SUPPORTING PREPARATION FOR INDONESIA'S 2024 ELECTION:

BUILDING A PREDICTIVE MODEL WITH PUBLIC OPINION FROM TWITTER DATA AND THE IMPACT OF KEY PHENOMENA, VALIDATED BY PUBLIC ELECTION SURVEYS

Goran Fadhil Basyar

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

2109544

COMMITTEE

Ir. Federico Zamberlan

LOCATION

Tilburg University

School of Humanities and Digital Sciences

Department of Cognitive Science & Artificial Intelligence

Tilburg, The Netherlands

DATE

Jan 14th, 2023

WORD COUNT

8665

ACKNOWLEDGMENTS

I would like to express my gratitude for Ir. Federico Zamberlan, for the guidance and support through my time making this thesis

SUPPORTING PREPARATION FOR INDONESIA'S 2024 ELECTION:

BUILDING A PREDICTIVE MODEL WITH PUBLIC OPINION FROM TWITTER DATA AND THE IMPACT OF KEY PHENOMENA, VALIDATED BY PUBLIC ELECTION SURVEYS

Goran Fadhil Basyar

Abstract

Elections stand as an important events that shape the trajectory of a nation's future and supporting the prediction will tap to its potential to enhance the reliability of electoral predictions, thereby contributing to informed decision-making processes. Traditional election forecasting models in political science generally take preference in poll surveys and economic growth at the national level as the predictive factors. However, frequent and detailed polling is costly. With the rise of internet especially social media, particularly Twitter, researchers have explored its potential to reflect the political landscape. In the recent decades, the exponential growth of social media has drawn enormous research interests from various disciplines. Existing studies suggest that social media data have the potential to reflect the political landscape. Particularly, Twitter data have been extensively used to predict election outcomes around the world. However, most of the studies correlate twitter sentiment directly and solely with the election results without the proper feature explain ability and incorporating other external factors which can hardly be regarded as predictions. To develop a more advance approach this study incorporates unexplored factors such as the impact of natural disasters, economic crises, biased policies, and other influential phenomena. The study seeks to contribute a better understanding of these factors by identifying and assessing their unique contributions through the incorporation of essential features in the predictive model in the scope of Indonesian election. By doing so, the research aims to offer a more comprehensive and accurate prediction of electoral outcomes, transcending the limitations of existing methods which later can be developed further for a national election in the future. This research went a thorough exploration of the evolution of Natural Language Processing (NLP) techniques. Beginning with early basic NLP models and progressing to sophisticated models like BERT, the analysis extends beyond mere evaluation scores. It delves into how these NLP models influence predictions and examines their interpretability. Results show that the proposed method of incorporating the important phenomenon can contribute to the prediction model with a better explain ability which can be reflected by important feature extraction which also help making sense of a transparent prediction model so later the nation can be prepared to the future candidate leadership to a better future.

Data Source, Ethics, Code, and Technology (DSECT) statement:

0. Source/Code/Ethics/Technology Statement Example

The dataset for this research is a collection of dataset sources tailored to support the prediction of 2024 presidential election in Indonesia using sentiment analysis of Twitter data and the incorporation of key phenomenon. It combines Twitter data, external event data from Wikipedia, and a pooling comparison dataset for validation which extracted from Wikipedia. The code used for this research can be found on my GitHub https://github.com/GoranFad/Thesis/. There is no assistance used in the process of writing this thesis, including generative language models,

grammar and spell checker, or any other tools.

1. Introduction

Elections, democracy, and politics stand as the foundation of societal development, shaping the destiny of nations and influencing the course of history. The democratic history of Indonesia, begin by a progression from dictatorial rule to a democracy, making a crucial backdrop to the upcoming elections. Indonesia's progress toward democracy gained momentum in 1998, leading to the establishment of a dynamic and diverse political system (Liddle, 2000). The forthcoming elections represent a pivotal moment in this democratic narrative, where the electorate plays an important role in shaping the nation's trajectory. As Indonesia embraces a democratic model, the elections serve as a foundation of citizen empowerment and collective decision-making. This democratic evolution emphasize the accurate election predictions, as they not only reflect the will of the people but also contribute to the compelling and strengthening of Indonesia's democratic institutions. The diverse social, economic, and environmental conditions considered in the predictive model mirror the complexity of Indonesia's democratic situation, emphasizing the importance of a holistic understanding to make sure the integrity and inclusivity of the electoral process.

Recognizing the impact these elements have on the fabric of our societies, this thesis embarks on a crucial journey into the realm of electoral studies, with a particular focus on the vibrant and diverse nation of Indonesia which will be held pretty soon (Figure 1.1. below are the candidates from left to right : Prabowo Subianto, Anies Baswedan and Ganjaro Pranowo). The main goal is to construct a predictive model that not only enhances the electoral process but also delves into exploring new areas beyond the usual limits of sentiment analysis on social media platforms as well as the important national daily phenomenon which later can also be proposed and developed for the better tools for the future election.

Figure 1. 1 2024 Presidency Candidates

In the democratic nations, elections are crucial, serving as the foundation for the expression of public will and the determination of governance trajectories. This research acknowledges the fundamental importance of elections and endeavors to contribute meaningfully to the existing body of knowledge. It aims to go beyond the common comparison of sentiment analysis on platforms like Twitter and with traditional polling methods. Instead, it expand into unexplored dimensions, examining how important daily national events such as natural disasters, economic crises, corruption and biased policies impact the predictive accuracy of the model, acknowledging the multifaceted nature of electoral dynamics.

As Pereira aptly points out, citizens often tend towards governments that align with

positive economic or social outcomes during their tenure, while they are likely to disapprove those perceived as responsible for bad performance (Pereira, 2019) This insight shows how political, economic, and social factors in the electoral landscape, forming the background of this research

While sentiment analysis on Twitter has been explored as a potential substitute to traditional polling methods, this study seeks to expand the view by incorporating different external factors that may not be comprehensively captured by social media users alone. Moreover, it aims to uncover the evolving landscape of natural language processing (NLP) techniques, examining their predictive performance and interpretability—from initial models to state-of-the-art approaches like BERT.

In essence, this thesis tires to fill critical gaps in existing literature, offering a better understanding of electoral predictions by integrating various factors. By comprehensively evaluating the influence of evolving NLP techniques, the research seek to provide valuable insights that contribute to the refinement of electoral prediction models, addressing the limitations of current methodologies and advancing the understanding of the complex dynamics nature in election forecasting. In doing so, the study recognize the importance of elections, democracy, and politics in the complex picture of society moving forward

The main research question of this study that are trying to be answered:

> *Does incorporating the inclusion of key phenomena, representing nationwide daily important events or occurrences, as features in the predictive model correlate with their impact on election predictions? To what extent does the contribution has as revealed through the analysis of feature importance lists generated after model training?*

The sub-question can be listed separately, as such:

> RQ1 *How does the choice of featurization method such as LSA, Word2vec and BERT applied to tweet analysis and key national phenomenon differ on the predictive performance of the employed models?*
>
> Both LSA,Word2vec, BERT have a different capabilities for processing or understanding a text. This research aim to figure out how picking one of these methods affects how well our models predict things with the use of tweets along with the key national event. The research will investigate its performance and interpretability. By comparing these methods, it will reveal which method is better suit to the election problem.
>
> RQ2 *How can the key phenomena help to better understand the model interpretability?*
>
> This research aim to understand the addition of key phenomena will play a crucial role in improving the interpretability of models. In machine learning, interpretability stands for the ability to understand and explain why a model makes predictions. By closely analyzing the features contribution for its explanation through a word extraction from key phenomena, the study tries to examine whether it will hep to understand the prediction better or not.

2. Related Work

Several studies have been done to predict the election in several countries which helped support this study.

Liu et al. aimed enhance election forecasting by employing Twitter sentiment analysis as an alternative to conventional polling methods. They collected Tweets and county-level socio-economic data from September 26th to November 8th, 2016, and used this dataset to predict the 2016 US presidential election result. Firstly, they used a deep learning model to get the sentimental input to the model along with the economic variable such as GDP etc which later incorporates several regression model with cross validation to the target variable to predict the percentage of the election. The study shows that Twitter sentiment analysis could yield more accurate predictions compared to traditional polling data in some extent. This research did an advance model establishment, however, it is still lack of explainability.

Budiharto et al. focused on conducting sentiment analysis on Twitter data related to the 2019 Presidential election, utilizing top hashtags. The process began with crawling tweets against specific hashtags to gather relevant data, followed by preprocessing and cleaning steps such as removing Twitter handles, eliminating punctuation, numbers, and special characters, and discarding short words. Additional measures like tokenization and stemming were applied. The refined tweets were then subjected to polarity analysis using TextBlob, calculating the difference between positive and negative words in each text relative to the total count of sentiment words. The combined sentimental results were used to determine the candidate with the highest positive sentiment. In conclusion, Twitter proved to be a valuable tool for poll and opinion mining, particularly in predicting political election outcomes. (Budiharto & Meiliana, 2018).

Pereira's research in political science studied voting behavior, especially the Economic Voting Theory, which resulted that citizens tend to support governments with positive outcomes and reject those with poor performance. The research validates the Economic Voting Theory, but also notes voter uncertainty in choosing opposition candidates (Pereira, 2019).

Khan et al. looked into how social media platforms such as Facebook and Twitter can be used to predict elections by understanding people's political opinions. They focused on sentiment analysis, studying people's emotions rather than analyzing social networks. They conducted a systematic mapping study on Twitter election predictions from January 2010 to January 2021, identifying 787 related studies. After applying specific criteria, they narrowed it down to 98 primary studies. The findings revealed that most studies used sentiment analysis, followed by volume-based and social network analysis approaches. (Khan, et al., 2021). None of them not tapping the realm of the availability of important event.

Hasan et al studied the text and opinions on social media platforms using machine learning techniques, including sentiment and subjectivity analysis. They recognized a pressing need for an advanced approach in sentiment analysis for elections, despite the utilization of various machine-learning methods and tools. To address these challenges they utilized Texblob, SentiwordNet, W-WSD to directly getting the Negative, Neutral and positive sentiment to get the prediction. (Hasan, et al., 2018). Again, the study only utilizing tweets and no further explain ability to understand which word has a major contribution for the model.

Tricahyo and Isa, looked into how Indonesia picks its president, using a thing called Word2Vec. They tried out different ways of figuring out people's feelings, like K-Nearest Neighbors, Naïve Bayes, Decision Tree, Random Forest, and Support Vector Machine. The goal was to compare these methods and see which one gives the most accurate results. They gathered data about presidential candidates from Twitter, with about 640 entries in Bahasa Indonesia using keywords like Prabowo, Sandi, Jokowi, and Ma'ruf. The analysis revealed that the Random Forest Classification method exhibited the highest accuracy, reaching 98.33%, while the Support Vector Machine method demonstrated the lowest accuracy at 81.96% (Tricahyo and Isa, 2020).The research solely used the twitter data without further explanation of the words that important for making the model prediction especially explanation from other external factors.Moreover, it only use 1 NLP method which is Word2vec without any comparison with other NLP.

The existing literature provides valuable insights into the use of Twitter data sentiment analysis and election prediction methodologies, however, most of the studies direct sentiment analysizer or using majority of positive, neutral or negative view without comprehensive explianability of

text extraction and it overlooks the exploration of crucial phenomena or noteworthy events that contribute to prediction accuracy and interpretability. Additionally, in the context of election prediction, the majority of studies concentrate on constructing models directly using Twitter data, with limited emphasis on extracting meaning for underlying predictions. This study aims to leverage the abundant Twitter data, incorporating key national phenomenon data and conducting a thorough analysis of important words. Through this approach, the research seeks to better understand the dynamics of public sentiment, key event and its correlation with election outcomes in Indonesia with a comprehensive feature extraction for more human understanding for the model through word extraction.

3. Methodology & Experimental Setup

3.1. Data Scraping

The research begin with data scraping. In the data collection phase, Twitter Harvest, an open tool by Helmi Satria (Helmi Satria Satria, 2023), was employed to scrape Twitter data. The selection criteria involved tweets with likes and retweets surpassing 100, ensuring the weighty of posts that are both representative and important within the Twitter user community.
The compilation of key phenomenon and survey data consisted of extracting pertinent information from Wikipedia, spanning the years 2020 to 2023. This process resulted in the creation of a comprehensive dataset encompassing surveys and notable events during this specified timeframe.

3.2. Dataset Description

- Twitter Sentiment Data: Twitter data is systematically collected, focusing on specific, keywords ("Prabowo", "Ganjar", "Anies"), and mentions of the major candidates (Prabowo, Ganjar, and Anies). Data collection spans several years, starting from January 2020 until November 2023. There will be approximately 6000 Tweets collected per candidate, in this study, 1 candidate will be selected which is Prabowo that have 5725 tweets from January 2020 until November 2023.
- National Phenomenon /Event: The research includes important events and crises related to the election. The incorporation of national event data offers context for understanding how external factors influence sentiment and election predictions. The dataset will be 767 instances for the period of January 2020 until November 2023.
- Survey Polling: A candidate electability survey dataset is created to validate the accuracy of sentiment analysis predictions. This dataset includes election outcomes, the number of candidates and Polling firm. It serves as a benchmark for assessing the performance of sentiment analysis models. It has 157 rows and 4 Columns

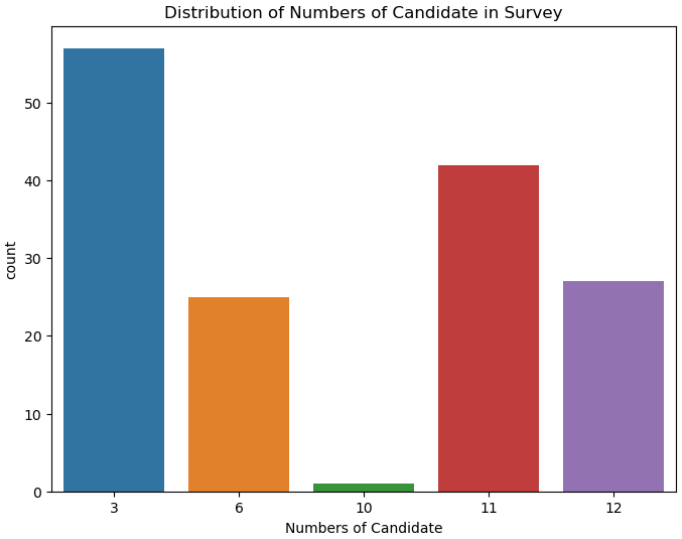

Figure 3. 1 Number of the Candidates within the Survey

Number of candidates vary from 3 to 12 candidates which dominated by 3 number of candidates which stands for the final candidate numbers of the 2024 Indonesian elections.

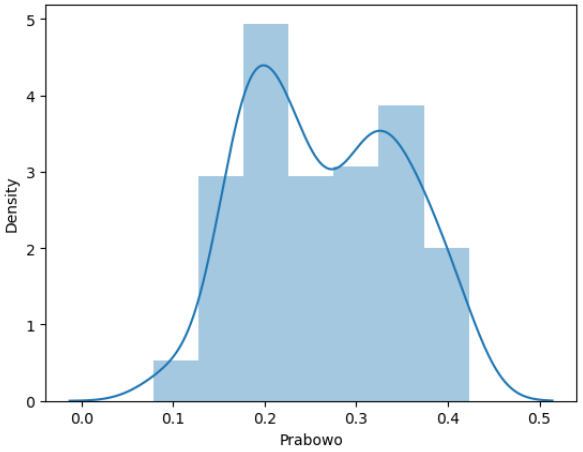

Figure 3. 2 Prabowo Electability by Density

The target variable of the candidate electability are in percentage and around 20%-30% being the majority electability for Prabowo Candidate which changed around time which can be seen in detail later for result explanation.

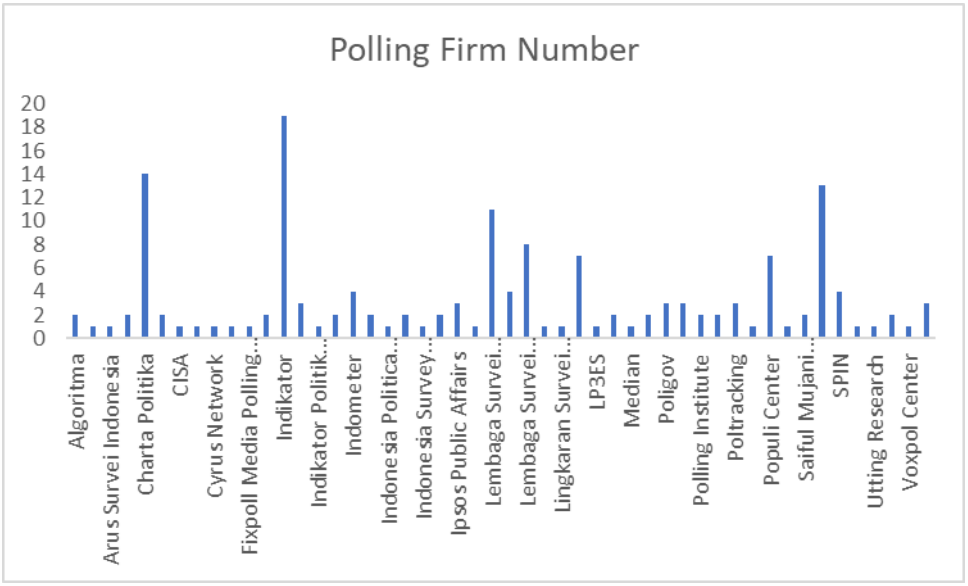

Figure 3. 3 Different Polling Firm Number in the Survey

Polling firm is one of the information in dataset that later will be used in the model. There are 50 different polling firms which conducted the survey of candidates electability in Indonesia.

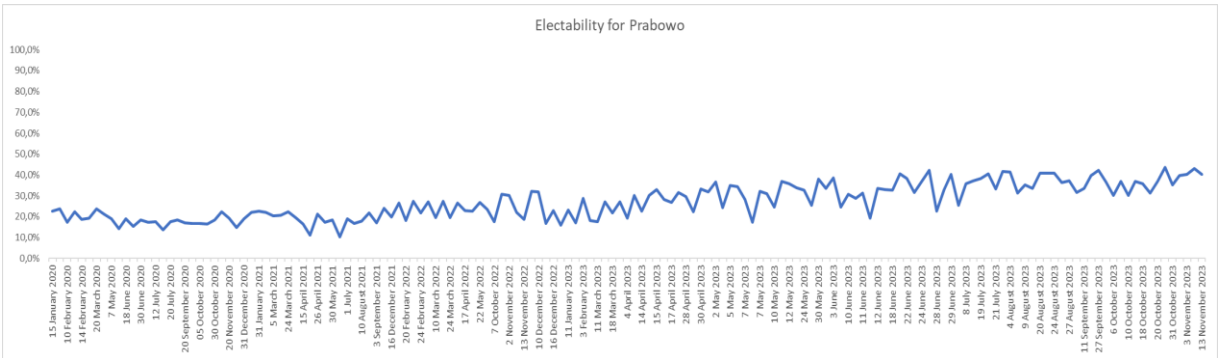

Figure 3. 4 Electability based on Time from 2020 until late 2023

As mentioned earlier, the dataset is based on timeseries from 2020 until 2023. From early 2020 there was some minor declining trend until the mid of 2021 and from there the trend was a minor increase until the early 2023. 56% of the data belong to 2023 year.

3.3. Data Cleaning and Preprocessing of independent variable.

Preprocessing of unstructured tweeter text as well as important events involves several steps. Text preprocessing procedures were implemented to enhance the quality of the data. Initially, text standardization is implemented, encompassing tasks such as converting text to lowercase, eliminating special characters, and removing numbers (Chandrasekar & Qian, 2016). Additionally, stop-word removal was applied to the entire text for both Latent Semantic Analysis (LSA) and Word2Vec since these words carry minimal information (Sarica & Luo, 2021) which later facilitating natural language processing (NLP) processes. For the tweet text since its in Indonesian language the stopword removal used an open library called Sastrawi .Notably, BERT, with its advanced capabilities, did not undergo stop-word removal, as these models, trained on extensive datasets, have demonstrated superior performance without the traditional preprocessing techniques like punctuation removal and stop word elimination (Alzahrani & Jololian, 2021).
In extracting candidate-specific data, the focus was placed on three candidates, with a detailed examination of survey results for Prabowo in the year 2023. Simultaneously, tweets containing the name "Prabowo" were identified and gathered, forming a dataset designed to capture sentiments and discussions specifically related to this candidate.
After the data has been cleaned and preprocessed, prior to applying individual NLP techniques, we undertake the segregation of tweets and events based on distinct date ranges from survey results. This separation is essential as survey results (target variable) typically span specific intervals, such as 33.4% survey responses for the period from June 12, 2023, to June 17, 2023, whereas important events and tweets are associated with a single date, like Tweet A being on June 12, 2023. Thus, we categorize the date range of surveys into a start date and an end date, representing the interval used to filter tweets and events specifically within that range. This approach ensures the alignment of relevant tweets and events with the corresponding survey data for accurate contextual analysis within defined timeframes.

3.4. Latent Semantic Analysis (LSA)

Latent Latent Semantic Analysis (LSA) is a high-dimensional linear associative model to analyze a collection of text and represents knowledge derived from it by obtaining words similarity and text documents (Wagire et al.,2020). Following the application of TF-IDF, the resulting data frame undergoes dimensionality reduction via Latent Semantic Analysis (LSA) using TruncatedSVD (Gupta and Patel, 2021). Initially, the TF-IDF generates an array with 405 columns. In determining the optimal number of components for Truncated Singular Value Decomposition (SVD), a thorough experimental exploration was conducted, considering both the cumulative explained variance and the performance of the model across varying numbers of components, ultimately leading to the selection of a specific number that strikingly balanced dimensional reduction and model effectiveness. In the process of LSA, after doing the TFIDF vectorizer and Truncated SVD, we average the score for each time interval to so it can give a feature for a particular target variable of the election survey LSA is well-suited for predicting candidate electability from Twitter and Wikipedia news data because it helps uncover hidden meanings and relationships in the text. By reducing the dimensionality of the textual features, LSA captures the essential semantic structure. This is crucial for understanding subtle nuances and latent themes in tweets and news articles related to political sentiment. Integrating LSA derived features into the Random Forest and KNN models enhances predictive performance by allowing them to grasp intricate patterns in public opinion expressed on social media and news, contributing to a more nuanced understanding of the factors influencing candidate electability.

3.5. Word2vec

Word2vec is a technique for natural language processing (NLP) published in 2013. The

word2vec algorithm uses a neural network model to learn word associations from a large corpus of text (Haritha et al., 2019). As mentioned by Word2vec, works really well, especially when compared to older methods like Latent Semantic Analysis (LSA), especially in tasks involving analogie (Gennaro et al, 2021).In the study, the cleaned sentence will be tokenized and will be vectorized for each of the word. Similar with LSA, this will undergo the average through time interval process to match to a particular survey result of target variables.
Word2Vec is particularly suitable for predicting candidate electability based on Twitter and news data due to its ability to capture the relationships between words in a vector space. By embedding words in a multi-dimensional space, Word2Vec represents contextual similarities and captures meanings, which is crucial when examining language from social media and news sources. Word2vec is good in understanding the context and sentiment of the text, allowing it to catch the subtleties and nuances of public opinion expressed in diverse and dynamic language. The embeddings generated by Word2Vec provide an advance representation of the textual content, making the predictive model to discover patterns and associations that may be indicative of public sentiment towards a candidate. Therefore, taking into account Word2Vec embeddings as features for fruther Random Forest and KNN models enhances the ability to interpret and predict the electability trends accurately.

3.6. BERT

BERT, a transfer learning model, operates without the need for preprocessing steps on unstructured tweet and model. Transfer learning techniques, which are independent of specific features, possess the capability to comprehend natural language within its context (Alzahrani & Jololian, 2021). BERT employs Masked Language Models (MLM) to enable pre-trained, comprehensive bidirectional representations. In MLM, a random subset of tokens within the input is masked, thereby challenging the model to predict the original vocabulary ID of the masked word based solely on its contextual cues (Devlin et al., 2018). As the tweet will be in Indonesian, IndoBERT will be used as it is a indonesian model of BERT ("IndoBERT", n.d.). As we know, BERT, is the most advance model(Devlin et al., 2018), so it will be used compare with other 2 NLP.Similar with 2 previous NLP used, after processing the cleaned text using BERT, it will be averaged within a timespan of the polling result.
BERT (Bidirectional Encoder Representations from Transformers) is well-suited for the electoral candidate prediction model due to its advanced contextualized embeddings and deep understanding of language semantics. Unlike traditional NLP models like LSA and Word2Vec, BERT captures bidirectional context in a text, allowing it to comprehend the advance relationships between words and their meanings. In the context of electoral sentiment analysis, BERT discover the intricacies of political discourse, identifying sentimental, and understanding the context surrounding candidate mentions on platforms like Twitter and Wikipedia. Its ability to capture small hints from written information makes it effective in predicting the electability percentage of candidates. Therefore, incorporating BERT embeddings as features in the Random Forest model enhances its capacity to leverage the rich contextual information present in the Twitter and Wikipedia data, thereby contributing to more accurate electoral predictions.

3.7. Timeseries Data set split and cross validation

The data set is split into 80% of training and 20% for testing purposes by time series.
Time series data inherently involves a temporal component, where observations are ordered based on time. By utilizing time series data split, the model recognize the sequential nature of the dataset, ensuring that the training set precedes the validation set in time (Cerqueira et al., 2020). This approach is crucial in capturing the temporal dependencies present in social media and Wikipedia news, as public sentiment and political candidates situation might change over time.

In addition to time series data split, employing cross-validation is essential for robust model evaluation. Cross-validation provides a comprehensive assessment of the model's performance by systematically splitting the data into training and validation sets multiple times (Cerqueira et al., 2020). Recognizing the dynamic nature of political landscapes and public opinion, cross-validation helps ensure that the model generalizes well to different periods, enhancing its capability in predicting electability beyond the specific time period used for training. This combined approach of time series data split and cross-validation enhances the overall validity of the predictive model, making it suited for real-world application in the context of electoral dynamics. Later, the best model will be chosen for the unseen data as can be seen in Figure 3.5.

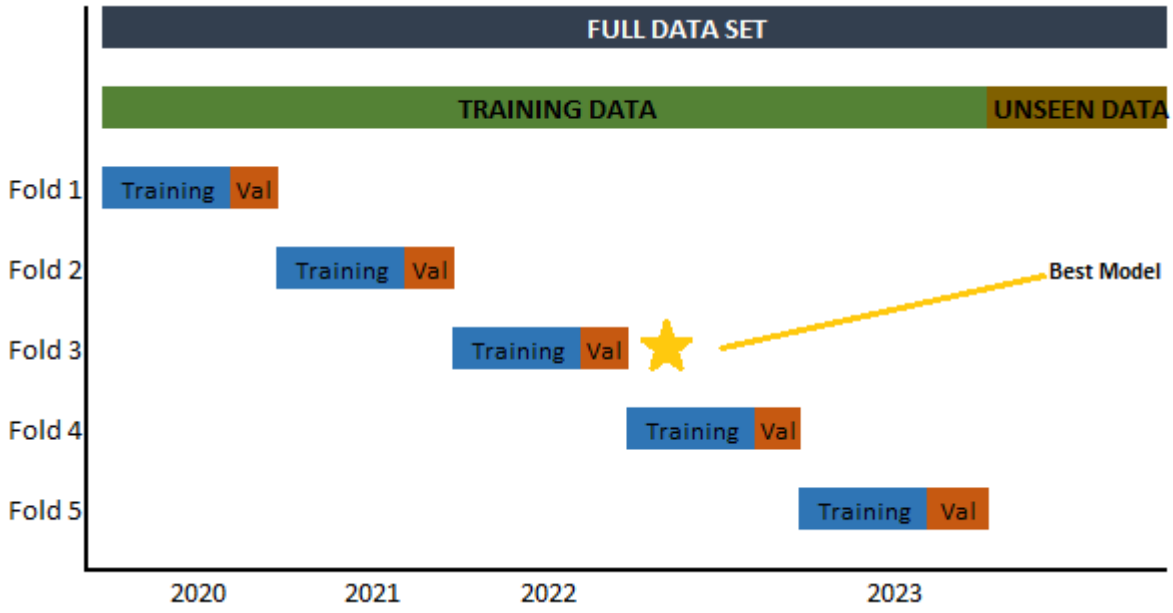

Figure 3. 5 Training and testing split cross validation methodology

3.8. Random Forest

Random Forest (RF) is a tree predictors, wherein for regression purposes,trees are cultivated based on random forests, guided by a random vector assigning numerical values rather than class labels to the tree predictor (Breiman, 2001). Random Forest perfectly in a high dimensional data (Biau and Scornet, 2016), it is because of the built-in capability for selecting the important features. This is also related having random forest model as its underlying that can select the important features using Random Forest (Darst et al., 2018).

Random Forest is well-suited for this election candidate electability prediction model due to its capability to handle complex relationships and interactions within a diverse set of features extracted from LSA, Word2Vec, and BERT embeddings. The ensemble nature of Random Forest, which compiled predictions from multiple decision trees, helps mitigate overfitting and enhances generalization performance. Moreover, Random Forest can effectively capture non-linearities and feature importance, crucial for understanding the complex patterns in social media and Wikipedia news. Its advance ability to handle noisy features and ability to handle large datasets make it suitable for the mixed data types and potential noise in Twitter and Wikipedia-derived features. The model's built-in capability to provide insights into feature importance allows analyzer to interpret and refine the importance of various linguistic and contextual elements in predicting electoral outcomes, making Random Forest a powerful for this case.

3.9. KNN

KNN, or k-nearest neighbors, is a model that doesn't assume the way the data is spread out (non-parametric). It doesn't try to figure out the exact pattern or distribution of the data which makes KNN easy to be used because it doesn't require an assumptions about the data structure; it just looks at the neighboring points to make predictions (Ray, 2019). Given one of the simplest method in ML, KNN works by finding the k nearest neighbors of a data, and afterwards generates predictions on the neighbors within the majority class in the k nearest neighbors (S. Zhang et al., 2018). Training data points are not used to do any generalization, which creates concerns in population size when using KNN (Song et al., 2017).
Using K-Nearest Neighbors (KNN) for predicting candidate electability based on Twitter and Wikipedia news data processed with three NLP techniques (LSA, Word2Vec, and BERT) can be beneficial in this context. KNN is a non-parametric and instance-based algorithm that relies on the similarity of data points. In the electoral context, KNN can capture the underlying patterns and similarities in the feature space, effectively identifying clusters of similar instances. Given the nature of social media data and news articles, where related information tends to be clustered, KNN can exploit these local patterns for prediction. The algorithm calculates the proximity of a new data point to its k-nearest neighbors, assigning a label based on the majority class within that neighborhood. This adaptability to local structures makes KNN particularly suitable for scenarios where the electability prediction may be influenced by specific clusters or patterns present in the input data.

3.10. Recursive Feature Elimination (RFE)

After being cleaned, pre-processed and processed by each of the NLP tools, the data set will consist of a very high dimensional. Introducing Recursive Feature Elimination (RFE) which is a dimensionality reduction technique commonly used in machine learning to enhance model performance and interpretability. It operates by iteratively removing the least crucial features from the dataset until an optimal subset is identified (Misra and Yadav, 2020). The process begins with training a model on the overall feature set and evaluating the importance of each feature. The least important features are then dropped, and the model is retrained on the reduced set. This recursive process continues until the desired number of features or optimal model performance is obtained. RFE is valuable in scenarios with high-dimensional data, as it helps to avoid overfitting and reduces complexity (Darst et al., 2018). It aims in identifying the most relevant features, improving model generalization, and facilitating a better understanding of the underlying relationships within the dataset.
Recursive Feature Elimination (RFE) is particularly suitable for this scenario of predicting candidate electability using a combination of NLP techniques such as LSA, Word2Vec, and BERT, alongside traditional polling data. RFE systematically evaluates the importance of features in the dataset by recursively removing the least crucial ones based on model performance. In this case, where dealing with diverse NLP features extracted from Twitter and Wikipedia news, RFE aids in identifying the most influential linguistic patterns and content that contribute crucially to predicting survey results. This stepwise elimination process ensures that only the most relevant features are retained, reducing dimensionality and potentially enhancing model interpretability. By incorporating RFE into the model and subsequently utilizing the Random Forest and KNN models, it can optimize the model's performance and identify the key linguistic and polling factors driving the predictions of candidate electability over time.

3.11. Hyperparameter tuning

Regarding hyperparameter optimization, both grid search (self-defined) and random search (Scikit-Learn) are explored. Grid search is explored to find the best parameters for for all 3 ML

models, where 5-fold Cross Validation is applied (Pedregosa et al.,2011). Being the most exhaustive approach to find the optimal hyperparameters, grid search goes through each possible combination of the hyperparameters. Grid search might require longer training time and high resource to find the optimal values. Random search, however, requires less time and resources, since hyperparameters are randomly selected to train the model with. Even though the absolute best set might not be explored, chances of finding a close best set are high. Results are usually sufficient (Agrawal, 2021).

Random Forest (RF):
The started the hyperparameter tuning for the Random Forest model by defining a parameter grid encompassing key parameters such as the number of estimators, maximum depth, minimum samples split, and minimum samples leaf. The model was then subjected to GridSearchCV, exploring the parameter combinations to identify the optimal set. The best hyperparameters were obtained based on the negative mean squared error, and the tuned Random Forest model was subsequently employed for predictions. With the following
estimators': 50, 100, 150
max depth: None, 10, 20, 30
min samples split: 2, 5, 10
min samples leaf: 1, 2, 4

K-Nearest Neighbors (KNN):
Hyperparameter tuning for the K-Nearest Neighbors model involved determining a parameter grid, including the number of neighbors, weighting scheme, and the distance metric. Utilizing GridSearchCV, the model was fine-tuned by systematically evaluating different hyperparameter combinations. The best set of hyperparameters, determined by the negative mean squared error, guided the selection of the best KNN model for subsequent predictions.
neighbors: 3, 5, 7
weights: 'uniform', 'distance'
p: 1, 2

3.12. Feature Importance

In the feature importance analysis using the Random Forest Regressor, the code not only ranks and prints the importance scores of the top features but also associates each feature with its corresponding index (Palczewska et al.,2013). Since, it was the average of a date range, the maximum value within the specific important component of the feature. For tweet-related and event-related features, this index is utilized to identify the most influential word in the associated term frequency-inverse document frequency (TF-IDF) matrix, showcasing the specific word that contributed the most to the feature's importance using get_feature_names_out . This additional step enhances the interpretability of the feature importance results, providing a more granular understanding of the influential components within the tweet and event features.

3.13. Evaluation

- Since it is a regression problem, R-square as a standard regression measurement will be chosen to evaluate the model. R-squared measures the proportion of the variance in the dependent variable that is predictable from the independent variable(s) (Chicco et al., 2021). R-squared provide insight into how well a model captures the variation in electoral predictions model. A high R-squared means that a portion of the variability in predictions is explained by the deployed model which is useful for a high-dimension

regression proble,

- The second evaluation method is Mean Average Error which has the natural measure of average error and well interpret ability which applies to our type of data-set (Willmott and Matsuura, 2005). MAE penalizes bigger errors more heavily, making it useful when it needs to emphasize and minimize the impact of larger prediction inaccuracies. This can be crucial in the context of electoral predictions, where accuracy is essential.
- The third one is MSE. The MSE either examine the quality of a predictor (i.e., a function mapping arbitrary inputs to a sample of values of some random variable), or of an estimator (i.e., a mathematical function mapping a sample of data to an estimate of a parameter of the population from which the data is sampled). In regards of prediction, understanding the prediction interval can also be useful as it provides a range within which a future observation will fall, with a certain probability which making it more robust to outliers.

3.14. Experimental Setup

To show the flow process of data science in this research, a chart is presented below:

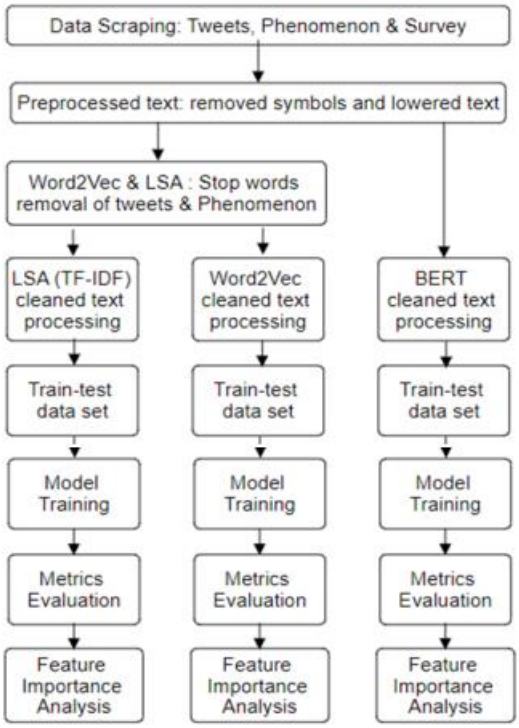

Figure 3. 6 Methodology flowchart

The initial process started with data scraping for tweeter (X) and the key phenomenon as well as the polling survey, in this case, the study focused one single candidate which is Prabowo which means the polling election percentage and the tweeter data will be based on Prabowo. Furthermore, Key Phenomenon and Tweet text features will be processed using LSA, Word2vec and BERT, the scores that came out from the NLP processing will be average based on the time range of the survey polling which will concatenate with number of candidate and polling firm from the survey data within the time range survey. Later to be fed to a selected ML model (RF and KNN) alongside with the polling firm as well as number of candidates as independent variable with the polling electability percentage as dependent variable from polling survey file. The datasets will be divided by 80% training set and 20% testing set based on its time order. Grid search will be used for hyperparameter tuning to find the best parameters with timeseries

cross validation of 5 folds along with Recursive Feature Elimination. The best models will then be applied to the unseen data/ test set. Random Forest and KNN models will undergo evaluation based on R-squared, MAE and MSE. More specifically for LSA, RF's feature importance score will be employed to select the most important features from each fold with a max score word extraction which later visualized through plot by time to check the progress of each feature importance.

4. Results

4.1. Overall Performance Evaluation

Timeseries						
	Random Forest			KNN		
	LSA	Word2vec	BERT	LSA	Word2vec	BERT
Mean Absolute Error (MAE)	0,073444689	0,0420707	0,0257050	0,0727575	0,0456500	0,0454913
Mean Squared Error (MSE)	0,008863557	0,0027533	0,0008479	0,0083241	0,0032256	0,0029536
R-squared (R2) Score:	-0,401841432	0,2474656	0,4563606	-0,31652215	0,1183733	0,1927376

Table 4. 1 Overall Performance Evaluation

In the results of this thesis, the analysis of three NLP techniques LSA, Word2Vec, and BERT revealed intriguing differences in their predictive performances for candidate electability percentages derived from Twitter and daily news from Wikipedia. Notably, the LSA-based model resulted in a negative R-squared value, suggesting a potential limitation in capturing the nuanced relationships within the dataset. Conversely, the Word2Vec and BERT models demonstrated positive R-squared values, showing their superior ability to handle the semantic intricacies and context-dependent nature of the data. Among the models applied, BERT consistently demonstrated best predictive capabilities across the board when integrated with Random Forest and k-Nearest Neighbors. The intricate contextual understanding and feature representation captured by BERT's pre-trained embeddings evidently contribute to its exceptional performance, showcasing its effectiveness in discovering different patterns within electoral data.

4.2. Fold base time performance

4.2.1.Yearly performance Evaluation

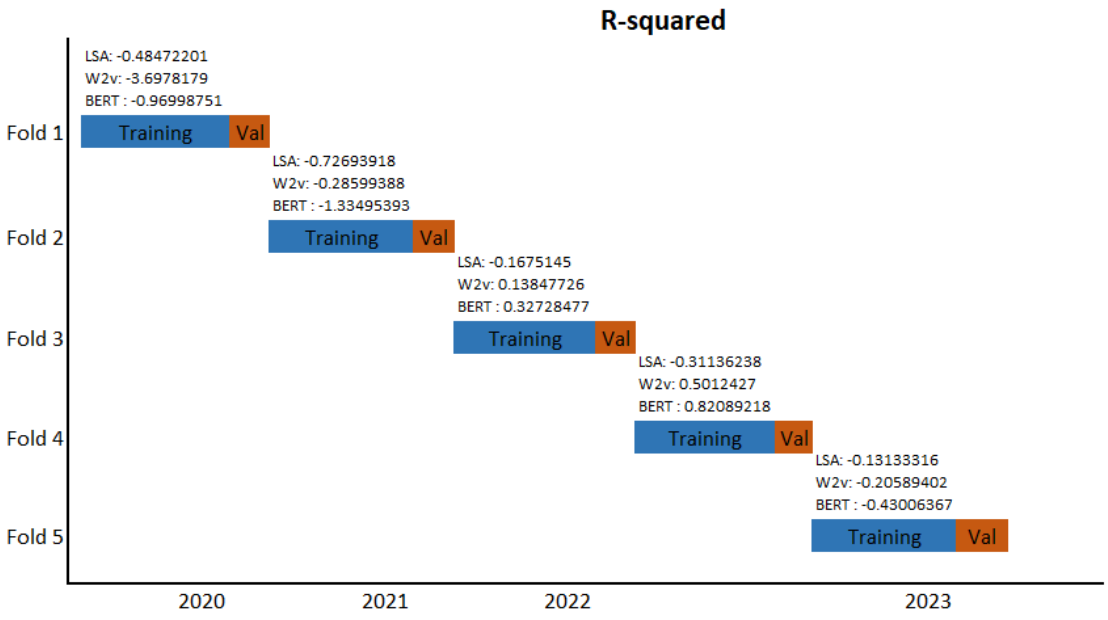

Figure 4. 1 Model Performance (RF only) by Rsquared

In figure 4.1. for Random forest fold performance, LSA showing a negative Rsquared across the fold which indicate the in capability to capture the timely dynamic changing nature of the data which has been stated above in the overall performance.
Below is each fold data condition, as we can see that the early stage of the political year, we can

see a substantial trend far way back from the election time. As it getting closer to the election period time, the trend is quite stable as we can see the in the last 20% of the test set which is just 3 months before the election time. In fold 1 there is a declining trend at the beginning and increasing trend at the middle. For fold 2,there was a steady line and in the middle there was a increasing trend. For fold 5, there was a major outlier at the beginning. In Fold 4, 5 and test set the trend was relatively stable.

As mentioned in the fold description we know that the early stage data has some trend which could not captured by the model, and the 2 fold near the test set it, it became better as it will come to near of the prediction, making it possible to predict the election. Therefore, in those 2 fold, it revealed as a good model especially the 4th fold. The negative rsquared can be seen in the fold chart which showing the noisiness during that particular time.

Similarly with Word2vec, BERT also having a best performance in 4th fold as it can captured the model quite well which can also be reflected in the target variable dataset description . The negative rsquared can be seen in the fold chart which showing the noisiness during that particular time.

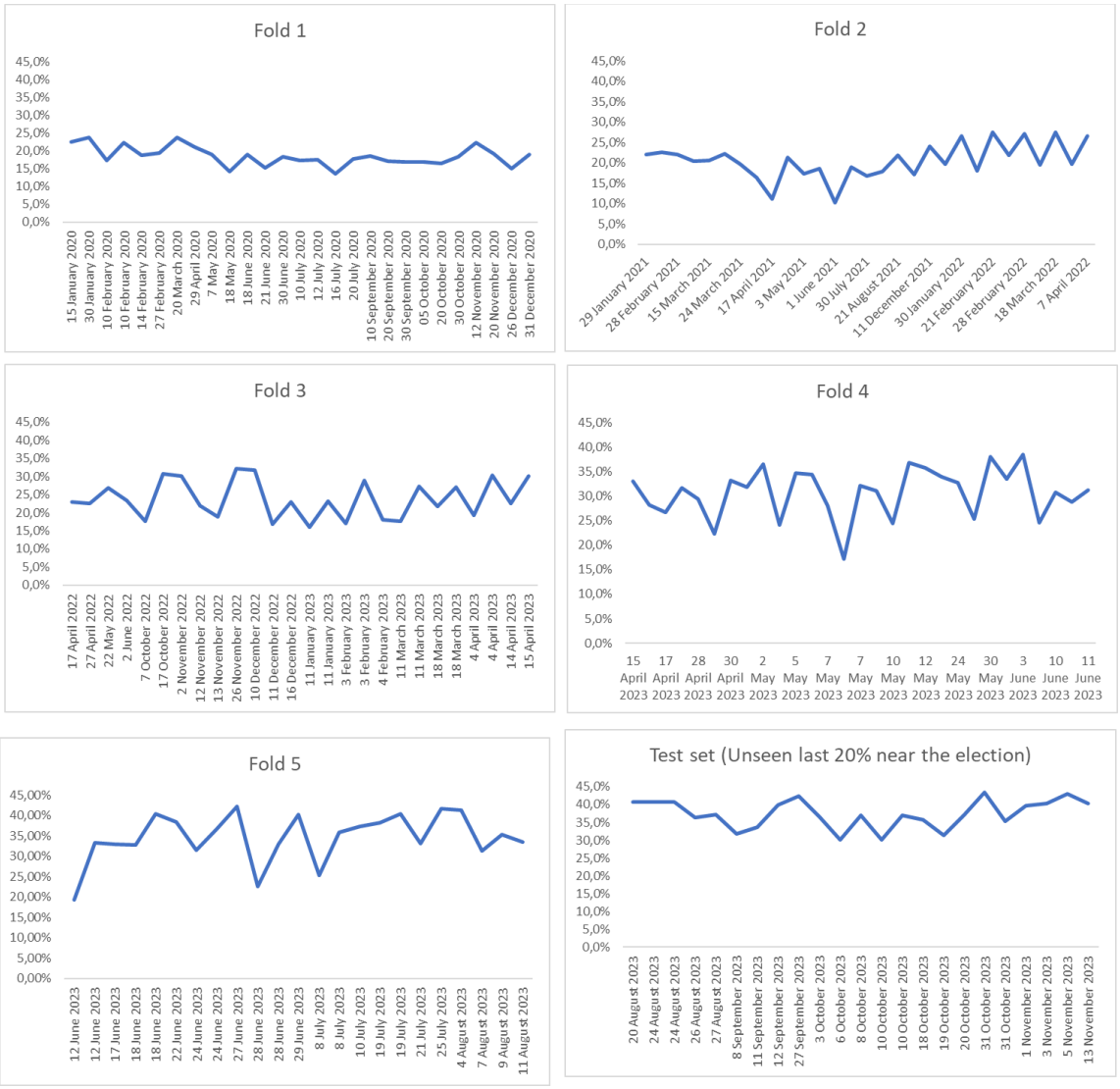

Figure 4. 2 Fold time dataset condition

4.2.2.Important Features for word extraction (For LSA)

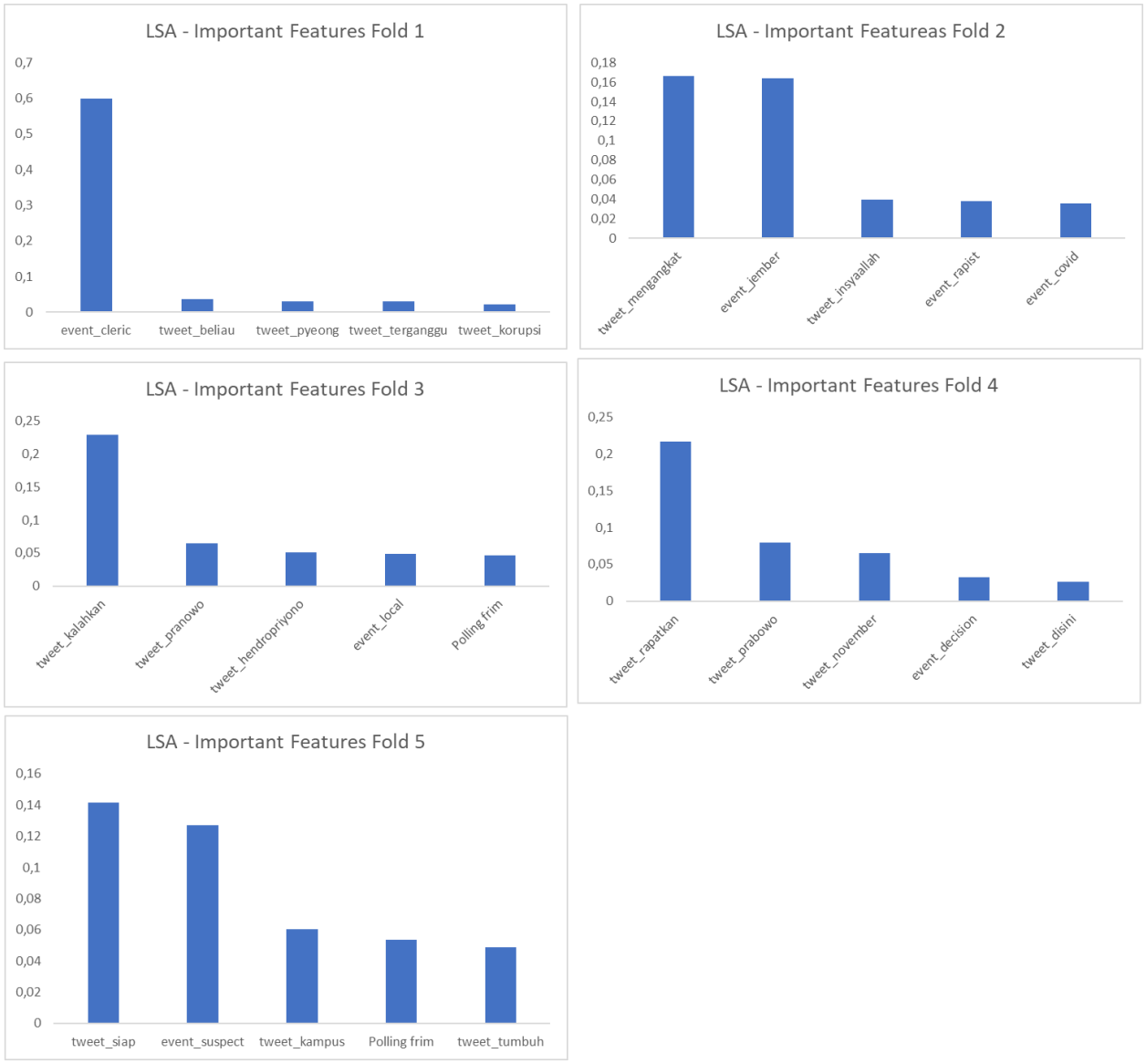

Figure 4. 3 LSA word feature importance by Fold

It became evident that key important event play an essential role in shaping the predictive performance of the models. The influence of specific events on the prediction model across the year was recognize through features such as event_cleric, event_rapat and event_jember. This observation emphasize the multifaceted nature of sentiment analysis, where not only inherent sentiments but also external factors like events crucially contribute to the accuracy and reliability of election prediction models. This also later, help to better understand on what important event contributing the trend.

5. Discussion

BERT outperform the 2 NLP with Word2vec comes the second best NLP. Even though BERT and Word2vec are word embeddings, however, as mentioned in the theory above, BERT capture longer range context, takes into account word position and having different embeddings based on context. Moreover, even though has the lowest performance, LSA, can reveal the real word behind the important features, which make it easier to analyze by time on what is happening during the time that affect the electability of the candidate.
The final results shows that the crucial factors incorporated into the model indeed contribute crucially, as evidenced by their impact on the feature importance. Notably, national events emerged as key contributors, with their presence demonstrated in the top 5 features.
Furthermore, from a model interpretation perspective, events enhance clarity in understanding, as revealed through the feature explanation. This clarity is achieved through the utilization of Latent Semantic Analysis (LSA), albeit with lower performance compared to BERT and Word2vec.
However, it is essential to acknowledge the limitations of this study. Firstly, the model does not incorporate fake news detection, posing a potential gap in its application scope. Additionally, there is a limitation regarding the noisiness of the dataset that need to be engineered in order to get a better score for several timeframe which has an outlier or temporal trend which hard to be captured. Moving forward, further exploration is warranted, particularly in delving deeper into the interpretability aspects of Word2vec and BERT so we can get a better understanding of the more advance technique.

5.1. LSA

In the context of this study, the LSA-based model exhibited a challenge in effectively capturing the complex semantic relationships present in the Twitter and Wikipedia dataset. LSA, relying on singular value decomposition to reduce dimensionality, may have struggled with the nuanced and context-dependent nature of language, particularly within the dynamic and evolving landscape of election-related discourse. The negative R-squared value observed with the LSA model implies a limited ability to extract and represent meaningful patterns within the data. The linear algebra-based approach of LSA might not have been as adaptive or nuanced as the other models in understanding the changing sentiments and language nuances present in the evolving dataset, thereby leading to its comparatively lower predictive performance.
LSA was used to get the better understanding of the prediction. For Fold 1, the most important meaningful words is cleric from “Controversial Indonesian Muslim cleric Muhammad Rizieq Shihab announced in a videotape that he will soon return to Indonesia “. This is very related to the candidate of Prabowo since he “used” Rizieq Shihab for the previous election for a propaganda which later was exiled in Saudi Arabia. There are only 2 different attitude of people in Indonesia towards Rizieq Shihab which either hate or love and getting to know that he will return back to Indonesia rising the sentiment towards Prabowo Subianto. Moreover, the tweet_pyeong was also an important features which stands for a famous Korean actor who played in a popular serial drama in Indonesia about a start up company, since Prabowo and Jokowi (current President) was a rival in previous election, however, in 2020, Jokowi appointed Prabowo as part of his cabinet which people mentioned that it was inspired by pyeong drama Later.
For Fold 2, tweet_mengangkat was the number one word extracted feature , “menagangkat” means appointed, where it is related to the appointment of vice chairman of Gerindra Party (Prabowo’s Party) who’s the grandson of Nadhaltul Ulama (well-known Islamic Organisation) founder which was seen as the representative of the Indonesian people.
For the Fold 3, tweet_kalahkan, “kalahkan” means defeat, which related to the Prabowo getting more support from the people since he got the support from the current president openly.
In Fold 4 became interesting, raptkan dan decision play an important, as he appointed Gibran as the vice president of Prabowo, which is super controversial, there is a lot of supporter who left him because of the continuous regime of the current president of Indonesia.
Lastly, in Fold 5, tweet_kampus was an important feature from the model, kampus stands for college/university, where there was a presidency candidate debate happening in one of the well known University in Indonesia which showing the real character of the President as the debate

was more informal so people can judge better.
The extracted really help us understand and verify the model in a human understanding which makes us making sense the prediction to a real world phenomenon that affecting the electability.

5.2. Word2Vec

Word2Vec has an advance ability to make sense of words even with a limited dataset as mentioned in the theory section as mentioned earlier. This flexibility suits the challenges that this study face in predicting elections, where getting a lot of labeled data is tough. So, not only did Word2Vec prove good at capturing time-related shifts, but it also revealed that it can handle small datasets well. This makes it a promising choice for similar prediction tasks when data is a bit scarce.
From the time perspective, we can see that the fold 3 and 4 gives the best performance as other fold was having a some increasing or decreasing trend that is quite hard to capture which is not happening in a near time of the election which means that a prediction can only be possibly made with a near time election event which is quite make sense in the nature of a timeseries prediction. Also if the data trend has been engineered it can give a hope to get a better prediction result.

5.3. BERT

In the context of this study, the BERT (Bidirectional Encoder Representations from Transformers) model emerged as a robust performer in predicting candidate electability percentages from Twitter and Wikipedia data. BERT's strength lies in its advanced transformer architecture, which allows it to consider the entire context of a word within a sentence, catching both preceding and succeeding words. This bidirectional understanding enables BERT to get intricate language variety and context-dependent relationships. Unlike traditional models that may struggle with word ambiguity, BERT perform in disambiguating meanings based on the surrounding context which makes it a best model out of 3. Moreover, BERT is pre-trained on vast amounts of diverse language data, giving it a strong cornerstone for understanding the complexity of election-related discourse. The positive R-squared value observed in the results suggests that BERT effectively leveraged its contextual understanding to make accurate predictions in this dynamic and evolving dataset.
Similarly, From the time perspective, we can see that the fold 3 and 4 gives the best performance as other fold was having a random trend that is quite hard to capture which is not happening in a near time of the election.

6. Conclusion

In conclusion, the study demonstrates the important contribution of incorporated factors, particularly key daily phenomenon/events to the model. Stressing how important national key events are, looking at which features matter helps us understand their crucial role. Making the model easier to understand is achieved by adding national key events, using LSA, even if it doesn't perform as well compared to other methods which later can be a promising future work for its performance improvement. In overall, key phenomenon has a contribution to the model prediction and has tremendous enhancement to understand the model by looking at the extracted word important features which also give the validity in terms of the domain knowledge validation. Furthermore, the 3 NLP method did serve a different result of the prediction in terms of the performance, making BERT as a best model. Recognizing the study's limitations, future research should address the absence of fake news incorporation, feature engineering the noisy data in few timeframe and explore additional avenues for enhancing the interpretability of Word2vec and BERT. This work serves as a foundation for future investigations into the complex dynamics of feature incorporation which contributed to the election prediction model in BERT (highest performance from 3 of the NLP model) as well as the better understanding of the interpretability of a time based model explanation on the context of a political time.

In terms of interpretability, BERT and Word2vec are not really as usable yet (future work) compared to LSA and due to its capability to extract the real word for its important features making us to understand the progress of the candidate electability. For future work, the optimization of the data for addressing the challenge posed by a noisy target variable in the early stages of the time series regression model was crucial for enhancing predictive performance. The application of smoothing techniques, such as moving averages and exponential smoothing, helped mitigate short-term fluctuations, revealing the underlying trends. These findings suggest that future research could delve deeper into these strategies to develop more robust and accurate time series regression models, especially in scenarios where noise poses a crucial challenge. In conclusion, The key national phenomenon does have a major contribution to the model not just in terms of prediction but also through the enhancement of human understanding which help us to verify the feature contribution through the help of word extraction with the real world experience, moreover, the different techniques applied (LSA, Word2vec, BERT) revealed their different capabilities in capturing the meaning of the text. This is a promising work to be done in the future which also can be applied globally for supporting the nation future development.

References

Alzahrani, E., & Jololian, L. (2021). How different text-preprocessing techniques using the bert model affect the gender profiling of authors.arXiv preprint arXiv:2109.13890.

Biau, G., & Scornet, E. (2016). A random forest guided tour. Test, 25, 197–227.

Breiman, L. (2001). Random forests. Machine learning, 45, 5–32.

Budiharto, W., & Meiliana, M. (2018). Prediction and analysis of Indonesia presidential election from twitter using sentiment analysis. Journal of Big data, 5(1), 1–10.

Cerqueira, V., Torgo, L., & Mozetiˇc, I. (2020). Evaluating time series forecasting models:An empirical study on performance estimation methods. Machine Learning, 109, 1997–2028

Darst, B. F., Malecki, K. C., & Engelman, C. D. (2018). Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. BMC genetics, 19(1), 1–6.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pretraining of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Di Gennaro, G., Buonanno, A., & Palmieri, F. A. (2021). Considerations about learning word2vec. The Journal of Supercomputing, 1–16.

Dogan, A., Birant, D., & Kut, A. (2019). Multi-target regression for quality
prediction in a mining process. 2019 4th international conference on computer science and engineering (UBMK), 639–644.

Fransiska, S., Rianto, R., & Gufroni, A. I. (2020). Sentiment analysis provider by. u on google play store reviews with tf-idf and support vector machine (svm) method. Scientific Journal of Informatics, 7(2), 203–212.

Gunn, S. R., et al. (1998). Support vector machines for classification and regression. ISIS technical report, 14(1), 5–16.

Gupta, H., & Patel, M. (2021). Method of text summarization using lsa and sentence based topic modelling with bert. 2021 international conference on artificial intelligence and smart systems (ICAIS), 511–517.

Hasan, A., Moin, S., Karim, A., & Shamshirband, S. (2018). Machine learning-based sentiment analysis for twitter accounts. Mathematical and computational applications, 23(1), 11.

Khan, A., Zhang, H., Boudjellal, N., Ahmad, A., Shang, J., Dai, L., & Hayat, B. (2021). Election prediction on twitter: A systematic mapping study. Complexity, 2021, 1–27.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. ACM computing surveys (CSUR), 50(6), 1–45.

Liddle, R. W. (2000). Indonesia in 1999: Democracy restored. Asian Survey, 40(1), 32–42.

Liu, R., Yao, X., Guo, C., & Wei, X. (2021). Can we forecast presidential election using twitter data? an integrative modelling approach. Annals of GIS, 27(1), 43–56.

Misra, P., & Yadav, A. S. (2020). Improving the classification accuracy using recursive feature elimination with cross-validation. Int. J. Emerg. Technol, 11(3), 659–665.

Palczewska, A., Palczewski, J., Robinson, R. M., & Neagu, D. (2013). Interpreting random forest models using a feature contribution method. 2013 IEEE 14th International Conference on Information Reuse Integration (IRI), 112–119. https://doi.org/10.1109/IRI.2013.6642461

Pereira, B. A. (2019). The impact of periods of crises on voting behavior in brazil [Doctoral dissertation, Ohio University].

Ray, S. (2019). A quick review of machine learning algorithms. 2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon), 35–39.

Sarica, S., & Luo, J. (2021). Stopwords in technical language processing. Plos one, 16(8), e0254937.

Satria, H. (2023). Crawl data twitter menggunakan tweet harvest - juli 2023 [Accessed on Decembee 1, 2023]. https://helmisatria.com/blog/crawl-data-twitter-menggunakan-tweet%20harvest

Sivakumar, S., Videla, L. S., Kumar, T. R., Nagaraj, J., Itnal, S., & Haritha, D.(2020). Review on word2vec word embedding neural net. 2020 international conference on smart electronics and communication (ICOSEC),282–290.

Tricahyo, V. A., & Isa, S. M. (2020). Classification of indonesian presidential campaign on

twitter using word2vec. International Journal, 9(4).
Wagire, A. A., Rathore, A., & Jain, R. (2020). Analysis and synthesis of industry 4.0 research landscape: Using latent semantic analysis approach. Journal of Manufacturing Technology Management, 31(1), 31–51.
Yang, X., Tan, L., & He, L. (2014). A robust least squares support vector machine for regression and classification with noise. Neurocomputing,140, 41–52.
Yoon, J. (2021). Forecasting of real gdp growth using machine learning models: Gradient boosting and random forest approach. Computational Economics, 57(1), 247–265.
Zhang, S., Cheng, D., Deng, Z., Zong, M., & Deng, X. (2018). A novel knn algorithm with data-driven k parameter computation. Pattern Recognition Letters, 109, 44–54.

Appendix A

Result for LSA random split with normal cross validation

	Metrics	2020	2021	2022	2023
RF	Mean Absolute Error (MAE)	0,011324444	0,027989898	0,033701333	0,05459259
	Mean Squared Error (MSE)	0,000286046	0,001057841	0,001791589	0,005292218
	R-squared	0,616973708	-0,774342083	0,298032752	0,228692685
KNN	Mean Absolute Error (MAE)	0,010641195	0,011916667	0,049276879	0,053894444
	Mean Squared Error (MSE)	0,000173116	0,000168917	0,002761726	0,005016716
	R-squared	0,768191813	0,716671908	-0,082079131	0,268845375

Appendix B

None