Journal of Informatics and Web Engineering

Vol. 5 No. 2 (June 2026) eISSN: 2821-370X

Analysis of Social Media Trends for Political Election Predictions

Wong Jing Hong[1], Lew Sook Ling[2], Tan Li Tao[3*]

[1]Faculty of Information Science and Technology, Multimedia University, Jalan Ayer Keroh Lama, 75450 Bukit Beruang, Melaka, Malaysia

[2]Centre for Advanced Analytics, CoE for Artificial Intelligence, Multimedia University, Jalan Ayer Keroh Lama, 75450 Bukit Beruang, Melaka, Malaysia

[3]Faculty of Law, University of Cambridge, The Sir David Williams Building, 10 West Road, Cambridge CB3 9DZ, United Kingdom

**corresponding author: (ltt28@cam.ac.uk; ORCiD: 0009-0004-7133-3045)*

Abstract - The traditional opinion polls are losing the capability to forecast election results because of the biasness in the sampling and the slowness in updating the polls. Although social media has a lot of real-time data, it has a considerable number of drawbacks, such as noise and demographic bias. This paper shows a new end-to-end forecasting pipeline, which is evaluated on a corpus of 1.75 million tweets related to the 2020 election in the U.S. Our approach adopts a dual-path sentiment extracted (VADER & RoBERTa) for a better accuracy and a new state level feature engineering to fix data bias. This plan transforms raw scores into 14 relative indicators, including sentiment differentials and volume ratios of tweets, which adjusts the regional activity imbalances. Average state-level prediction of a tuned Gradient Boosting Tree (GBT) classifier trained using these features was 70.6 per cent (ROC-AUC 0.69). Importantly, the cumulative prognosis was an exact duplicate of the Electoral College 306-232 majority. The feature analysis established that our relative indicators, especially the tweet volume ratio, were the strongest predictors that we engineered. The objectives in this paper are to present an alternative to traditional polling that is both powerful and easy to interpret in terms of bias reduction due to these relative characteristics. This framework has shown that it is a scalable, real-time process of political forecasting that has attained its goals of capturing the dynamics of the electoral process where conventional methods fail to do so.

Keywords—Election Forecasting, Sentiment Analysis, Social Media, Machine Learning, Gradient Boosting Trees.

Received: 16 July 2025; Accepted: 23 November 2025; Published: 16 June 2026

1. INTRODUCTION

Prediction of political elections is still a major issue in the field of politics and data science. Conventional public opinion polls have been the main tool in this job done over the decades. Their reliability, however, has been called into question increasingly, as major and systematic failures have occurred in recent high-profile elections. The vulnerability became especially evident in the 2016 and 2020 presidential elections in the United States, where the

pre-election polls largely and erroneously overestimated the popularity of the Democratic candidate [1]. These errors are associated with underlying issues in contemporary survey research, such as reducing response rates, the inability to get a genuinely representative sample of the electorate, and social-desirability biases, which may distort answers [2]. This undermining of faith in established procedures impels the search of new sources of data and new patterns of analysis.

To address these shortcomings, scholars have resorted to social media networks such as Twitter, which present an enormous and constant flow of real-time user-generated information. Such digital discussion provides a unique chance to understand the mood of society with the level of detail and in real-time, which is not possible through other means. Nevertheless, it is not so easy to use this data to make a prediction. The perennial challenge, according to Huberty [3], with using social media in forecasting is that social media users do not represent a representative sample of the voting population and the data is noisy and biased [4]. Therefore, although exciting, social media data demands an advanced technique to exclude noise and isolate useful predictive information.

The rest of the paper is structured in the following way. Section 2 provides a related work in election prediction by both traditional and social media-based approach. Section 3 provides our framework and methodology, including data preprocessing, feature engineering, etc. Section 4 provides experimental results, such as model performance as well as feature importance analysis [6]. The findings are discussed in Section 5 and the paper concluded with a summary and future work directions in Section 6.

This paper aims to build and evaluate a model of social-media-driven election prediction that does not depend on the traditional polling data. It tries to develop sophisticated state-level sentiment and activity measures and test them on machine learning models including Gradient Boosting Trees (GBTs) and Multi-Layer Perceptrons (MLPs) to determine how well they predict state-level outcomes as well as the overall U.S. Electoral College outcome.

2. LITERATURE REVIEW

2.1 Election Forecasting Approaches

Public opinion polls and econometric models of political fundamentals have long been considered the staple of traditional election forecasting. This approach is reflected in models of the U.S. election, including the one proposed by, which takes historical data, such as economic indicators and presidential approval ratings as predictive features [7]. The decreasing accuracy of the polling, which has shown severe systemic issues in 2016 and 2020 elections in the United States, has however prompted the desire to find alternative sources of data. In this paper, a framework that examines the sentiment of social media as a potential source of alternative data to be used in election forecasting, with the use of machine learning to model the sentiment of the people was proposed.

Social media has become the main alternative, and a whole new research direction is the prediction of elections based on such platforms as Twitter. Very early research tended to match the raw number of mentions of a candidate with the vote. Later studies, however, determined that sentiment is a more important element. Indicatively, suggested a more subtle measure of the positive to negative messages ratio to apply to the 2020 U.S. election [8]. Despite these developments, the profession is not an easy one. The challenges of social media data, such as demographic biases or the noise, have caused many studies to yield controversial or non-generalisable results, which makes some question the overall feasibility of its use in forecasting In this paper, a new feature engineering pipeline is offered, which aims at generating relative, not absolute, predictive signals to reduce the impact of data noise and bias in order.

2.2 Sentiment Analysis and Machine Learning Models

Sentiment analysis is the essence of forecasting on social media. One of the most frequent baselines is a comparison between lexicon-based methods and machine learning models. The tool is highly popular and lexicon-based, specifically trained on social media text (VADER). It has been demonstrated that it is a solid baseline performer in election scenarios. At the same time, the Transformer-based deep learning models, including BERT and its variations, like RoBERTa, are becoming more common because they better capture the context in complicated political language [9]. In this paper, a two-fold sentiment analysis framework is built, which is a strategically designed combination of the performance of a lexicon-based tool (VADER) and the contextual richness of a

transformer-based model (RoBERTa) that has resulted in the development of a more reliable set of sentiment indicators.

In literature, different machine learning classifiers are used in the last prediction task. Gradient boosting models, such as XGBoost, are very often selected because of their good performance with structured, tabular data generated through text analytics [10]. The work by Feng et al. which employed GBT and MLP models, provide the helpful overview of their performance as demonstrated in Table 1.

Table 1. Summary Table for GBT and MLP Model

Model	CV Accuracy (No Sentiment Analysis)	CV Accuracy (With Sentiment Analysis)	2020 Accuracy	2020 EV Prediction (DNC vs. GOP)
GBT	95.4%	95.4%	96%	279 vs. 259
MLP	91.4%	94%	96%	338 vs. 200

As it can be seen in the table, the hybrid model developed by Feng et al. provided a high accuracy (96 percent) using GBT and MLP classifiers by incorporating the traditional data and the sentiment data of social media. It is interesting to note that their results indicate that sentiment data enhanced the performance of the MLP model to a large extent [11]. This shows the possibility of applying powerful machine learning algorithms such as the GBT and MLP in this field.

The present research expands on these founded research directions using a dual-sentiment analysis model (VADER and RoBERTa) and evaluation of two different types of strong machine learning algorithms (GBT and MLP) on a new feature set, state level.

3. PROPOSED FRAMEWORK AND METHODOLOGY

3.1 Dataset Description, Scope, and Limitations

Datasets 1 - US Election 2020 Tweets [13]. The corpus has a total of about 1.72 million tweets scraped on 15 October 2020 to 8 November 2020 through the Twitter API and snsscrape. The data is released as two CSV files (hashtag_donaldtrump.csv and hashtag_joebiden.csv) with fields at the tweet level such as timestamp (created_at), tweet_id, full text, and the number of likes and retweets, as well as metadata about the tweet author (e.g., user location, verification), and language code. The following is the data that we can build our state level activity and sentiment indicators on. Preprocessing involved lower-casing and simple text cleaning (removing URLs, mentions and emojis), the elimination of retweets and duplicates, and filtering to English in cases of a language tag. Geotags were used to map tweets to states and in their absence, tweets were mapped to states by parsing location strings corresponding to states as they were reported by users; tweets that could not be assigned to a state were dropped.

Dataset 2 - 2020 Presidential Election Results by State [14]. The data (scraped off Associated Press) is the official 2020 state-level results in one file (voting.csv) containing 51 rows (50 states + District of Columbia) and 8 columns, such as the name/abbreviation of the state, and the percentage of the votes won by Donald Trump and Joe Biden, and a 0/1 column which gives the winner of the state. We treat these figures as ground truth to compare our state-level predictions to, and to use state results to calculate the national Electoral College outcome.

Scope - The research is confined to the 51 U.S. jurisdictions at the state level and 15 October to 8 November 2020. The social-media corpus is limited to tweets with candidate-related hashtags (#DonaldTrump/#Trump and #JoeBiden) and the ground-truth dataset provides results on the state (not county or district) level. Our model hence forecasts one label each state; Electoral College totals are calculated by adding those state forecasts.

Limitation - Twitter users do not represent the electorate, and hashtag-based collection can over-represent the most active or organized accounts; even after filtering, bot or campaign activity can be present. The volume of tweets is not evenly distributed among states, which adds variance to the states with a low population of tweets. The user location-based mapping is not perfect and may create errors in assignment. Sentiment Analysis of short texts is prone to sarcasm, irony and the use of domain slang. The English-only filter disallows the non-English political discourse. Finally, the ground-truth dataset is state-aggregated, and it does not reflect within-state heterogeneity and split allocations (e.g. by congressional district).

3.2 Overall Workflow

The overall mechanism of the suggested framework is a clear pipeline, which includes data acquisition to final prediction and assessment, as shown in Figure 1. Such a sequential organization of research is typical of data-driven studies on election forecasting, where the stages of data processing, feature engineering, and machine learning modelling may be rather complicated [12]. The process is started with the acquisition of two primary data sources, a huge-scale raw Twitter data concerning the 2020 U.S. presidential election (taken on Kaggle) and the official state-level election data. The Twitter data has a size of about 1.75 million tweets comprising of text, user metadata, and timestamps and, where present, geographical information.

Data preprocessing is done in detail to improve the quality and relevance of data. This involves deleting non-English tweets, deleting duplicates, and stripping out irrelevant or spam tweets and normalizing the text in the usual way by lowercasing, tokenizing, and noise removal. Such measures are essential in the reduction of the noise and bias that is inherent in social media data. As much as it is possible, tweets are geolocated at state-level in the United States using available metadata or other user data which is necessary to analyse the data at a state level.

The second step is advanced sentiment analysis where both, a rule-based sentiment analysis algorithm VADER, specifically designed to work with social media text and a transformer-based deep learning model RoBERTa, which can capture complex contextual sentiment will be used. A tweet is then run through each model to obtain the various sentiment scores (positive, negative, neutral and compound) which gives a more detailed and deeper view of the opinion of the people.

After sentiment scoring, all the tweet-level features are aggregated in a specific four-week window before the election. Aggregation is done on the state and candidate level, which allows building feature sets at the state level. New methods of feature engineering have been used, like the computation of sentiment polarity differences between candidates, ratio of tweet volumes referencing each candidate and other engagement indicators. These are engineering features that are meant to capture the intensity, as well as relative dynamics of online discourse and sentiment changes at the state level. This method of developing relative indicators is meant to address the shortcomings of just carrying out volume-based measures as observed in earlier studies.

The resulting combined feature dataset is then combined with the ground-truth election results, with the result that features, and target labels are aligned in a supervised learning context. Using the structured dataset, two classes of machine learning classifiers are trained and tested, GBT that can effectively work with structured tabular data and MLP that is a neural network model that can model nonlinear relationships. These two types of powerful classifiers are selected because they are successfully used in other similar forecasting models.

The cross-validation is used during model training to have a robust evaluation of the predictive performance with a preference to state-level accuracy, ROC AUC, and F1-score. The results at the state level are finally aggregated to model the national result in Electoral College. This allows the comparison of the results predicted by the model and the real outcome of the 2020 U.S. presidential election, which demonstrates both the usefulness and effectiveness of the suggested framework. Figure 1 shows the overall system workflow of this research.

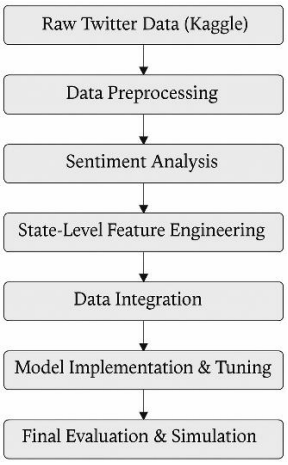

Figure 1. Overall System Workflow

3.3 Data Acquisition and Preprocessing

Two Kaggle open datasets were used. The main dataset consisted of about 1.75 million tweets concerning the 2020 presidential candidates in the United States, Donald Trump, and Joe Biden [13]. The secondary dataset, voting.csv, gave the formal election results, which were used as the ground truth target variable (winner) [14]. The preprocessing of the raw tweet text pipeline was strict. As usually done in social media analysis, a set of cleaning functions was used to remove the irrelevant noise, i.e. URLs, user mentions, and punctuation. Then the tokenisation, stop and lemmatisation were applied on the text to normalise it using Natural Language Toolkit (NLTK) library to reduce the words to their fundamental or dictionary form.

3.4 Dual-Method Sentiment Analysis

A two-fold approach was used to offer a sound sentiment analysis. A lexicon-based tool VADER (Valence Aware Dictionary and sEntiment Reasoner) was used because it is sensitive to sentiments used in social media [15]. Second, a more complex contextual nature was to be captured, so a model with RoBERTa (cardiffnlp/twitter-roberta-base-sentiment-latest) was employed [16]. Transformer-based deep learning models such as RoBERTa are now commonplace since it has been found that they are more useful in capturing context in complex political language.

3.5 State-Level Feature Engineering

The most fundamental methodological procedure in this work was to transform tweet-level data into a well-organized, forecasting state-level feature set. Tweets were initially matched to a U.S. state by parsing the unstructured user_location field and after this, the metrics were aggregated per candidate per state in the four-week period before the election.

This was done to go beyond the mere volume counts that were applied in the early forecasting studies, which are known to be highly limited. Rather than using simple aggregations but influenced by evidence that more subtle measures such as the proportion of positive to negative messages are useful, this paper constructed a full set of relative features [17]. These were difference features (e.g. vader_compound_diff = Biden score - Trump score) and ratio features (e.g. tweet_count_ratio = Biden volume / Trump volume). In the experiments, a last feature set of 14 features, so called Refined Feature Set, was built to model.

3.6 Predictive Modelling and Evaluation

The final 51x14 state-level feature matrix was fed to two other separate classifiers: GBT and MLP. GBT is a tree-based ensemble model, which was selected due to its high performance on structured, tabular data, which was confirmed by the studies on similar text analytics tasks [10]. A contrasting paradigm was chosen, an MLP, to see whether a neural network could better represent the non-linear relationships. A broad hyperparameter search was performed on each model to determine the best model configuration using GridSearchCV with 5-fold cross-validation strategy, and the search was performed to maximise accuracy. Model evaluation was done with an extensive array of metrics, comprising Accuracy, ROC AUC, Precision, Recall, and F1-Score, to provide a complete evaluation of the predictive performance.

4. EXPERIMENTS AND RESULTS

4.1 Experimental Setup

To respond to the goals mentioned, the findings indicate that the suggested dual-sentiment, state-level feature engineering strategy makes a significant difference in the accuracy of the election prediction, which is illustrated by the mean state-level accuracy of 70.6% and an impeccable prediction of the Electoral College. The importance of the engineered features is also supported by feature analysis, and the comparison demonstrates that GBT model is more effective than MLP in this task. All these findings validate the research objectives of using social media data to perform reliable forecasting of election results that is interpretable.

Each of the experiments was carried out in a Google Colaboratory environment with Python 3. Its execution depended on a few major libraries such as Pandas to process the data, NLTK to preprocess the text and scikit-learn to model machine learning. All models were fitted and tested on the final state-level dataset, composed of 51 samples (states and D.C.) and including the 14-feature "Refined Feature Set" as well as the 12-feature "Engineered Feature Set". We applied a common method of machine learning validation tasks, 5-fold cross-validation, to make sure that the model performance would be well-evaluated [11]. The GridSearchCV was the method of systematic search of the best hyperparameters of both GBT and MLP classifiers, where accuracy was the main scoring method.

4.2 Model Performance Comparison

To understand which model performed better on what feature set, a comparative analysis was conducted. Performance measures of the optimal-tuned parameters of each model across-validated are summarised in Figure 2.

Model Configuration	Mean CV Accuracy	Mean CV ROC AUC	Macro Avg F1-Score	Weighted Avg F1-Score
GBT (Refined Features, Tuned)	**0.7036**	**0.6947**	**0.7040**	**0.7050**
MLP (Engineered Features, Tuned)	0.6873	0.6707	0.6830	0.6840
MLP (Original Features, Tuned)	0.6491	0.6813	0.6470	0.6470

Figure 2. Comparative Performance of Predictive Models

As depicted in Figure 2, GBT model which was trained using the 14 refined features yielded best performance in all the four evaluation measures. It achieved the best mean cross-validated accuracy (0.7036), which implies that it got the winner at the state level right in nearly 70.4 percent of the cases. It also had the highest ROC AUC (0.6947) indicating greater discrimination between winning and losing classes, highest Macro Avg F1-score (0.7040) and Weighted Avg F1-score (0.7050).

On the other hand, the MLP whose features were engineered to be competitive performed competitively but remained below the GBT model until the end of the run whereas the MLP whose features were not tampered with performed worst of all.

The excellent results of the GBT model confirm earlier results [18] that ensemble methods based on trees tend to perform better than neural networks on structured, tabular data, especially when the number of features is not large, but well-engineered. The competitive results of MLP with engineered features demonstrate that neural architecture can still detect meaningful patterns, however, they are not as helpful when the dataset is small and not very complex.

Figures 2 are made using features calculated over the US Election 2020 Tweets dataset [13] and compared to the official results at the state level as provided by the 2020 US Presidential Election Results by State dataset [14].

4.3 Best Performing Model

Since it shows better performance, the tuned GBT model with optimized features was chosen to be analysed further. The feature importance of the model was also analysed to determine the factors that largely contribute to its predictions, which is a direct answer to one of the main project goals. Figure 3 visualises the relative significance of the top 10 features.

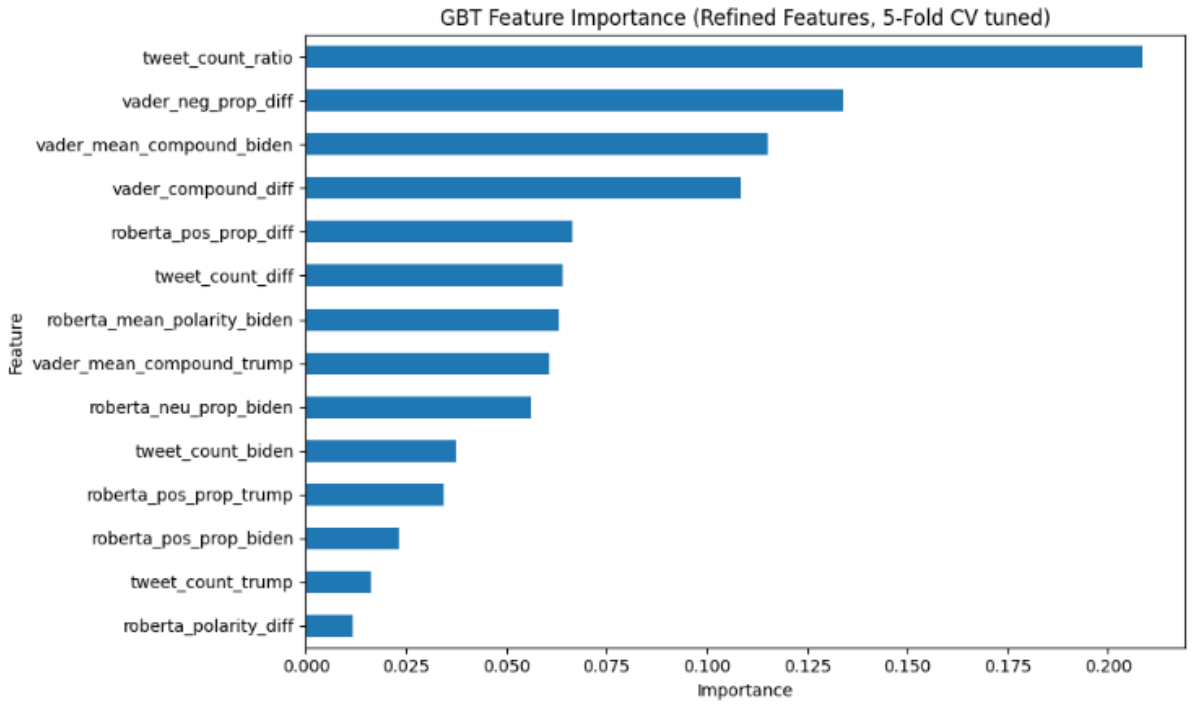

Figure 3. Top 10 Feature Importance from Best GBT Model

The rankings of feature importance of the tuned GBT model (Figure 3) show that the relative volume of tweets containing the names of the two candidates, tweet_count_ratio was the most predictive variable by far, and all other variables paled in comparison. The excellence of this metric is that relative indicators of online activity are more predictive than the absolute number of tweets, which can be exaggerated by a concerted action or by bots.

Next, several difference-based sentiment measures (vader_neg_prop_diff, vader_mean_compound_biden, vader_compound_diff) proved very influential. The overall implication of these findings is that comparative sentiment dynamics (i.e. the extent to which one candidate is mentioned in a more negative or positive way than the other) are stronger predictors of electoral outcomes than standalone sentiment measures of each candidate.

The values of features presented in Figure 3 are obtained based on the US Election 2020 Tweets dataset [13], and their predictive accuracy is verified with reference to the official results of the 2020 US Presidential Election Results by State dataset [14].

To determine the usefulness of the model in practice, the model was applied in predicting the winner in each of the 51 states/districts. These predictions at the state level were further summarized according to the electoral college system in the U.S. The simulation result was a very close national outcome as indicated in Figure 4.

Candidate	GBT Model Prediction	Actual 2020 Result
Joe Biden	306	306
Donald Trump	232	232

Figure 4. GBT Model Simulation Vs. Actual 2020 Election Result

4.4 Comparison with Previous Work

The benefit of the ratio-, and difference-based features is consistent with previous study [3], which advised against using raw volume-based predictors as the sole predictor in social media forecasting. Moreover, the overall superiority of GBT over MLP is in line with findings [18] that gradient boosting has a high baseline performance in classification tasks with modest and high-quality feature sets.

4.5 Limitations and Practical Implications

Despite the positive results, the findings can only be used in the context of the case of the 2020 presidential elections in the United States and cannot be directly transferred to other elections. The ranking of the feature importance might be biased to more active states regarding the online community. In addition, the ground truth dataset resolution at the state level constrains the performance evaluation since it does not represent an intra-state variation.

Concretely, the results indicate that relative and comparative characteristics rather than absolute ones should be given primary consideration in future social media-based forecasting models, and that gradient boosting techniques can be a competent option in case of well-engineered tabular data.

5. DISCUSSION

5.0 Contributions of This Paper

In this discussion, the findings of the experiment conducted in Chapter 4 are interpreted in the light of the previous literature, methodology, and the overall implications of the social media-based forecasting of elections. The current article deals with not only the common inaccuracy of conventional public opinion polls, but also with the inherent complexity of trying to derive high-quality signals out of massive social media data [5]. To address these issues, we construct and test an end-to-end pipeline of analysis that will process noisy Twitter sentiment and convert it into actionable state-level features that can be used to predict elections. The main contribution of the work is the combination of two sentiment analysis techniques which are used to extract more and more precise sentiment indicators based on the data: the lexicon-based method of VADER and the deep contextual model RoBERTa. Having said that, we introduce a new feature engineering approach that goes beyond the straightforward sentiment counts and produces relative quantities including sentiment polarity differences and candidate-specific ratios of tweet volume per each of the states. Through these optimized features, our optimized GBT algorithm hits a cross-validated accuracy of around 70.6 percent on state-level forecasts, and, most importantly, exactly replicates the national electoral college score (306 232) in the 2020 U.S. presidential race as well. Overall, this study shows that further processing and engineering of social media sentiment can serve as a very strong support signal in election forecasting, which can be an effective addition to the traditional polling approach.

5.1 Summary of Key Findings

The experimental outcomes of the paper substantiate the effectiveness of the suggested forecasting framework in a research area where most of the studies have been discovered to have controversial or non-generalisable outcomes. Using the optimized model of 14 state-level features, GBT model with tuning was the best-performing model with a mean 5-fold cross-validated accuracy of about 70.4%. The most dramatic result, the state-level predictions of the model were exactly matched by the ultimate national 306-232 Electoral College result of the 2020 U.S. election. Moreover, the feature importance analysis showed the relative numbers, especially the ratio between the volumes of tweets of candidates (tweet_count_ratio), were the most important predictors of the state-level outcomes. All these findings support the main goals of the study and ensure a solid empirical rationale of the contributions to the field of social media-based forecasting of election outcomes, which will be discussed further in the paper.

5.2 Contribution 1: A Robust Forecasting Pipeline Independent of Traditional Data

The main contribution of this research is the creation and confirmation of a powerful, state-level election forecasting pipeline that can run without depending on the conventional polling data. Over the past few election cycles, the accuracy of traditional polling has been called into serious question as there has been a high degree of failure in forecasting the results, this being most often an underestimation of the backing of a particular candidate. Such failures are explained by the problems such as the nonresponse bias and social-desirability effects, which provide a reason to look at alternative sources of data in this paper, this implies that a new model of forecasting should not only interpret the sentiment but also include the ways to address these issues with data specifically.

Our model provides such alternative. In contrast to such models as the one suggested by Sinha et al., which are based on historical and economic fundamentals, our model is designed to reflect the real-time pulse of public opinion as it is directly based on the analysis of unstructured social media data. More to the point, it is opposed to such hybrid approaches as the one by Feng et al., which attained high accuracy by integrating sentiment analysis with classic data sources such as census data and polling averages. Effective as it is, however, such hybrid approaches remain dependent on the very polling data that is turning out to be fallible. The contribution brought forward by our work is substantial as it demonstrates how even a pipeline based exclusively on the engineered features of social media can generate forecasts of great accuracy at the national level, thus providing a means of prediction that does not require the conventional data feeds. The validity of this approach can be verified by the performance of GBT model that attained an accuracy of about 70.4 percent with cross validation of the state-level prediction.

5.3 Contribution 2: Advanced Feature Engineering with Relative Metrics

The second important contribution of this study is that the feature engineering process is more advanced than the early social media-based prediction models that were merely limited to counting the volumes and sentiments. Although in the earlier works the correlation between the raw count of mentions and the electoral success was usually sought, our results show that the relative, comparative measures are much more predictive. This is clearly confirmed by our feature importance analysis (see Figure 3), where tweet_count_ratio (the ratio between the volumes of tweets between the candidates) corresponds to the single most significant feature in the most accurate GBT model.

Moreover, other features that are based on the difference, including vader_neg_prop_diff (a difference in the proportion of negative sentiment) are also among the most important predictors. This method is an extension of, but a highly expanded version of more recent methods, like the positive-to-negative message ratio proposed by Yavari et al., as it proposes a whole package of 14 engineered indicators. The advantage of these relative measures is that they automatically control the enormous differences in the number of online activities on behalf of various states and various candidates. This is a direct answer to one of the so-called perennial challenges of the discipline, namely the demographic and geographic bias of raw social media data that renders it an unrepresentative sample of the electorate. The relative dynamics of online discussion reduce noise and make our procedure more effective in deriving a meaningful predictive signal by examining large amounts of rather noisy data.

5.4 *Contribution 3: A Successful Proof-of-Concept for National Electoral Simulation*

Lastly, this experiment is an effective demonstration of the viability of the concept of using state-level election predictions based only on social media data to simulate the national-level result. This is one of the key benchmarks, because the correct prediction of the Electoral College as opposed to the national popular vote is the key measure in the study of presidential elections in the U.S. [2]. Although the state-level accuracy of the model of about 70.4% is consistent with the difficulty and noise in the prediction of individual state races, the fact that the model perfectly replicates the 306-232 Electoral College result (see Figure 4) is a very significant result.

This result implies that the model might not perform well in predicting the individual states but the features designed and the GBT algorithm were very efficient in reflecting the macro-level political picture and the general distribution of voter sentiment at the national scale. It shows that the state-level errors were not systematically biased so as to distort the national assessment; but instead, they tended to balance each other out, resulting in a properly balanced

national forecast. This is a perfect example of national simulation that was done without any conventional data inputting, and it shows the potential of social media analytics as a means not only of sentiment measurement, but also of a high-level electoral forecasting framework.

6. CONCLUSION

This paper has been able to design, apply, and validate an end-to-end forecasting pipeline to utilize social media data to forecast an election at the state level. We mainly make three contributions. First, we showed that a model based on the signals in social media alone, without using the traditional polls, can effectively reproduce the national election result. Second, we proposed a new mechanism of feature engineering, and showed that comparative measures, that reveal interactions between candidates, e.g., the ratio of tweet volume and the difference in sentiment, are a more predictive variable than the absolute measures. Lastly, the flawless replication of the 2020 U.S. Electoral College outcome in the model is an effective demonstration of concept of the framework feasibility. Although these findings are encouraging, generalizability of the framework is limited by the limitations of social media data, such as the small size of the sample of states (N=51) or the possibility of demographic bias of Twitter users. Researchers should thus work on how to reduce such shortcomings in the future. The best way forward would be to create a hybrid model that would combine our designed social media aspects with classic data points, including state-wide polls and economic indicators, that may yield stronger and more accurate projections. Moreover, the model would be more generalizable in case the dataset would be extended to cover more election cycles [19] and data provided by other social media platforms [20]. These will open pathways towards a stronger and more encompassing forecasting system on elections.

ACKNOWLEDGEMENT

The authors would like to acknowledge Faculty of Information Science & Technology and Multimedia University in giving the resources and academic atmosphere needed to conduct this study.

FUNDING STATEMENT

The authors received no funding from any party for the research and publication of this.

AUTHOR CONTRIBUTIONS

Wong Jing Hong: Conceptualisation, Data Curation, Methodology, Software, Validation, Writing – Original Draft;
Lew Sook Ling: Conceptualisation, Methodology, Review – Original Draft;
Tan Li Tao: Conceptualisation, Review – Original Draft.

CONFLICT OF INTERESTS

No conflict of interest was disclosed.

ETHICS STATEMENTS

All this research is based on a publicly accessible Kaggle dataset that includes tweet IDs and little other metadata associated with the 2020 U.S. presidential election, and publicly released, state-level election-result data. The Kaggle dataset was initially created based on the public API of Twitter according to the Developer Policy of the platform. As per the Terms of Service of Twitter, and the Kaggle data-license restrictions, we:

- Downloaded tweets in the form of IDs only to analyse them and not to distribute the entire text or information containing any personal identifiable information (PII).

- Pre-processed by removing the user handles, URLs, and other direct identifiers, because the step ensured that all subsequent analyses were conducted using text that was fully anonymized.
- Always carried out aggregate, state level only aggregate, state level analyses were carried out- no attempt was made to profile or re-identify individual users.

Since the study represents secondary analysis of publicly available, anonymized data in social media and no direct contact with human subjects, it did not need the formal approval of an Institutional Review Board (IRB) according to the current regulations. The research follows the principles of the Committee on Publication Ethics (COPE) and all the applicable data-protection laws in the process of data acquisition, processing, analysis, and dissemination.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] A. Gelman, "Failure and Success in Political Polling and Election Forecasting", *Statistics and Public Policy*, vol. 8, no. 1, pp. 67–72, Jan. 2021, doi: 10.1080/2330443X.2021.1971126.

[2] D. S. Hillygus, "The Evolution of Election Polling in the United States', *Public Opinion Quarterly*, vol. 75, no. 5, pp. 962–981, Dec. 2011, doi: 10.1093/poq/nfr054.

[3] M. Huberty, "Can we vote with our tweet? On the perennial difficulty of election forecasting with social media', *International Journal of Forecasting*, vol. 31, no. 3, pp. 992–1007, Jul. 2015, doi: 10.1016/j.ijforecast.2014.08.005.

[4] K. D. S. Brito, R. L. C. S. Filho, and P. J. L. Adeodato, "A Systematic Review of Predicting Elections Based on Social Media Data: Research Challenges and Future Directions", *IEEE Transactions on Computational Social Systems*, vol. 8, no. 4, pp. 819–843, Aug. 2021, doi: 10.1109/TCSS.2021.3063660.

[5] P. Chauhan, N. Sharma, and G. Sikka, "The emergence of social media data and sentiment analysis in election prediction", *J Ambient Intell Human Comput*, vol. 12, no. 2, pp. 2601–2627, Feb. 2021, doi: 10.1007/s12652-020-02423-y.

[6] D. Rousidis, P. Koukaras, and C. Tjortjis, "Social media prediction: a literature review", *Multimed Tools Appl*, vol. 79, no. 9–10, pp. 6279–6311, Mar. 2020, doi: 10.1007/s11042-019-08291-9.

[7] P. Sinha, A. Verma, P. Shah, J. Singh, and U. Panwar, "prediction for the 2020 united states presidential election using linear regression model", 2020.

[8] A. Yavari, H. Hassanpour, B. Rahimpour Cami, and M. Mahdavi, "Election Prediction Based on Sentiment Analysis using Twitter Data", *International Journal of Engineering*, vol. 35, no. 2, pp. 372–379, Feb. 2022, doi: 10.5829/ije.2022.35.02b.13.

[9] A. Khan, H. Zhang, N. Boudjellal, A. Ahmad, and M. Khan, "Improving Sentiment Analysis in Election-Based Conversations on Twitter with ElecBERT Language Model", *Computers, Materials & Continua*, vol. 76, no. 3, pp. 3345–3361, 2023, doi: 10.32604/cmc.2023.041520.

[10] K. Afifah, I. N. Yulita, and I. Sarathan, "Sentiment Analysis on Telemedicine App Reviews using XGBoost Classifier", in 2021 *International Conference on Artificial Intelligence and Big Data Analytics*, Bandung, Indonesia: IEEE, Oct. 2021, pp. 22–27, doi: 10.1109/ICAIBDA53487.2021.9689762.

[11] G. Feng, H. Cai, K. Chen, and Z. Li, "A Hybrid Method of Sentiment Analysis and Machine Learning Algorithm for the U.S. Presidential Election Forecasting", Dec. 09, 2023, arXiv: arXiv:2312.05584, doi: 10.48550/arXiv.2312.05584.

[12] Y. Mejova, "Sentiment Analysis: An Overview", *Comprehensive Exam Paper*, University of Iowa.

[13] C. Hui, "US Election 2020 Tweets," Kaggle. Accessed: Jul. 15, 2025. [Online]. Available: https://www.kaggle.com/datasets/manchunhui/us-election-2020-tweets

[14] C. Macpherson, "2020 US Presidential Election Results by State," Kaggle. Accessed: Jul. 15, 2025. [Online]. Available: https://www.kaggle.com/datasets/callummacpherson14/2020-us-presidential-election-results-by-state

[15] C. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text", *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, no. 1, Art. no. 1, May 2014, doi: 10.1609/icwsm.v8i1.14550.

[16] N. C. Dang, M. N. Moreno-García, and F. De La Prieta, "Sentiment Analysis Based on Deep Learning: A Comparative Study", *Electronics*, vol. 9, no. 3, p. 483, Mar. 2020, doi: 10.3390/electronics9030483.

[17] W. El-Hajj and H. Hajj, "An optimal approach for text feature selection", *Computer Speech & Language*, vol. 74, p. 101364, Jul. 2022, doi: 10.1016/j.csl.2022.101364.

[18] P. Mishra, D. Punia, G. Sikka, A. K. Sharma, and K. Sikka, "Evaluating Various Techniques for Twitter Sentiment Analysis for Election Results", in *2024 First International Conference on Pioneering Developments in Computer Science & Digital Technologies (IC2SDT)*, Delhi, India: IEEE, Aug. 2024, pp. 52–57. doi: 10.1109/IC2SDT62152.2024.10696204.

[19] M. Tabany and M. Gueffal, "Sentiment Analysis and Fake Amazon Reviews Classification Using SVM Supervised Machine Learning Model", Jou*rnal of Advances in Information Technology*, vol. 15, no. 1, pp. 49–58, 2024, doi: 10.12720/jait.15.1.49-58.

[20] Z. Zhou, M. Serafino, L. Cohan, G. Caldarelli, and H. A. Makse, "Why polls fail to predict elections", *J Big Data*, vol. 8, no. 1, p. 137, Dec. 2021, doi: 10.1186/s40537-021-00525-8.

BIOGRAPHIES OF AUTHORS

Wong Jing Hong received his Bachelor of Information Technology (Hons.) in Artificial Intelligence from Multimedia University, Melaka, Malaysia. During his first Final-Year Project, he worked in a team to design and develop an end-to-end food delivery system comprising three fully functional modules (admin, customer, and rider), which included order management, real-time tracking, and simple route optimization. This experience reinforced his practical knowledge of Python, MySQL, and RESTful API development, and inspired his interest in broader applied machine learning solutions. His current research focuses on natural language processing and sentiment analysis for data-driven election forecasting, while he continues to develop his skills in building scalable, user-centered software systems. He can be contacted at email: 1211203886@student.mmu.edu.my.

Lew Sook Ling is an Associate Professor at the Faculty of Information Science and Technology, Multimedia University (MMU), Malaysia. She has been with MMU since 2001 and received her Ph.D. in 2013. She is also a Senior Member of IEEE. Her current research interests include educational technology, business analytics, image processing, and machine learning. She actively contributes to academic development through publications and collaborative research in emerging technologies. She can be contacted at email: sllew@mmu.edu.my.

Tan Li Tao is pursuing a Bachelor of Arts (Honours) in Law at the University of Cambridge, United Kingdom. He is interested in examining the interaction of information technology and the social sciences. He declares that this work was conducted independently while enrolled at the University of Cambridge, and the views expressed are his own. He can be contacted at email: ltt28@cam.ac.uk.