Ganjar Pranowo, Anies Baswedan and Prabowo Subianto (excluding accounts of their political parties, campaign teams and others).

1.6. Thesis Structure

The outline of this thesis is organized as follows:

- Chapter I describes the background of the problem that will be discussed in the thesis;
- Chapter II describes the literature review of the thesis;
- Chapter III describes the methodology of the thesis;
- Chapter IV determines the experiment results and discussion;
- Chapter V determines the conclusion of the thesis and the future work.

CHAPTER V CONCLUSIONS AND RECOMMENDATIONS

5.1. Conclusion

This research does not claim to have made the first discovery regarding presidential election predictions, but it can overcome several shortcomings of previous election prediction methods such as sentiment analysis and of course traditional surveys.

The obstacles related to sampling have been tackled. Utilizing social networks for sample collection is counterbalanced by training the models with conventional polls. The representation issue with Twitter, being a proxy for all platforms, is mitigated by the proposal's ability to harness data from multiple social media sources, focusing on those most frequented by candidates and the public. This system is also adaptable to emerging platforms. The shift in gauging social media efficacy, from counting discussions about a candidate to assessing the engagement and interactions with their content, means that there is less data to gather: merely thousands of posts from a handful of candidates as opposed to millions from the entire populace. There is no need to specify keywords for data retrieval. Data can be gathered over an extended timeframe, like 10 months prior to elections, and the flexibility in selecting various data window durations eliminates arbitrary decision-making on the data collection span. Additionally, the limited data quantity simplifies processing and reduces computational demands.

Regarding the challenges in model creation, the issue of being highly vulnerable to volume manipulation was tackled by individually training and forecasting each candidate's outcomes. In this manner, the model was tailored to the unique patterns of a candidate’s followers, incorporating the activities of any potential network of BOTs or sponsored campaigns, if present.

5.2. Recommendations

This research combined expertise from fields such as electoral forecasting, social media analysis, and machine learning, underscoring distinct contributions to each domain. In the realm of electoral forecasting, vital to many democracies, we introduced a method that supplements conventional polling techniques. As shown

in our research, this approach can be employed for real-time predictions on days lacking polls or integrated into established polling methodologies, utilizing both interview and social media data. It further allows political figures and parties to gauge their standing solely based on their social media data and preceding surveys, offering a rapid, cost-effective way for daily prognostications.

In terms of social media research, our study explored the potential of identifying non-linear relationships between online activities and real-world consequences. Such insights are often challenging for social scientists due to hurdles in accessing and managing the extensive data required. Further research might provide insights into the utilization of social media by candidates in these nations, the level of user engagement, and the polarization experienced during the elections.

REFERENCES

Alzubi, J., Nayyar, A., & Kumar, A. (2018). Machine Learning from Theory to Algorithms: An Overview. *Journal of Physics: Conference Series*, *1142*, 012012. https://doi.org/10.1088/1742-6596/1142/1/012012

Asshiddiqie, J. (2009). *Pengantar Ilmu Hukum Tata Negara*. Rajagrafindo Persada. https://books.google.co.id/books?id=lIrtzwEACAAJ

Bimber, B. (2014). Digital Media in the Obama Campaigns of 2008 and 2012: Adaptation to the Personalized Political Communication Environment. *Journal of Information Technology & Politics*, *11*(2), 130–150. https://doi.org/10.1080/19331681.2014.895691

Brito, K., & Adeodato, P. J. L. (2023). Machine learning for predicting elections in Latin America based on social media engagement and polls. *Government Information Quarterly*, *40*(1), 101782. https://doi.org/10.1016/j.giq.2022.101782

Brito, K. D. S., & Adeodato, P. J. L. (2020). Predicting Brazilian and U.S. Elections with Machine Learning and Social Media Data. *2020 International Joint Conference on Neural Networks (IJCNN)*, 1–8. https://doi.org/10.1109/IJCNN48605.2020.9207147

Brito, K., Silva Filho, R. L. C., & Adeodato, P. (2022). Please stop trying to predict elections only with Twitter. *DG.O 2022: The 23rd Annual International Conference on Digital Government Research*, 88–95. https://doi.org/10.1145/3543434.3543648

Cen, L., Vu, Q. H., & Ruta, D. (2019). *Efficient Support Vector Regression with Reduced Training Data*. 15–18. https://doi.org/10.15439/2019F362

Cholisin, C. (2009). Mengembangkan partisipasi warga negara dalam memelihara dan mengembangkan sistem politik Indonesia. *Jurnal Civics: Media Kajian Kewarganegaraan*, *6*(1).

Crossley, A. M. (1937). Straw Polls in 1936. *Public Opinion Quarterly*, *1*(1), 24. https://doi.org/10.1086/265035

Dessart, L. (2017). Social media engagement: A model of antecedents and relational outcomes. *Journal of Marketing Management*, 1–25. https://doi.org/10.1080/0267257X.2017.1302975

Dinesh, A., Anitha Selvasofia, S. D., Datcheen, K. S., & Rakhesh Varshan, D. (2023). Machine learning for strength evaluation of concrete structures – Critical review. *Materials Today: Proceedings*, S221478532301982X. https://doi.org/10.1016/j.matpr.2023.04.090

Edgar, T. W., & Manz, D. O. (2017). Machine Learning. In *Research Methods for Cyber Security* (pp. 153–173). Elsevier. https://doi.org/10.1016/B978-0-12-805349-2.00006-6

Francia, P. L. (2018). Free Media and Twitter in the 2016 Presidential Election: The Unconventional Campaign of Donald Trump. *Social Science Computer Review*, *36*(4), 440–455. https://doi.org/10.1177/0894439317730302

Gayo-Avello, D. (2012). No, You Cannot Predict Elections with Twitter. *IEEE Internet Computing*, *16*(6), 91–94. https://doi.org/10.1109/MIC.2012.137

Hall, W., Tinati, R., & Jennings, W. (2018). From Brexit to Trump: Social Media's Role in Democracy. *Computer*, *51*(1), 18–27. https://doi.org/10.1109/MC.2018.1151005

Hallock, W., Roggeveen, A. L., & Crittenden, V. (2019). Firm-level perspectives on social media engagement: An exploratory study. *Qualitative Market Research: An International Journal*, *22*(2), 217–226. https://doi.org/10.1108/QMR-01-2017-0025

Hamiruddin, H. (2021). SURVEI DAN KONSULTAN POLITIK: MEMBANGUN POPULARITAS DAN ELEKTABILITAS POLITIK. *Al-Din: Jurnal Dakwah dan Sosial Keagamaan*, *7*(1), Article 1. https://doi.org/10.35673/ajdsk.v7i1.1702

Hinch, J. (2017). *#MakeAmericasPollsGreatAgain: Evaluating Twitter as a tool to predict election outcomes*.

Hou, A. C. Y., & Shiau, W.-L. (2019). Understanding Facebook to Instagram migration: A push-pull migration model perspective. *Information Technology & People*, *33*(1), 272–295. https://doi.org/10.1108/ITP-06-2017-0198

Joyce, B., & Deng, J. (2017). Sentiment analysis of tweets for the 2016 US presidential election. *2017 IEEE MIT Undergraduate Research Technology Conference (URTC)*, 1–4. https://doi.org/10.1109/URTC.2017.8284176

Laor, T. (2022). My social network: Group differences in frequency of use, active use, and interactive use on Facebook, Instagram and Twitter. *Technology in Society*, *68*, 101922. https://doi.org/10.1016/j.techsoc.2022.101922

Nasrullah, R. (2015). Media sosial: Perspektif komunikasi, budaya, dan sosioteknologi. *Bandung: Simbiosa Rekatama Media*, *2016*, 2017.

Nguyen, N.-H., Abellán-García, J., Lee, S., Garcia-Castano, E., & Vo, T. P. (2022). Efficient estimating compressive strength of ultra-high performance

concrete using XGBoost model. *Journal of Building Engineering*, *52*, 104302. https://doi.org/10.1016/j.jobe.2022.104302

O'Connor, B., Balasubramanyan, R., Routledge, B., & Smith, N. (2010). From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. *Proceedings of the International AAAI Conference on Web and Social Media*, *4*(1), 122–129. https://doi.org/10.1609/icwsm.v4i1.14031

Oppenheimer, B. I., Stimson, J. A., & Waterman, R. W. (1986). Interpreting U. S. Congressional Elections: The Exposure Thesis. *Legislative Studies Quarterly*, *11*(2), 227. https://doi.org/10.2307/439877

Pamungkas, S. (2009). *Perihal pemilu* (Cet. 1). Laboratorium Jurusan Ilmu Pemerintahan dan Jurusan Ilmu Pemerintahan, Universitas Gadjah Mada.

Puntoadi, D. (2011). *Menciptakan Penjualan via Social Media*. Elex Media Komputindo. https://books.google.co.id/books?id=g85BDwAAQBAJ

Republika. (n.d.). *Ini Tiga Lembaga Survei Terbaik Prediksi Pilkada DKI, Versi IWD*. Retrieved July 26, 2023, from https://news.republika.co.id/berita/nasional/pilkada/17/02/27/om17x3372-ini-tiga-lembaga-survei-terbaik-prediksi-pilkada-dki-versi-iwd?

Rudin, C., & Wagstaff, K. L. (2014). Machine learning for science and society. *Machine Learning*, *95*(1), 1–9. https://doi.org/10.1007/s10994-013-5425-9

Sheldon, P., & Bryant, K. (2016). Instagram: Motives for its use and relationship to narcissism and contextual age. *Computers in Human Behavior*, *58*, 89–97. https://doi.org/10.1016/j.chb.2015.12.059

Sindonews. (n.d.). *Survei Poltracking Indonesia Paling Akurat Prediksi Pilpres 2019*. Retrieved July 26, 2023, from

https://nasional.sindonews.com/berita/1407956/12/survei-poltracking-indonesia-paling-akurat-prediksi-pilpres-2019

Solihah, R., & Witianti, S. (2017). Permasalahan dan Upaya Mewujudkan Pemilu Demokratis di Indonesia Pasca Reformasi. *Jurnal Bawaslu*, *3*(1), 13–33.

Srinivasan, S. M., Sangwan, R. S., Neill, C. J., & Zu, T. (n.d.). *Power of Predictive Analytics: Using Emotion Classification of Twitter Data for Predicting the 2016 U.S. Presidential Elections*. *8*(1).

Sukma, R. (2022, December 22). *Decline of Democracy*. https://app.komp.as/ktG9cr5fD4VPVY6D6

Sulistyono, A., & Jakaria, J. (2022). ANALISIS PENGARUH SOCIAL MEDIA ENGAGEMENT TERHADAP RELATIONSHIP QUALITY YANG DIMEDIASI OLEH FAKTOR – FAKTOR RELATIONSHIP MANAGEMENT. *Equilibrium : Jurnal Ilmiah Ekonomi, Manajemen dan Akuntansi*, *11*(1), 53. https://doi.org/10.35906/equili.v11i1.987

Syafei, M., & Darajati, M. R. (2020). Design of General Election in Indonesia. *LAW REFORM*, *16*(1), 97–111. https://doi.org/10.14710/lr.v16i1.30308

Tsai, W.-H. S., & Men, L. R. (2017). Consumer engagement with brands on social network sites: A cross-cultural comparison of China and the USA. *Journal of Marketing Communications*, *23*(1), 2–21. https://doi.org/10.1080/13527266.2014.942678

Tumasjan, A., Sprenger, T., Sandner, P., & Welpe, I. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. *Proceedings of the International AAAI Conference on Web and Social Media*, *4*(1), 178–185. https://doi.org/10.1609/icwsm.v4i1.14009

We Are Social. (2023). *Digital 2023: Global Overview Report*. https://wearesocial.com/uk/blog/2023/01/digital-2023/

We Are Social, & Meltwater. (n.d.). *Digital 2023: Indonesia*. Retrieved July 22, 2023, from https://datareportal.com/reports/digital-2023-indonesia

Xu, L., Yan, X., & Zhang, Z. (2019). Research on the Causes of the "Tik Tok" App Becoming Popular and the Existing Problems. *Journal of Advanced Management Science*, 59–63. https://doi.org/10.18178/joams.7.2.59-63

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology*, *9*(2, Pt.2), 1–27. https://doi.org/10.1037/h0025848

Zolghadr, M., Niaki, S. A. A., & Niaki, S. T. A. (2018). Modeling and forecasting US presidential election using learning algorithms. *Journal of Industrial Engineering International*, *14*(3), 491–500. https://doi.org/10.1007/s40092-017-0238-2

Zuhro, R. S. (2019). Demokrasi dan Pemilu Presiden 2019. *Jurnal Penelitian Politik*, *16*(1), 69. https://doi.org/10.14203/jpp.v16i1.782