AI-based Twitter Framework for Assessing the Involvement of Government Schemes in Electoral Campaigns

Aarzoo Dhiman and Durga Toshniwal

1. Aarzoo Dhiman

aarzoo@cs.iitr.ac.in

Department of Computer Science and Engineering

Indian Institute of Technology Roorkee, Uttarakhand-247667, India

2. Durga Toshniwal

durga.toshniwal@iitr.ac.in

Department of Computer Science and Engineering

Indian Institute of Technology Roorkee, Uttarakhand-247667, India

Corresponding Author

Aarzoo Dhiman

Research Scholar

Department of Computer Science and Engineering

Indian Institute of Technology Roorkee, Uttarakhand-247667, India

Email: aarzoo@cs.iitr.ac.in

Phone: +91-7876145644

Abstract

The government schemes (also known as programs and plans) or *social welfare policies* can be defined as the set of assistance and aids provided by the country's governance body. These schemes focus on the improved well-being of needful citizens. Some researchers have shown that introducing such policies and schemes has had an electoral impact in democratic countries. These earlier studies relied upon the post-poll and public survey data to reach conclusions. However, this data source has limitations and has to be collected manually, which makes it time-consuming and costly. The readily available internet inculcates the sharing of opinions freely on social media, facilitating government-citizen interactions. These interactions may show fluctuations in frequency and intensity on social media with the success and failure of some government schemes. Thus, this research proposes utilizing the Twitter data related to the government welfare schemes during the election duration to uncover the spatial and temporal relationships between the tweets' information diffusion pattern and political elections. To start with, we perform tweet classification to identify the target communities or groups and multiple user-engagements by employing deep learning-based pre-trained language representation (LR) models. The scarcity of labeled data limits the application of the supervised classification models on real-time data. Thus, we propose Mod-EDA, a text augmentation method to upscale the labeled data for reduced overfitting. Going further, we propose two modules, where the classified tweets are studied to investigate the scheme tweets' information diffusion pattern in correspondence to the election duration in terms of *the voting phase* and *the electing parties*, respectively. The proposed framework is evaluated for a case study of the 2019 Indian general elections. This study depicts that the voting phases and election duration trigger high government schemes related tweet generation. However, it is not affected by the location of the voting phase. The generation of complaints and negative tweets in one voting phase is covered with the positive news in subsequent voting phases. It is also seen that there is a strong influence of the ruling party on the scheme-related Twitter data generation.

Keywords: Tweet Classification, Tweet Clustering, Sentiment Analysis, Text Augmentation, Geo-location Analysis

2010 MSC: 00-01, 99-00

1. Introduction

The availability of affordable and high-speed Information Communication Technologies (ICTs) has led to the emergence of dense digital/virtual networks for information dissemination. This enables the stakeholders belonging to various diverse domains such as healthcare, business, education, government, and agriculture to connect, communicate, and collaborate for decision-making. Over the years, these social media platforms have facilitated the government-citizen interactions, leading to the rise of e-governance (Dwivedi et al. (2017), Alryalat et al. (2017), Agostino (2013), Rana et al. (2015)). The combination of social media and e-governance ensures the tangible benefits to both sides **(i)** the government, by providing the visibility, increased donorship (JACKSON (2017), Lutz (2009)) and trust (Bertot et al. (2012)), and **(ii)** the citizens (government's stakeholders) by providing transparency, enhanced public participation, and upgraded quality of service (Singh et al. (2020b)). While the government is utilizing social media to promote their achievements and engage in discussions with the citizens, the citizens are trying to express their opinions and perceptions on social media platforms to make the government accountable (Driss et al. (2019)).

Government-sponsored schemes and campaigns are a primary means to protect and promote the country's citizens' social, economic, and personal development. Depending on the country, these schemes may include programs focussing on employment generation, poverty alleviation, supporting loans & scholarships, providing education, housing & healthcare facilities, etc. Many government entities own social media accounts that share information and other news regarding these schemes and policies. In return, the citizens can further share this content directly or provide their views, perceptions, feedback, and experiences regarding these schemes.

As the government schemes, campaigns, and policies can directly affect the citizens' social, economic, and health status, the success and failure of these schemes can become the agendas for the competing electoral parties during elections. Different studies of election post-poll surveys imply that the generation of social policies had an impact on the electoral decisions of the citizens (Deshpande et al. (2019), De Koster et al. (2013)). However, these studies suffer from limited data availability. Furthermore, the manual data collection makes it a laborious and expensive task.

This research work proposes utilizing social media data, i.e., Twitter data, which has real-time availability, faster access, and low labor cost. Over the years social media data proved to be beneficial to perform various government-citizen centric tasks including the study of government-citizen interactions (Bertot et al. (2010)), e-governance (Rana et al. (2015)), predicting election results (Tumasjan et al. (2010)), and analyzing the political discourse (Hua et al. (2020)) etc. There is a limited study of the social media data to investigate the impact of election campaigns on government schemes and related information diffusion patterns. Twitter data is well suited for such socio-economic, behavioral analysis-based studies of different topics because of the ease of availability, short and context bounded messages and huge amount (Weller et al. (2014)). Twitter data has helped to provide deep insights on various topics such as studying the social-economic behavior (Huang & Wong (2016)), linguistic characteristics (Preotiuc-Pietro et al. (2016)), traffic analysis (Li et al. (2019)), disaster events (Li et al. (2018)) and public and mental health study (De Choudhury et al. (2013)), etc. These analyses help to capture the type of information shared and the groups of society influenced by them.

In this research work, we study the Twitter data (approx. 1.4 million tweets) from December 2018 to

November 2019 pertaining to various schemes launched by the Indian government as the case study. Though the Twitter user-base comprises only $\approx$ 17 million users[1] while having the third-largest user-population all over the world, approx. 18% of the Indian population relies on Twitter for news (Aneez et al. (2019)). These schemes may be oriented towards various communities and societies, e.g., women empowerment, maternal and child health, poverty, etc. It is crucial to identify the target audience and their opinion towards the topic to perform further analysis. Thus, a classification module has been proposed that classifies tweets based on target groups and types of engagements. This has been performed with the help of contextual deep learning-based pre-trained language representation (LR) models such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al. (2018)), Embeddings from Language Models (ElMo) (Peters et al. (2018)) and Universal Sentence Encode (USE) (Cer et al. (2018)) etc. However, the real-time data collection has the limitation of the scarcity of labeled data, further limiting the application of the above-described models. Thus, to overcome this limitation, we have further proposed and employed the text augmentation model Mod-EDA (Dhiman & Toshniwal (2020)) to increase the size of the labeled dataset. Earlier, text augmentation models did not consider the contextual relationship among tweets to perform the text perturbations. The proposed approach Mod-EDA handles this by fine-tuning GloVe (Pennington et al. (2014)) generated domain-specific word embeddings to perform class-wise insertions and substitutions in the text. The proposed Mod-EDA has an average improvement of 1.76% from the baseline model EDA (maximum 2.03% with BERT) and 1.82% (maximum 2.98% with BERT) from data without augmentation in testing F1-score values on the benchmark dataset.

In the second step, the government scheme data has been analyzed in correspondence to the duration of an election. Due to the vast size, networks, user groups and communities, etc., simply collecting and describing queries may not be adequate to collect all insights. With the help of various text analysis models, interpreting and analyzing the perceptions of various stakeholders can be made more intuitive. This step has primarily focused on two critical aspects, which can affect the concentration and frequency of the tweets, i.e., *the voting phase* and *the electing party*. We have proposed two analytical modules to perform these two steps. First, the relationship between the government scheme tweets and the voting phases of the election has been assessed. We have employed contextual word embedding-based clustering to identify an event in the subset of data collected near the voting phase along with their geolocation analysis. This can help understand whether the voting date and voting location impact the tweeting concentration or pattern of the government schemes. Second, correlation between the government scheme tweets and the election outcomes has been calculated. This can help us understand the ruling/competing parties' role in the *type of information* dissemination. Therefore, a better understanding of government schemes' role in the election campaign can be assessed using Twitter data. This study depicts that the government schemes related

[1] https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/

tweet generation increases with the voting phases and election duration. However, it is not affected by the location of the voting phase. The generation of complaints and negative tweets in one voting phase is covered with the positive news in subsequent voting phases. It is also seen that there is a strong influence of the ruling party on the scheme-related Twitter data generation. The proposed framework studies the popular government schemes in terms of the generic themes, such as target audience, type of engagements, voting phases and the electing parties. These themes make it possible for the model to apply to case-study of other democratic countries.

2. Related Work

The abundance, faster availability, and ease of access to the social media data provides excellent opportunities for solving many data-driven problems. In recent years, social media has shown success in uncovering many linguistic, social, informational, and behavioral patterns of the public in different scenarios. Social media has been used to solve many data-driven real-time problems such as disaster response and management (Kim et al. (2018), Imran et al. (2020)), public health maintenance (Jiang & Yang (2017)), mental health characterization (Gkotsis et al. (2017)), controversy detection (Garimella et al. (2018)), crime trend prediction (Aghababaei & Makrehchi (2018)), and election campaign manipulations (Jungherr (2016), Ferrara et al. (2020), Chauhan et al. (2021)), etc. The data mining based analytics techniques have facilitated these tasks. These methods may use different text mining-based machine learning, transfer learning and deep learning tools such as social network analysis (Li et al. (2018)), Latent Dirichlet Allocation (LDA) (Blei et al. (2003)) based content analysis (Kim et al. (2018)), time-series analysis (McCleary et al. (1980)), Long Short-Term Memory (LSTM) (Hochreiter & Schmidhuber (1997)) based models such as Embeddings from Language Models (ELMo) (Peters et al. (2018)), and Word2Vec (Mikolov et al. (2013)) based models such as FastText (`https://fasttext.cc/`) etc.

2.1. Social Media in E-governance and Policy Making

Electronic-governance (E-governance) is the application of Information Communication Technologies and internet infrastructures to improve government services, information-exchange and government-citizen interactions (Garson (2006), Addo & Senyo (2021)). Here, the stakeholders can be the citizens, businesses, or government ministries. The two primary benefits of social media usage in governance are increased government-citizen interactions and transparency (Bertot et al. (2010)). It is believed that citizens' involvement via social media interactions can improve the process of government policy and decision-making. Furthermore, (Criado et al. (2013)) suggests that even though the goals of using social media for governance include cost-cutting, efficient management, and better employee performance & satisfaction, the primary application of social media is innovating the government-citizen interaction measures. Bekkers, Edwards, and De Kool (Bekkers et al. (2013)) concluded in their study that social media is a tool to monitor the public reactions for policy-making and external communications. Social media can also act as the source of collecting ideas, opinions, and ideologies as a tool for policy-makers (Bonsón et al. (2015), Stamati et al. (2015), Marzouki et al. (2021)).

Some research works show that increased social media activity from the government does not necessarily mean a higher engagement (Bonsón et al. (2017)). The increased social media activity may occur due to a large amount of promotional and informational posts. Such posts tend to increase transparency, leading to trust in the government. The researchers have proposed many analytical frameworks to facilitate government-citizen interactions to perform improved policy-making. In (Charalabidis et al. (2014)), Charalabidis et al. proposed the 'passive crowdsourcing' based framework to assess citizens' perspectives, opinions, issues, and arguments for the different government activities. They used sentiment analysis and argument summarization for qualitative information aggregation. In (Driss et al. (2019)), Drissa, Mellouli, and Trabelsi proposed a framework to facilitate government-level policy-making by using the data from Facebook. As these platforms are rich in text data, the authors used text semantic analysis to extract information and knowledge from Facebook posts. In 2020, (Singh et al. (2020b)) Singh et al. utilized the Twitter data in a cloud computing framework to study the effect of public participation in the deployment of Goods and Services Tax (GST) proposed by the Indian government in 2017. In 2021, Marzouki et al. (Marzouki et al. (2021)) measured the Twitter users' engagement on the sustainable development goals to identify the trends and evolution of topics for improved policy-making.

These research indicate that the government's policy-makers provide policy-related information and ideas to the citizens on various social media platforms. On the other hand, the citizens provide their opinions and views on the same. Thus, social media can act as a platform to compare and contrast the citizens' and government officials' generated information to gain better insights.

2.2. Twitter and Election Campaign

Many studies focus on analyzing the information diffusion patterns, public perceptions and user roles during the election period to understand the role of social media in election campaigning. This analysis further helps in developing models that aid improved election outcome predictions (Jungherr (2016)). Twitter has been recognized as the platform for acculturation of ideologies (Grover et al. (2019)) and the political deliberation (Tumasjan et al. (2010)). According to Monte Lutz (Lutz (2009)), a report published by Edelman, an American public relations and marketing consultancy firm, the primary reason behind the win of Barack Obama during the 2008 US presidential elections was the powerful engagement with the common public using social networks. Since the success of Barack Obama's *online campaigning* during 2008 US presidential elections (Lutz (2009)), many political leaders have followed this strategy as a tool to win elections (Plasser & Plasser (2002)). This had also inspired Indian Prime Minister (PM), Narendra Modi, during the 2014

Indian general elections (Kamat (2014)). Stier et al. (Stier et al. (2018)) showed that different social media platforms such as Facebook and Twitter are used for different types of target audiences and purposes. Many studies (Pal & Panda (2019), Rao (2019)) have shown that social media played a crucial role in the win of PM Narendra Modi's lead party BJP. In this research work, the Twitter data of government schemes have been used; thus, a literature review of the study of election campaigns using Twitter has been provided.

In 2015, Kagan et al. (Kagan et al. (2015)) used Twitter sentiment classification to predict the 2013 Pakistan election and 2014 Indian general elections. Later in 2016, Ahmed et al. (Ahmed et al. (2016)) used the Twitter data of the 2014 Indian elections to analyze the topical, functional, and interaction strategies of campaigning of different parties. This included investigating Twitter usage for campaigning, parties' target user groups and types of information shared by the different parties. Later in 2017, (Ahmed et al. (2017)) the authors extended their work and analyzed the Twitter usage patterns of the political parties during the election period. Authors divided parties into minor & major and investigated the media attention received and given to them. The authors proposed that social media removes resource inequality among the parties. Later, Badawy et al. (Badawy et al. (2018)) investigated the hostile actors on social media discussions that attempt to manipulate public opinions. They studied the extent and effects of Russian interference in the US congress election campaign. They performed text analysis to show that most of the trolled were promoting conservative causes. In 2019, Grover et al. (Grover et al. (2019)) used the Twitter data of the 2016 US presidential elections to investigate the role of the nature of social media discussions in the acculturation of people's ideologies and polarization of voting preferences of people. In 2020, Nelimarkka et. al. (Nelimarkka et al. (2020)) compared the candidate-constituent interactions on Facebook and Twitter for 2015 Finnish parliamentary elections. They concluded that Twitter was primarily used for sharing and seeking information and opinions, while Facebook was used for more formal campaigning.

The literature indicates that social media is an intermediary platform for citizen-government interactions. It is believed that this platform is used to spread information which government officials can use for policy-making. The Twitter platform has shown a significant involvement in the election campaigning process. Many countries and political parties have spent millions on their election campaigning on social media. As Twitter can be recognized as a common platform for sharing information among government officials and citizens, it will be interesting to investigate how the electing parties utilize the government schemes and policy-related tweets during their election campaigning process. The proposed research provides a text-analysis framework to analyze the spread of government schemes and policy-related tweets during the election period. This will help us understand the impact of election campaigning on Twitter usage for government scheme-related information.

The primary task of the proposed research work is to perform the tweet classification to label the tweets in terms of their target audience and the type of user engagements. This utilizes the proposed text augmentation module helpful in upscaling the labeled data. The following section provides a brief introduction to the text augmentation literature.

2.3. Text Augmentation

Data augmentation is the process of performing some perturbations in the data sample, such that its label remains meaningful while introducing some randomness in the data sample. The randomness helps in generating a new data sample with some known label, thereby increasing the size of the labeled data. Data augmentation is supposed to improve the classification and prediction accuracy of the models, commonly used in image processing and speech recognition tasks. The process of data augmentation for natural language processing tasks is called *text augmentation*. Text augmentation deals with increasing the size of the annotated text data such that the distribution and semantic meaning of the sentence remains intact. Shorten et al. (Shorten et al. (2021)) have provided a survey on the text augmentation approaches.

Primarily, text augmentation approaches can be divided into two types. First, the models that use deep learning models based on sentence generations (Kobayashi (2018)), Generative Adversarial Networks (GAN) (Hu et al. (2017)) and language translations (Zhao et al.). These models try to generate the sentences automatically (Jia & Liang (2017)) using predictive models for noise generation (Xie et al. (2017)) and synonym replacements (Kobayashi (2018)). Liu et.al. (Liu et al. (2020)) utilized reinforcement learning to perform text augmentation. However, these models suffer from heavy hardware requirements, which are time-consuming because of the underlying powerful deep learning and machine learning models. Second, the models use rule-based approaches to perform simple word or phrase-level perturbations in the text data. Easy Data Augmentation (EDA) (Wei & Zou (2019)) is one of the sentence edit models that use simple character and word level perturbations in the sentences while keeping the target labels intact. EDA uses random insertion, substitution, swap, and deletion for this process. A similar approach has been provided in (Niu & Bansal (2018)), where they categorize the replacements into two types, i.e., *should-change* which includes negation placement and antonym replacement, and *should-not change* which includes random swap, stop word dropout, text paraphrasing, and grammar errors. (Papadaki (2017)) introduces an interpolation and extrapolation-based approach, where they tweak the word embeddings using the word vectors from nearest neighbors. In the case of interpolation, word embeddings tend to move towards, and in extrapolation word embeddings tend to move away from the neighbors. However, these models fail to capture the sentiment, context, and domain-specific knowledge while making the sentence's word and character level edits. In 2020, Wei et al. (Wei et al. (2020)) deployed EDA to increase the dataset size for classifying the questions related to COVID-19 collected from 13 online sources. We propose Mod-EDA, an extension of EDA, because of the simplicity and effectiveness of the model on smaller training datasets.

2.4. Feature Representation

The proposed work utilizes the state-of-the-art word embedding models for feature representation of the tweets. This subsection gives a brief introduction to these models for a theoretical foundation.

2.4.1. Bidirectional Encoder Representation from Transformers (BERT)

BERT (Devlin et al. (2018)) stands for Bidirectional Encoder Representation from Transformers. It is an open-sourced neural network-based natural language processing pre-trained model introduced by Google Inc. in 2018. This model is based on the Transformer model (Vaswani et al. (2017)) that processes all the words in a sentence parallelly using the attention-based mechanism, which is different from earlier CNN (Krizhevsky et al. (2012)) and LSTM (Hochreiter & Schmidhuber (1997)) based models that perform sequential processing for predictions. BERT is based on Transformers and thus requires much less memory and better computation performance with GPUs and TPUs because of task parallelization. This helps in providing better accuracy than the sequential models. The pre-trained model is publicly available on TensorFlow Hub[2].

2.4.2. Embeddings from Language Models (ELMo)

ELMo (Peters et al. (2018)) stands for Embeddings from Language Models, an NLP framework developed by AllenNLP, publicly available on TensorFlow Hub[3]. ELMo creates word vectors from the text data on top of a two-layer bidirectional LSTM based language model. It is based on a character-level convolutional network. ELMo takes care of the context of a word while creating the word vectors. The bidirectional model creates two intermediate vectors for a word in the forward and backward pass, respectively. The weighted average of intermediate word vectors is the final representation returned by ELMo. This model can create different word vectors for the same word depending on the context.

2.4.3. Universal Sentence Encoder (USE)

USE (Cer et al. (2018)) stands for Universal Sentence Encoder, a transformer-based language representation model developed by Google. Unlike BERT and ELMo, which convert the words into their vector representations, USE creates vector representation for the whole sentence. The pre-trained Universal Sentence Encoder is publicly available to use on TensorFlow Hub[4]. The output vector returns a fixed-length vector representation of 512 dimensions. The USE model has shown the best performance on the text semantic similarity tasks.

3. 2019 Indian General Elections Case Study: Motivation

India has been ranked the third country in maximum Twitter user-population as it consists of 17 million Indian users. According to the New Delhi-based Center of Media Studies, this general election was marked as the world's most expensive election with a cost of around $8 billion (Bloomberg (2019)). This amount

[2]https://www.tensorflow.org/hub/tutorials/bert_experts

[3]https://tfhub.dev/google/elmo/3

[4]https://www.tensorflow.org/hub/tutorials/semantic_similarity_with_tf_hub_universal_encoder

has grown by around six times since 1998[5] and was even higher than the 2016 US presidential elections expenditure of \$6.5 billion. In his statement, Chairman of the Center for Media Studies, N. Bhaskara Rao[6], said that most of the jump in spending came with social media, travel, and advertising. Social media spending was estimated to be worth Rs. 50 billion, which is dramatically higher than that of spend of Rs. 2.5 billion in 2014[7]. Thus, utilizing India specific Twitter data as the case study can be highly insightful. In this section, we motivate the use of social media data for Indian election campaigning.

3.1. Indian Government and Social Media

In 2006, the Government of India put forward a National e-government plan (NeGP) comprising 27 mission mode projects to fully utilize the electronic media to provide all government services online to the citizens (MeitY (2018)). Later in 2011, four more projects were added to NeGP, which focused on health, education, a public distribution system (PDS) and posts (Indian Post) as per the second administrative reform commission report titled- “Promoting e-governance- The smart way forward” (MeitY (2020)). This later induced the ‘Digital India’ campaign[8] launched by India’s government in 2015 to make India a digitally empowered community. The emergence of e-governance helped alleviate the flaws of the traditional policy-making and governance infrastructure, based on statistics generated by government agencies and international bodies (Severo et al. (2016)). Researchers have identified the public and real-time social media data as the perfect resource to improve policy-making, which can further promote and maintain the social, economic and health status of the citizens (Zinnbauer (2015), Bertot et al. (2012), Landsbergen (2010)). This allows the politicians to discuss the policy proposals, receive feedback and measure political discontent (Zhuravskaya et al. (2020)).

In April 2012, as a part of NeGP, the Ministry of Electronics and Information technology (Meity) provided a framework and guidelines for citizens engagement through social media to “enable various government agencies to create and maintain their social media strategies”[9]. The report states that policy-making can be more citizen-centric with the help of social media engagements between government and citizens. Many popular government schemes and campaigns launched by the Indian government were successful over the years. For example, in June 2018, WHO recognized India’s success in reducing the Maternal Mortality Ratio by 77% (WHO (2018)). One of the reasons behind this success was considered to be the successful implementation of government schemes and campaigns such as The Pradhan Mantri Surakshit Matritva

---

[5] https://cmsindia.org/sites/default/files/2019-05/Poll-Expenditure-the-2019-elections-cms-report.pdf

[6] https://www.dnaindia.com/analysis/column-social-media-and-elections-2728093

[7] https://economictimes.indiatimes.com/news/elections/lok-sabha/india/why-indias-election-is-among\-the-worlds-most-expensive/articleshow/68367262.cms?from=mdr

[8] https://digitalindia.gov.in

[9] https://www.meity.gov.in/writereaddata/files/Social%20Media%20Framework%20and%20Guidelines.pdf

Abhiyan (PMSMA)[10] and Janani Shishu Suraksha Karyakaram (JSSK)[11] and the social media reach of these schemes and campaigns. Similarly, in September 2019, Prime Minister Narendra Modi received Global Goalkeeper Award[12] for Swachh Bharat Mission[13] from Bill and Melinda Gates Foundation (NDTV (2019)). The consideration of social media is one of the reasons behind this success which motivates us to evaluate citizens' perceptions of these schemes.

3.2. Twitter and Election Campaigning

Primarily, the Twitter based election studies have been focused on the developed countries or nations such as United States (US) (Grover et al. (2019), Grinberg et al. (2019), Srinivasan et al. (2019), Hua et al. (2020)) and United Kingdom (UK) (Burnap et al. (2016), Anstead & O'Loughlin (2015)). However, Twitter has been successfully utilized as the resource for extracting deep insights in many other countries having different demographics, resources and political infrastructure. Some of these countries include Malaysia (Sun et al. (2019)), Colombia (Pedro-Carañana et al. (2020)), Russia (Miller (2019)), Denmark (Derczynski et al. (2019)), Indonesia (Budiharto & Meiliana (2018)), and Austria (Seethaler & Melischek (2019)) etc.

Over the years, many Twitter data-based Indian elections prediction frameworks have been built. In (Singh et al. (2020a)), authors analyzed the Twitter data for Punjab (State in India) Assembly elections to predict the election results. This work can be considered interesting because Punjab comprises just 2.19% of the total Indian population. Thus, Twitter has proven to be a helpful election prediction resource at a very fine-grain level also. Similarly, in (Srivastava et al. (2015)), authors analyzed the Delhi (Capital of India) Assembly Elections 2015 and proposed a sentiment analysis-based framework to predict the election results. Delhi also comprises approx. 2.194% of the Indian total population. Twitter's success in predicting elections and performing election-based studies on coarse and fine-grained levels motivates us to utilize Twitter as the data source to conduct this research work.

3.3. Indian General Election 2019

Indian General Elections were held in seven phases from 11 April to 19 May 2019 to constitute 17^{th} Lok Sabha, i.e., House of People, the lower house of Parliament of India. The results were declared on 23^{rd} May 2019. During this election period, the voters turn out was 67 %, which was the highest number overall and the highest participation was of women voters. The primary candidate parties who stood for election were National Democratic Alliance (NDA), led by Bharatiya Janata Party (BJP) and United Progressive Alliance (UPA), with the largest member party Indian National Congress (INC). NDA (BJP) won the 2019 Indian General Elections with 303 out of 543 seats.

[10] `https://pmsma.nhp.gov.in`

[11] `https://www.nhp.gov.in/janani-shishu-suraksha-karyakaram-jssk_pg`

[12] `https://www.gatesfoundation.org/goalkeepers/about-event/awards/`

[13] `https://swachhbharatmission.gov.in/sbmcms/index.htm`

Various studies analyze Twitter data during the election period to gather multiple insights. As a recurring activity, the launch of different schemes and policies tended towards social welfare, public health, general well-being and citizens' security take place all over the tenure of any government. Many schemes targeting different societies/communities of India are launched every year. The proposed research work is based on the hypothesis that such schemes' success and failure may play a vital role in deciding the public's stance for a given party during elections. The ruling party may tend to use this scheme-based information and promotions for their election campaign promotion. The opposing parties may tend to generate awareness regarding these government schemes' successes and failures to create some influence. This study focuses on exploring the government schemes-related tweets generated during elections to assess their relationship. We accomplish this by examining the government scheme data in terms of two parameters, i.e., **(i)** change in the government scheme-related tweet spread in before, during, and after the election duration, and **(ii)** correlation between the Twitter data and the election outcomes in the form of the seats won per state. The details of the steps followed are discussed in the sections below.

4. Methodology

The vast amount and unstructured format of the Twitter data require text-based machine learning models for knowledge extraction. In the proposed research work, the process to understand the relationship between the government schemes and electoral dynamics has been divided into four parts, **(i)** *Tweet collection and pre-processing*: it deals with the collection of data and converting the raw format into the usable format. This includes feature extraction and text cleaning. **(ii)** *Location Extraction*: it deals with geo-parsing the tweets to collect the location information from the tweets. **(iii)** *Tweet Classification*: this part proposes a module to classify the tweets into the target groups i.e. *general health, poverty-related, mother & child care, and Election* and the type of user-engagement in the tweet i.e. *informational, promotional, appreciation, and complaint*. This module proposes two steps/sub-modules, first, *text augmentation* where word-level perturbations are performed on the tweets to increase the size of the labeled dataset, and second, the application of supervised learning based tweet classification to label the tweets with the corresponding target groups and the type of user-engagement. **(iv)** *Government Scheme versus Electoral Dynamics*: this part proposes two analytical modules to study the change in the government schemes related tweet diffusion patterns with the change in the election campaign process. The first module studies the tweet diffusion pattern with respect to the voting phases in the election, and the second module studies the tweet diffusion pattern with respect to the primary electing parties. Figure 1 depicts the block diagram representation of all the modules/sub-modules of the proposed framework. A detailed explanation of all the modules has been provided further in the section. The data collection and pre-processing steps have been explained in Section 5.1 for a better understanding of the case study on Indian Elections.

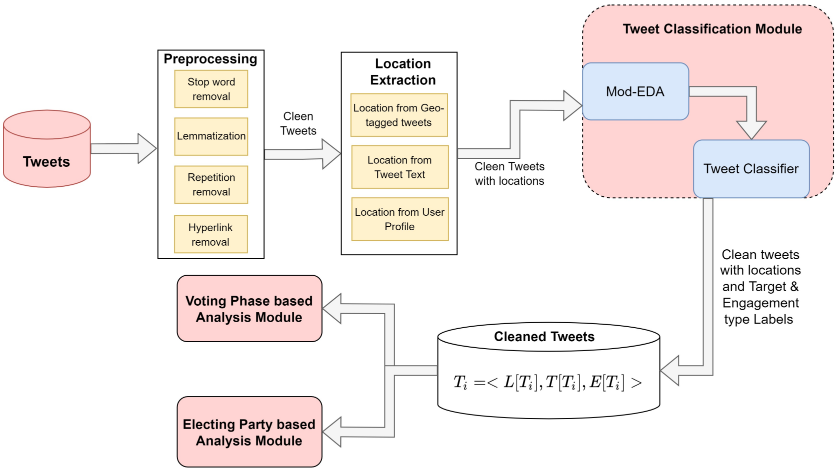


Figure 1: The block diagram description of all the parts of the proposed framework. Here, tweet T_i is represented as a tuple $< L[T_i], T[T_i], E[T_i] >$ after the tweet classification step, where, $L[T_i]$ is the location, $T[T_i]$ is the Target label and $E[T_i]$ is the Engagement type for the tweet T_i.

4.1. Location Extraction

A general Twitter data collection results in approx 2% of the geo-tagged tweets. This can not be considered enough for generating location-based results. There are primarily three measures of extracting the location information from different attributes in a tweet i.e. **(i)** the *geo-tagged tweets*, **(ii)** *tweet text* and **(iii)** *user profile* . The *geo-tagged tweets* contain the latitude-longitude information in the tweet itself. The *tweet text's* location is the location extracted through the mention of a place/location in the tweet text. This information can be extracted using a list of the names of cities/towns and states in the country. The *user location* is the location extracted through the profile of the tweet poster. The user profile contains the location of the user account. The resultant geo-location of the tweet can be issued in the priority {*geo-tagged tweet (i.e., lat, long)* > *text mention* > *user-location*}; where geo-tags have the highest priority. If a tweet was geo-tagged, i.e., it contains the latitude and longitude, it is set as the location information of the tweet. If the tweet is not geo-tagged (i.e., no latitude-longitude is available), the location extracted from the tweet text is considered the tweet location. Finally, if both latitude/longitude and tweet text locations are not available, then the user profile location is considered the tweet location. All the resultant locations are further parsed using the Google Geocoding API[14] to convert location names into latitudes and longitudes and vice-versa. For a tweet T_i, its location is saved as $L[T_i]$.

4.2. Tweet Classification Module

After the text pre-processing and location extraction from the tweets, the next component focuses on the classification of the tweets in terms of the target groups and the type of user-engagements. The analysis of

[14] `https://developers.google.com/maps/documentation/geocoding/start`

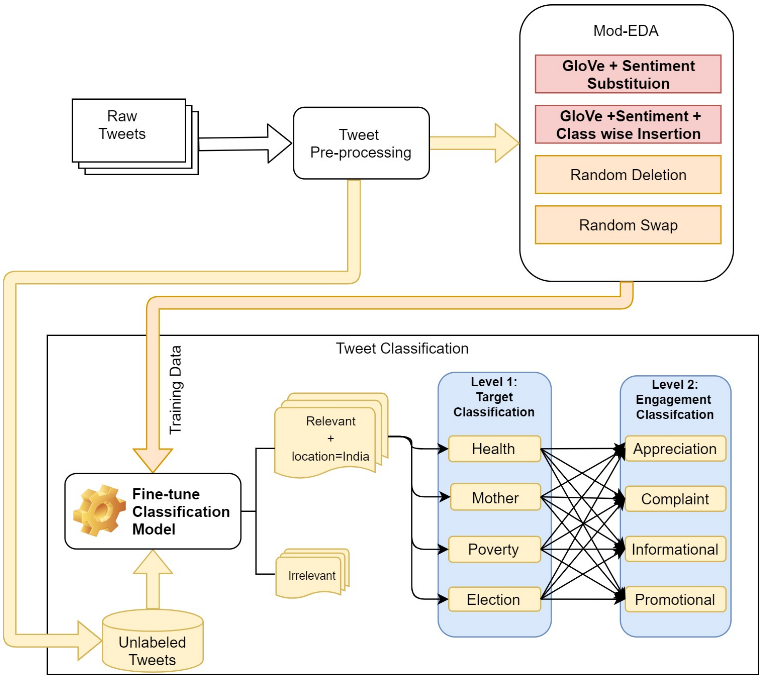


Figure 2: The block diagram description of the *Tweet Classification Module* which classifies the government scheme tweets on the basis of target group and the type of engagement

the Twitter data in terms of *target groups*, i.e., the sections of the community focused by the scheme and the *type of user-engagement*, i.e., the kind of views presented by the tweet poster, can help in providing better insights into the awareness and involvement of the public and effectiveness of these schemes. We propose to employ the deep learning-based language representation models: BERT (Devlin et al. (2018)), ElMo (Peters et al. (2018)) and USE (Cer et al. (2018)) for the tweet classification using supervised learning. However, the supervised models require an ample amount of annotated data to prevent the model from over-fitting. The annotation process is a time-intensive and laborious process, and thus, it becomes a hindrance in using the supervised models on the real-time data sets and applications. Therefore, to overcome this problem, we propose Mod-EDA (Dhiman & Toshniwal (2020)), a text augmentation method, which increases the size of the data set by making word-level perturbations in the text. The proposed method retains the contextual domain-specific information while making the perturbations, which helps in maintaining the accuracy of tweet classification. The proposed tweet classification module has been provided in Figure 2 as a block diagram for better understanding. A detailed explanation of the text augmentation approach Mod-EDA and the tweet classification has been provided further in the section.

4.2.1. Modified Easy Data Augmentation(Mod-EDA)

Mod-EDA (Dhiman & Toshniwal (2020)) is an extension of the text augmentation model EDA (Wei & Zou (2019)). EDA (Wei & Zou (2019)) is one of the sentence edit models that use random insertion, substitution, swap, and deletion to introduce character and word level perturbations in the sentences while keeping the target labels intact. EDA fails to consider the domain-specific contextual knowledge and the class sentiments while making the perturbations. Mod-EDA proposes to retain the domain-specific knowledge as well as some randomness in the text during augmentation. The primary steps involved in Mod-EDA include *Modified Substitution, Modified Insertion, Random Deletion*, and *Random Swap.* Random Deletion randomly deletes some words from a sentence, and the Random swap performs the swaps in the position of two or more words. These two have been taken directly from EDA without any changes. However, in Modified Substitution and Modified Insertion, domain-specific knowledge has been exerted in the form of word embeddings, generated with the help of GloVe (Pennington et al. (2014)) and sentiment polarity of the words & sentences calculated using the AFINN Library (Nielsen (2017)). We use the 200-dimensional GloVe vectors pre-trained on 2 Billion tweets available at (Pennington) having vector representations for 27 Billion tokens. These vectors are in a ready-to-use format which makes them easy and faster to use. Algorithms 1 and 2 summarize the steps proposed for Modified Substitution and Modified Insertion. A brief overview of the steps involved in Mod-EDA is provided below:

Modified Substitution. This step proposes to substitute a word based on the prior knowledge retained from the dataset under consideration. The contextual knowledge makes the substitution domain-specific. We also keep track of the sentiment of the word to make an informative substitution. Algorithm 1 provides the step-by-step procedure followed. Here, we have used GloVe generated embeddings as given in (Dhiman & Toshniwal (2020)). These embeddings are further used to generate the top t similar words to the original word ($cand\_list \leftarrow WE.most\_similar(word, topn = t)$). Next, the word having the minimum difference in the sentiment polarity with the original word is selected as the replacement ($cand\_sub \leftarrow min(polarity(cand\_list) - sent\_word)$). AFINN Library (Nielsen (2017)) has been used to calculate the sentiment polarity of the words.

Modified Insertion. Similar to the Modified Substitution, Insertion also chooses the word to be inserted, which matches the context of the class under consideration. Algorithm 2 provides the step-by-step procedure followed. First, a list of class-specific frequent keywords are generated ($freq\_keys \leftarrow most\_frequent\_keys(text\_list)$). Next, a set of t keywords is selected at random from the frequent keyword list ($freq\_keys$), which is further populated with the set of most similar keywords to the keyword in $freq\_keys$ ($cand\_list \leftarrow WE.most\_similar(cand\_list)$). Finally, the word to be inserted is selected based on the most similar sentiment polarity to the sentence ($cand\_ins \leftarrow min(polarity(cand\_list)\text{-}sent)$).

4.2.2. Tweet Classifier: Classifying Target Groups and User Engagements

The text augmentation performed in the previous step results in an increased number of labeled tweets. These augmented tweets are further processed using the language representation models for classification

Algorithm 1: Modified Substitution

Data: Tweets $T = \{t_1, t_2, \ldots, t_n\}$, k,t

Result: A list of tweets with substitution

```
WE ← GloVe(T)
for tweet t_i in T do
    t_i ← remove_stopwords(t_i)
    t_i ← remove_punctutations(t_i)
    orig_word ← tokenize(t_i).random(k)
    for word in orig_word do
        cand_list ← WE.most_similar(word, topn = t)
        sent_word ← polarity(word)
        cand_sub ← min(polarity(cand_list) − sent_word)
        t_i ← replace(word, cand_sub)
    end
end
```

Algorithm 2: Modified Insertion

Data: Tweets $T = \{t_1, t_2, \ldots, t_n\}$, k,t

Result: A list of tweets with insertions

```
WE ← GloVe(T)
for class in classes do
    text_list.append(t_i if label[t_i] == class)
    freq_keys ← most_frequent_keys(text_list)
    for t_i in text_list do
        sent ← polarity(t_i)
        orig_id ← len(tokenize(t_i)).random(k)
        cand_list ← freq_keys.random(t)
        cand_list ← WE.most_similar(cand_list)
        cand_ins ← min(polarity(cand_list) − sent)
        t_i[orig_id] ← cand_ins
    end
end
```

in the relevant categories. This section describes the steps followed for performing the tweet classification. The Tweet Classification box in Figure 2 provides a block diagram description of this step.

The self-collected Twitter dataset contains many visible as well as invisible impurities. The visible impurities can be easily removed with the pre-processing steps (as depicted in section 5.1). However, there may exist some invisible or logical impurities, which makes the task challenging. It may happen that some government schemes' names may be similar to some common popular keywords, which can add redundant data during collection. For example, the data collection for a government scheme named 'NITI AAYOG,' may pull the tweets for an actor named 'Niti Taylor.' Thus to remove such pragmatic ambiguities caused by polysemy of words, we leverage SOTA deep learning-based language representation (LR) models to facilitate text classification. The application of LR models requires enough labeled data for training. Thus, initially, a set of tweets are labeled having equal proportions of *relevant* and *irrelevant* tweets. A tweet is labeled *relevant* if it belongs to any of the government schemes; otherwise, it is labeled *irrelevant*.

The LR models employed for the training are: BERT (Devlin et al. (2018)), ElMo (Peters et al. (2018)) and USE (Cer et al. (2018)). The models are fine-tuned on all the tweets containing an equal number of relevant and irrelevant tweets. The vector representation generated by these LR models is fed into the simple Feed Forward Neural Network with one hidden layer and two nodes at the output layer. We used Rectified Linear Unit (ReLU) as the activation function. The Sigmoid function was applied to derive the probability distribution among classes at the output layer. The values of the hyper-parameters Maximum Sequence Length (MSL), Batch Size (BS), Embedding Size (ES), Hidden Layer Size (HLS), number of epochs, and Learning rate (μ) used are given in Table 1. The values of these parameters were chosen where the accuracy was best. The accuracy of the classifier has been evaluated on 10% (400 tweets) of the dataset (test-set), where 70% (2800 tweets) of the data has been used as training set and 20% (800 tweets) of the data has been used as the validation set. We made use of early stopping to ensure our model is not overfitting. We use the standard definition of accuracy i.e. $\frac{TP+TN}{TP+TN+FP+FN}$, where, TP is True Positives, TN is True Negatives, FP is False Positives, and FN is False Negatives. The details on the size of dataset used for learning have been provided in Section 5.2.1. The learned model is further used to predict the relevance of the tweets for the whole dataset. The tweets labeled irrelevant are discarded from further processing. We follow the same hyper-parameters and classifications steps to perform the tweets' Level 1 and Level 2 classification. The details on the Level 1 and Level 2 are provided further in the section.

There are various studies that judge the success of any government scheme in terms of the *proposed beneficiaries* and the *ease of access* in providing public service delivery (Mishra & Attri (2020)). Here, we propose to extend their study in the context of Twitter data by labeling the tweets with four primary *focus groups/ Target user-groups* and the four primary *types of user-engagements*. The rationale behind the choice of these classifications is as follows. These schemes may be oriented towards various communities and societies, e.g., health, mother/child care, poverty, etc. The government schemes are primarily focused on

Table 1: The values of hyper-parameters used for all the combinations of embeddings with the Feed Forward Neural Network. The abbreviation used are MSL: Maximum Sequence Length, BS: Batch Size, ES: Embedding Size, HLS: Hidden Layer Size, Epochs: number of epochs, and μ: Learning rate.

LR Model	MSL	BS	ES	HLS	Epochs	μ
BERT	75	64	768	512	500	1e-5
ELMo	75	64	1024	256	50	1e-3
USE	-	64	512	256	10	1e-4

the common public comprising of lower and middle-income sections. The sections of community focused by the scheme and the kind of views presented by the tweet poster will help in extracting deeper insights into the awareness and involvement of the public. Thus, to cover most of these needful groups, we classify the tweets to identify the target communities/groups into *general-health, poverty-related, mother & child care, and Election*. The label *Election* here refers to the government scheme related tweets that also mention election-related information. This is referred as *Level 1: Target Classification* in the Figure 2.

Next, to identify the type of information shared, the tweets are labeled into four types of user engagements, i.e., *informational, promotional, appreciation, and complaint*. In correspondence to the sentiment analysis, it can be said that the appreciation-related tweets refer to positive sentimental tweets, and complaint-related tweets refer to negative sentimental tweets. However, general sentiment analysis studies over the government schemes data have shown that most tweets show a neutral sentiment (Dhiman & Toshniwal (2018)). Thus, two categories were created to gain more out of the neutral tweets, i.e., informational and promotional. The informational tweets generally refer to the tweets that spread information in news, statistics, facts, and figures. The promotional tweets tend to present and promote the positive outcomes or after-effects of these tweets. This is referred as *Level 2: Engagement Classification* in the Figure 2. The type of engagement classification will help provide the citizens' perception of these schemes and policies. As a result of this classification, each tweet will have two classes associated with it. Combining both the categories, i.e., target groups and the type of engagement, will provide the perception and level of engagement of these target groups.

Let tweet T_i has a Target group $T[T_i]$ (using Level 1 classification) and Engagement type $E[T_i]$ (using Level 2 classification) as depicted in Figure 1. This labeled data is fed through Mod-EDA to generate up-scaled labeled data for Level 1 and Level 2 classes. The tweet vectors are generated for these augmented labeled tweets. These vectors are further used to train the Feed Forward Neural Network. We train the separate models for Level 1 and Level 2 classification. The configuration of the model used is the same for the relevance classification as given in Table 1.

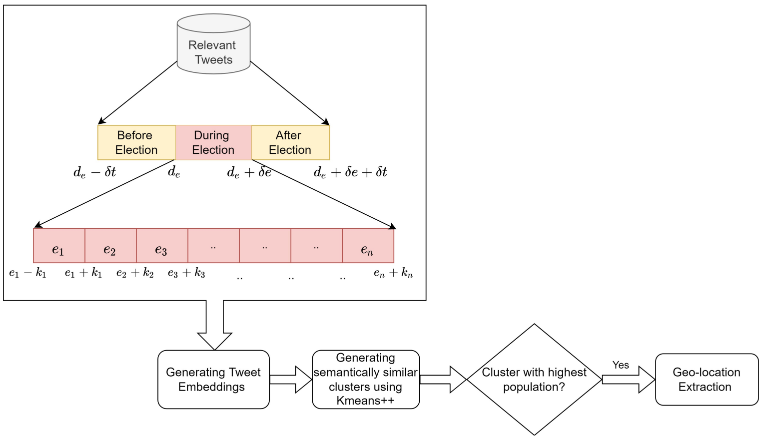


Figure 3: The block diagram description of the *Voting Phase based Analysis Module* proposed to analyze the effect of voting phase on government scheme tweets

4.3. Government Schemes versus Electoral Dynamics

Various studies analyze the Twitter data during the election period to gather various insights. However, as a recurring activity, the launch of various schemes and policies tended towards social welfare, public health, general well-being, and citizens' security take place all over the tenure of any government. Many schemes targeting different societies and communities are launched every year. The success and failure of such schemes play a major role in deciding the stances of the public for a given party during elections. (Deshpande et al. (2019), De Koster et al. (2013)) Different parties also tend to use this scheme-based information for their election campaign promotion. Thus, a study to analyze the role of social media in election campaigns using these schemes can help uncover many patterns. Such a study will help to analyze how social media has an impact on scheme-based election campaigning. To the best of our knowledge, no such study addresses the government launched schemes and policies during the election campaign using social media. In this section, we propose two Twitter analytical modules that investigate the change in tweet diffusion patterns with respect to the voting phases (Section 4.3.1) and the electing parties (Section 4.3.2).

4.3.1. Voting Phase based Analysis Module

As per the Center of Media Studied report- *Poll Expenditure, The 2019 Elections*, "Around 40 percent voters acknowledged receiving poll-related messages on their mobile phone just before the polling day." There are research studies that study the impact of these welfare schemes on the voting decisions of the citizens (Deshpande et al. (2019)). However, very few studies focus on the change in the pattern of information diffusion during the election campaigns. Social media is one of the sources where the freedom of opinion sharing may influence the decisions of the citizens (Rana et al. (2015)). Thus, in this section, we aim to study the change in frequency of tweets for different locations before and after the election duration. We study this in terms of the target groups and the type of engagements on fine-grain locations, and the voting phases.

Figure 3 provides the block diagram description of the module proposed to analyze the effect of the voting phase on government scheme tweets. The voting phase's effect on the government scheme tweets has been studied by splitting the Twitter data into three phases for three time windows. These windows have been named *before elections* phase i.e., data collected from $d_e - \delta t$ to d_e (δt days), *during elections* phase i.e., data collected from d_e to $d_e + \delta e$ (δe days) and *after election* phase i.e., data collected from $d_e + \delta e$ to $d_e + \delta e + \delta t$ (δt days). The time-windows have been chosen to keep the data set size uniform. Here, d_e is a date and its value is chosen in accordance to the start of the voting phase, δe is the number of days between the start and end of the election voting phases and δt is the number of days for which the before and after election phase is to be studied.

Let the elections are held in n voting phases i.e. $\{e_1, e_2, \ldots, e_n\}$, where e_i is the date of voting of the i^{th} phase. Depending on the number of seats and population, a state could have polled in one or more phases. So, the *during election* phase dataset is further divided into n chunks corresponding to the n election phases for a fine-grain level. The time window has been chosen, such that there is an approximately equal number of days between two consecutive voting phases. It can be seen from the Figure 3 that first chunk starts at date $e_1 - k_1$ and ends at $e_1 + k_2$, which also acts as the start date for the second chunk and so on. The value of k_i is calculated as the median of $\{d | d$ is date such that $d > e_i$ and $d \leq e_{i+1}\}$ such that there are equal number of days before and after the voting phase. If there are even number of days between the date e_i and e_{i+1}, any user-defined date can be chosen for the chunking.

Next, all the chunks are analyzed individually to assess information spreading for different voting phases. First, the vector representations for individual tweets are generated by employing BERT. The vector representation of the tweets has been created because we want to cluster the semantically similar tweets. The tweets (as vectors) have been clustered using k-means++ (Arthur & Vassilvitskii (2006)). k-means++ is an extension of the k-means clustering algorithm (Han et al. (2011)), which uses a smart centroid initialization technique. The initial choice of centroids in clustering affects the number of iterations required and the uniformity of the clusters. The clustering approach k-means++ has been chosen for its simplicity, efficiency, and high speed. The elbow curve method has been used to determine the number of clusters to be extracted for each chunk. The clustering leads to the generation of topic-level clusters (Yang et al. (2018), Agarwal et al. (2019)), i.e., the clusters sharing similar information. Furthermore, the most populated clusters in each chunk have been extracted. The most populated clusters focus only on the most popular content spreading for the given election phase. The geo-locations of these clusters help to provide the location aggregation of these topics. The target group and type of engagement level classification done in the previous section have been utilized to investigate the type of content shared during different voting phases and how this content changes with the change in the voting phase's location. Section 5.3.1 discusses the results gathered by this

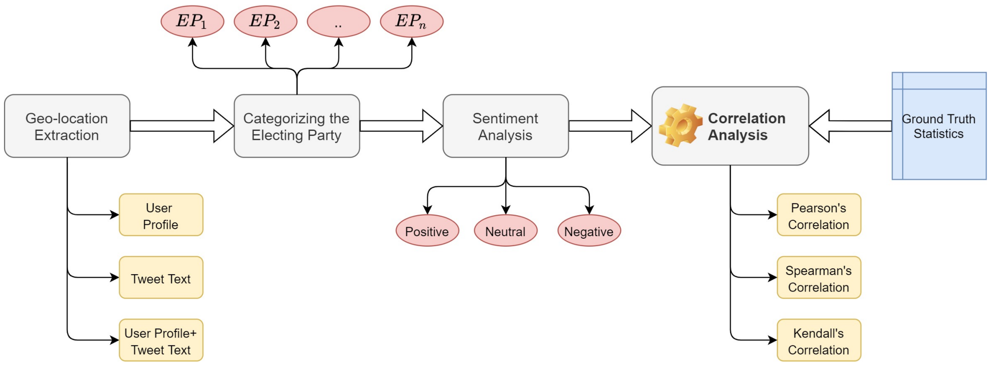


Figure 4: The block diagram representation of the steps followed to calculate the correlation with the electing parties for the *Electing Parties based Analysis Module.*

module.

4.3.2. Electing Parties based Analysis Module

The earlier research has shown that the citizens' trust in a government acts as a precursor of re-election. This trust is influenced by the ease of access and the benefits of the social welfare schemes and policies (Mishra & Attri (2020) Bjørnskov (2010)). The successful implementation of these welfare schemes and the increased citizen trust becomes the basis for claiming credit for state and central government in the election campaigns (Tillin & Pereira (2017) Deshpande et al. (2019)). There are various studies in the literature which try to identify the impact of these welfare schemes on the voting pattern of the citizens (Deshpande et al. (2019) De Koster et al. (2013) Hunter & Power (2007) Zucco Jr (2013)). This section analyzes the government schemes related tweet diffusion pattern in terms of the primary electing parties. Some studies try to predict the election results using the tweet sentiments for the competing parties. However, this section investigates the correlation between the government scheme-related tweets and the election results. As the primary focus of this analysis is to study if the government scheme-related tweets affect the election results, we only study the correlation and do not refer to it as the outcome prediction. Here, we cluster the tweets mentioning different electing parties from different locations. The cumulative sentiments per location per electing party correlate with the ground truth data to investigate any association. The details of the proposed module are given further in the section.

Figure 4 provides the block diagram illustration of the proposed module to investigate the correlation between the government schemes related tweets and the election outcomes in terms of the electing parties. We propose to gather insights on the finer geo-locations. This is accomplished by segregating the tweets in terms of their locations. The geo-locations can be extracted as explained in Section 4.1. In addition to extracting locations from the *tweet geo-tags*, *tweet text* and the *user-profiles*, we propose to combine the user-profile and text provided locations. Approximately 2-3 % of tweets have geo-tagged locations; thus we combine it with all the other approaches without providing a separate block for geo-tagged tweets in Figure 4. Let tweet T_i is any tweet in the corpus. After this step this tweet will have a location associated with it, i.e. $T_i =< L[T_i] >$, where $L[T_i] = \{L_{Text}[T_i], L_{UserProfile}[T_i], L_{Text+Profile}[T_i]\}$, is the location information available via the tweet text, user profile and the combination of the both. The values of $L_{Text}[T_i]$, $L_{UserProfile}[T_i]$, and $L_{Text+Profile}[T_i]$ can be same, different or empty depending on the tweet and the user. The tweet without any of the location information is discarded from further processing.

Let there be n major electing parties i.e. $\{EP_1, EP_2, \ldots, EP_n\}$. The next step is to segregate the tweets in terms of the electing parties. A simple keyword search-based approach can be used to label a tweet with a corresponding electing party. Here, we try to search for keywords specific to any electing party in any tweet. A tweet may mention keywords belonging to more than one electing party. This situation is resolved by maintaining the number of occurrences of the related keywords. Let any electing party EP_j can be defined as a collection of keywords $EP_j = [KW_1, KW_2, \ldots, KW_m]$. For i^{th} tweet T_i, if it mentions the keyword $EP_j[k] = KW_k$, the count of EP_j is incremented by 1. The tweet T_i is assigned the electing party EP_j with the maximum count. After this step, the tweet has two associated entities i.e. $T_i =< L[T_i], EP[T_i] >$, where $EP[T_i]$ denotes the name of the assigned electing party.

Yaqub et al. (Yaqub et al. (2020)) have shown that the combination of geo-locations and sentiment analysis can reflect upon the ground public opinion for the election data. Thus, sentiment labeling of the tweets is performed. The sentiments used were positive, negative, and neutral. The lexicon sentiment libraries are easy and faster to use as compared to the supervised learning methods. Thus, the tweets are labeled using the AFINN lexicon library (Nielsen (2011)). The AFINN python module returns a Real number as a sentiment score for any given sentence. The tweet is labeled positive for a positive score, negative for a negative score, and neutral for a score equal to 0. The sentiment label helps extract the tweet poster's polarization for the respective electing party for a given location. After the sentiment analysis, each tweet is labeled with a three-item tuple, i.e. $T_i =< L[T_i], EP[T_i], S[T_i] >$, where $S[T_i]$ is the associated sentiment such that $S[T_i] = \{Positive, Negative, Neutral\}$.

Finally, the correlation between the sentiment scores per electing party for different locations and the ground truth data. Here, ground truth data is the number of wins per electing party per location. We employed three correlation measures, i.e., Pearson correlation (Sedgwick (2012)), Kendall correlation (Kendall (1938)), and Spearman correlation (Zar (2005)) to calculate the correlation scores among the number of tweets and number of seats won per state. The choice of three correlation measures helps to explore the correlation measure which gives the highest values. Section 5.3.3 shows the results for the proposed analysis.

Table 2: Statistical description of the Twitter dataset collected from December 2018 to November 2019

Sr. No.	Attributes	Value
1	Total number of Keywords	106
2	Total number of tweets	1,368,378
3	Number of relevant tweets	737,939
4	Example Keywords	#IPledgeFor9, #MaternalHealth, #NHPIndia, JananiSurakshaYojna, Ujjwalayojna, PoshanAbhiyan, IndiaMGNREGA, MGNREGA, Sukanya Samriddhi Yojna, betibachaobetipadhao, Deen Dayal Antyodaya Yojana, Rashtriya Mahila Kosh
5	Example Tweets	RT @NHPINDIA: If not treated properly, #Gangrene may lead to a fatal infection. RT @NHPINDIA: Resist your temptation to consume any form of #Tobacco. Quitting is not easy, but it's worth it! #SayNOToTobacco #SwasthaBharat RT unfoundation: 1 out of 7 facts about #MaternalHealth: around 303,000 girls and women die every year.; RT @NHPINDIA: Food choices exhibit a major role in your oral health.Read to know what are the types of food that are beneficial for your oral health

5. Results and Discussions

In this section, the proposed framework is validated for the case study on Indian Elections 2019 and the Twitter data related to some government welfare schemes launched by the Indian government. The details of data collection and pre-processing have been provided in Section 5.1. The proposed research work is broadly divided into two parts, i.e., **(i)** *Tweet Classification Module* which includes Mod-EDA and Text Classification, and **(ii)** *Government Schemes versus Electoral Dynamics* which includes the study of the government scheme related tweets in terms of the voting phases and the electing parties. We discuss the results for the respective modules in Section 5.2 and Section 5.3 respectively.

5.1. Dataset Description: Data Collection and Preprocessing

The dataset comprises the tweets collected between December 2018 to November 2019, containing one or more government scheme-related keywords and hashtags. A list of government scheme names and hashtags related to poverty alleviation, general public health and maternal health etc., was compiled from various sources[15]. This list comprised of 106 keywords and hashtags employed for data collection using Twitter Streaming API. For example, PoshanAbhiyan, IndiaMGNREGA, #IPledgeFor9, #NHPIndia, JananiSurakshaYojna, Ujjwalayojna, #MaternalHealth, MGNREGA, Sukanya Samriddhi Yojna, Deen Dayal Antyodaya Yojana, betibachaobetipadhao, and Rashtriya Mahila Kosh, etc. are some of the keywords used. The Twitter streaming API collected approx. 1.4 million tweets, stored in the JSON format. A statistical description and sample tweets of the Twitter corpus are given in Table 2.

[15] `https://www.india.gov.in/my-government/schemes`

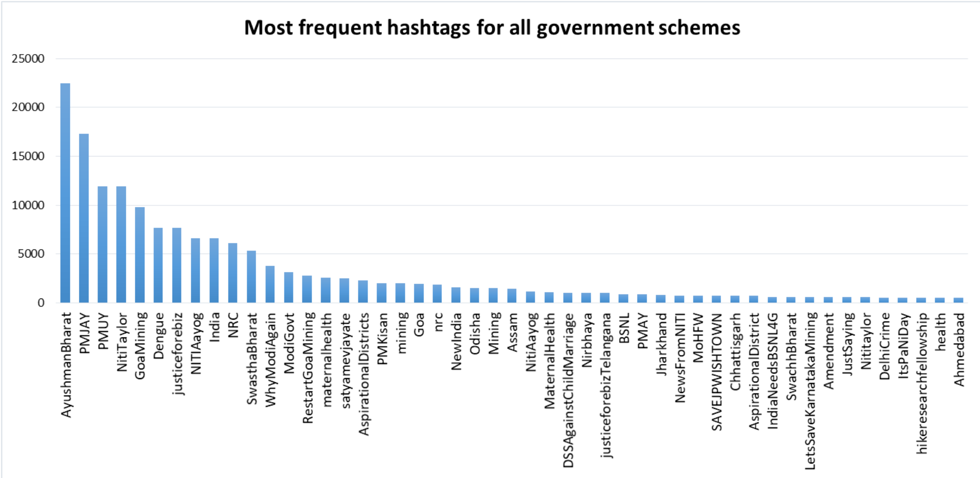


Figure 5: The top 50 most frequently used hash-tags and keywords extracted from the corpus.

The Twitter Streaming API returns the raw data comprising a wide range of attributes such as text, timestamp, place, user details, user mentions, etc. In this step, the Twitter data has been pre-processed to retain the relevant information from the raw data. All the features collected for a tweet are not relevant to the current study. Thus, to minimize the space requirement, only the most relevant features have been saved. The features used are, i.e., timestamp, text, place, and user name from the tweets and the retweets. In the raw format, the tweet text consists of many impurities that may hinder the application of the text analysis and language representation models. The tweets have been cleaned to remove these impurities from the tweet text. First, the tweets are tokenized and the tokens are then lemmatized to extract the root form of the tokens. Furthermore, the tweets are cleaned by removing all the stop-words, hyperlinks, special symbols, recurring characters, and other non-decodable information from text. The dataset contained very general keywords, such as *'MaternalHealth'* and *'MotherCare'*, which collected data all over the world. Thus, to keep the tweets relevant to the current study, the tweets posted outside India have been removed. The Indian Twitter data may contain tweets from different regional languages such as Hindi, Gujarati, Tamil, Telugu, etc. However, to increase the re-usability of the proposed model, only the tweets written in the English language have been retained for further processing. The tweets belonging to other languages have been discarded. Furthermore, we performed location extraction from the tweets where approximately 92.4% were labeled with a corresponding location. Figure 5 provides the top 50 most frequent hash-tags extracted from the corpus. The most frequently used keyword is AyushmanBharat[16], a health insurance scheme, aka Pradhan Mantri Jan Arogya Yojana (PMJAY), which provides free access to healthcare for 40% of people in the country.

[16]`https://pmjay.gov.in`

5.2. Tweet Classification Module

In this section, the proposed module is explored for the collected Twitter dataset for Indian government schemes. First, a detailed description of the tweet labeling has been provided in Section 5.2.1. In Section 5.2.2, the performance of the proposed Mod-EDA has been evaluated using the ground truth dataset and the collected real-time Twitter dataset. Furthermore, a detailed analysis is performed over the classification results on the Indian government schemes Twitter data.

5.2.1. Tweet Labeling

First, 4000 tweets were labeled manually into two categories having an equal number of tweets, i.e., relevant and irrelevant. Three Master's students did this labeling, and the label with at least 60 percent occurrence was taken to be the final label. This labeled data is used to train the classification model for three language representation models, i.e., BERT, ELMo and USE. The classification accuracy for relevant and irrelevant models were 0.92, 0.91, and 0.90, respectively. The application of the trained model on the whole dataset resulted in approx 0.7 million tweets (with location). All the further work has been performed on approx 0.7 million tweets related to government schemes. The relevant 2000 tweets were further labeled for the Target groups and the Engagement type as explained in Section 4.2.2. The support count for the classes in Target groups was 528, 434, 380 and 658 for general-health, poverty-related, mother & child care, and Election, respectively. Table 3 provides a brief example of the tweets tagged with respective Target Group and Engagement Type. The support count for the classes in Engagement types was 451, 525, 571 and 453 for informational, promotional, appreciation and complaint, respectively. This labeled data is fed to Mod-EDA to generate the up-scaled dataset of the augmented tweets for both categories. The following section evaluates the performance of Mod-EDA for the ground truth data and the Indian government schemes Twitter data.

5.2.2. Tweet Classifier

This section provides a brief overview of the results of the classification performed. The performance of the proposed Mod-EDA is tested on the text classification task with BERT, ELMo, and USE. The results are an average of five random seeds. Before applying Mod-EDA on a Twitter dataset, the accuracy of Mod-EDA is compared with EDA on a benchmark dataset SST2 (Stanford SentiTreebank dataset 2013) (Socher et al. (2013)). SST2 dataset contains 68221 sentences labeled 0 or 1 for their sentiments. This dataset is used for validating the proposed model and compare the performance of the proposed model with the base Model EDA. We run BERT, ELMo, and USE on the dataset without any augmentation, with EDA, and with Mod-EDA.

The text augmentation approach creates nine similar instances for a single sentence. After applying Mod-EDA on the relevant tweets, the support count for the classes in Target groups was 4752, 3906,

Table 3: A sample of tweets tagged with categories from the Target groups and the Engagement types

Sr. No.	Tweet Example	Target Group	Engagement Type
1.	Rahul Gandhi @RahulGandhi offering the unemployed youth Rs 10,000 per month, is your idea of 'EMPOWERMENT' and Employment creation ?? @narendramodi @nitin_gadkari @NITIAayog	Election	Complaint
2.	Janandolan community engagement is key to make kuposhan mukt gaon. @NITIAayog	Health	Information
3.	They don't consider adharcard as a valid adress proof and ask for electricity bill Mandatory. Today they have behave rudely with me and ordered to show electricity bill for sukanya samriddhi yojana (for my niece)	Mother	Complaint
4.	With insurance coverage of upto 5 lakh for inpatient Ayushaman Bharat is bringing change in the lives of middle class and poor sections of society	Poor	Promotion

3420 and 5922 for general-health, poverty-related, mother & child-care and Election, respectively and for classes in Engagement types was 4059, 4725, 5139 and 4017 for informational, promotional, appreciation and complaint, respectively. If text augmentation is performed on all the tweets, this may cause overlap in the training and testing set, which may cause the model to overfit. Thus, to test the accuracy of the classification model, 90 percent of the data is separated from the whole corpus, and Mod-EDA and EDA are applied to it to generate an augmented corpus. Next, the model is trained on this 90 percent of the data by taking training and testing set in the 80-20 ratio. This 20 percent of 90 percent data is used to test the LR-feed forward model and the accuracy of this model is referred to as the validation accuracy. Later, the trained model is tested on the remaining 10 percent of the unseen data. The accuracy of the remaining 10 percent unseen data is referred to as the testing accuracy. This type of training-testing ensures that there is no overlap between the training and the testing data due to the proposed text augmentation. The accuracy of all the models were captured in terms of $Precision_i$ (Equation 1), $Recall_i$ (Equation 2), and $F1_i$ (Equation 3) for class i using 10-fold cross-validation. We use micro-average to calculate the aggregate of these metrics. The micro-average has been used because it adequately captures the class imbalance in multi-class classification setup. Table 4 shows that the testing Recall, Precision and F1-score for all the models is lower than the respective validation scores. This is because the testing data consists of purely

Table 4: Comparison of Mod-EDA, EDA and without augmentation on SST2 dataset.

	Validation Results				Testing Results			
	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy
BERT	0.7030	0.7498	0.7498	0.7165	0.7153	0.7704	0.7418	0.7319
EDA+BERT	0.9075	0.9342	**0.9207**	**0.9195**	0.7201	0.7589	0.7390	0.7320
Mod-EDA+BERT	0.9052	0.9257	0.9153	0.9144	0.7414	0.7677	**0.7543**	**0.7500**
ELMo	0.7130	0.7298	0.7213	0.7180	0.6824	0.7223	0.7018	0.6931
EDA+ELMo	0.9043	0.9234	**0.9137**	**0.9128**	0.7024	0.7223	0.7122	0.7081
Mod-EDA+ELMo	0.8924	0.9254	0.9086	0.9069	0.7134	0.7323	**0.7227**	**0.7191**
USE	0.7308	0.7198	0.7252	0.7273	0.6924	0.7223	0.7070	0.7007
EDA+USE	0.8923	0.9123	**0.9022**	**0.9011**	0.6902	0.7112	0.7005	0.6960
Mod-EDA+USE	0.8935	0.8933	0.8934	0.8934	0.7034	0.7223	**0.7127**	**0.7089**

unseen data (without any augmented overlaps in the training data). We consider testing results to be a more accurate measure than the validation results because these are less prone to overfitting. Table 4 shows that Mod-EDA has an average improvement of 1.76% from EDA (maximum 2.03% with BERT) and 1.82% (maximum 2.98% with BERT) from data without augmentation in testing F1-score values on the SST2 dataset. We further extend the same evaluation for the government schemes Twitter data.

Table 5 summarizes the testing recall, precision, F1-score, and accuracy of three state-of-the-art pre-trained language representation models BERT, ELMo, and USE with Mod-EDA on the government schemes Twitter data. The table shows that classification accuracy is best with BERT with an F1-score of 0.71 and an accuracy of 0.85. The fine-tuned BERT model was further applied to the entire Twitter corpus for Dec 2018 to Nov 2019 to predict the different categories of the tweets. The application of the learned model on the Twitter data assigned the Target group and the Engagement type for all the tweets. We further analyze the results of the classification to better understand the government schemes Twitter data.

$$Precision_i = \frac{M_{ii}}{\sum_j M_{ji}} \tag{1}$$

$$Recall_i = \frac{M_{ii}}{\sum_j M_{ij}} \tag{2}$$

$$F1_i = 2 \cdot \frac{Precision_i \cdot Recall_i}{Precision_i + Recall_i} \tag{3}$$

where, M is the confusion matrix such that a given row of the matrix corresponds to specific value for the "truth".

Table 5: Accuracy measurement of BERT, ELMO and USE for fine tuning the labeled dataset

		Mod-EDA +BERT			Mod-EDA+ELMO			Mod-EDA+ USE		
		Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Target Groups/ Communities	Election	0.71	0.71	0.71	0.75	0.64	0.69	0.67	0.57	0.62
	General-health related	0.9	0.93	0.92	0.85	0.95	0.89	0.81	0.93	0.86
	Poor people's health	0.9	1	0.95	0.69	1	0.82	0.82	1	0.9
	Mother-child health	0.78	0.7	0.74	0.92	0.6	0.73	0.71	0.5	0.59
	Accuracy			0.8557			0.8269			0.7789
Type of Engagement	Appreciation	0.66	0.57	0.61	0.65	0.67	0.66	0.44	0.5	0.47
	Complaint	0.65	0.85	0.73	0.62	0.77	0.69	0.64	0.69	0.67
	Information	0.65	0.63	0.64	0.44	0.47	0.46	0.39	0.53	0.45
	Promotion	0.6	0.63	0.61	0.55	0.56	0.55	0.7	0.53	0.61
	Accuracy			0.697			0.6432			0.5537

5.2.3. Discussion

The tweet classification results in tweet T_i having three-element tuple $< L[T_i], T[T_i], E[T_i] >$ i.e. Location, Target group and Engagement type. Figure 6 provides the box plot representation of the number of tweets for different classes. The plot shows the location-wise distribution of the number of tweets for different Target groups for different Engagement types. The figure depicts that the maximum proportion of tweets belong to the general health-related schemes and the minimum proportion of tweets belong to the mother & child care schemes. The high frequency of health-related schemes is evident due to a more popular target domain and the success and popularity of schemes targeted to the common public, i.e., health insurance schemes such as Ayushman Yojana (PMJAY)[17]. An approximately 40% increase in the internet user population with 25% internet penetration in rural areas adds up to this result. Figure 6 also depicts that most of the tweets are generally promotional and informational, and complaint-related tweets are the least occurring of all.

Furthermore, a geo-plot analysis of the tweets is performed to illustrate the location-wise involvement of the public in posting tweets related to the government schemes. Figure 7a shows the heatmap geographical plot of all the tweets for the whole duration. The red regions in the heatmap represent the places with higher tweet concentrations. This is further fine-grained by extracting the Target group and Engagement type information for India's most active cities. Here, most active means the cities with the highest amount of tweets generation. Figure 7 depicts the cities with high tweet generation with Target group and Engagement type-wise distribution. The composition of different Target groups and user-engagement types are shown as

[17] `https://www.pmjay.gov.in`

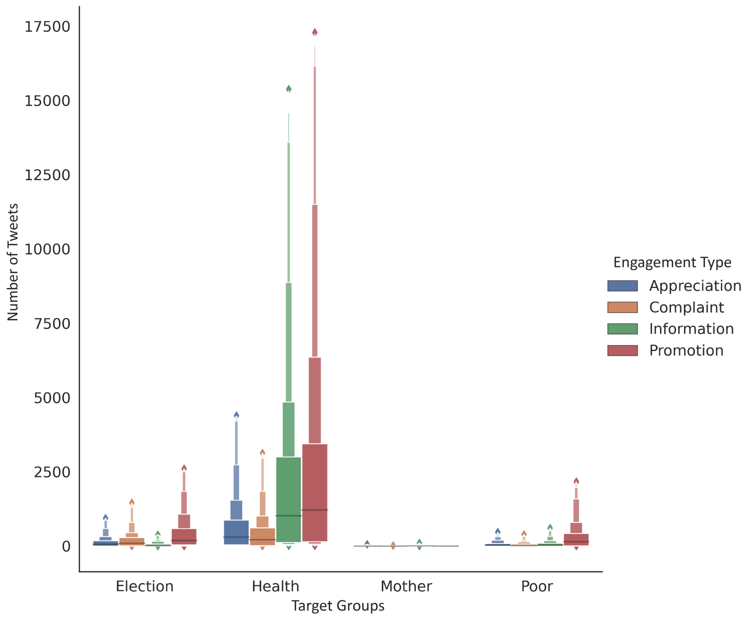
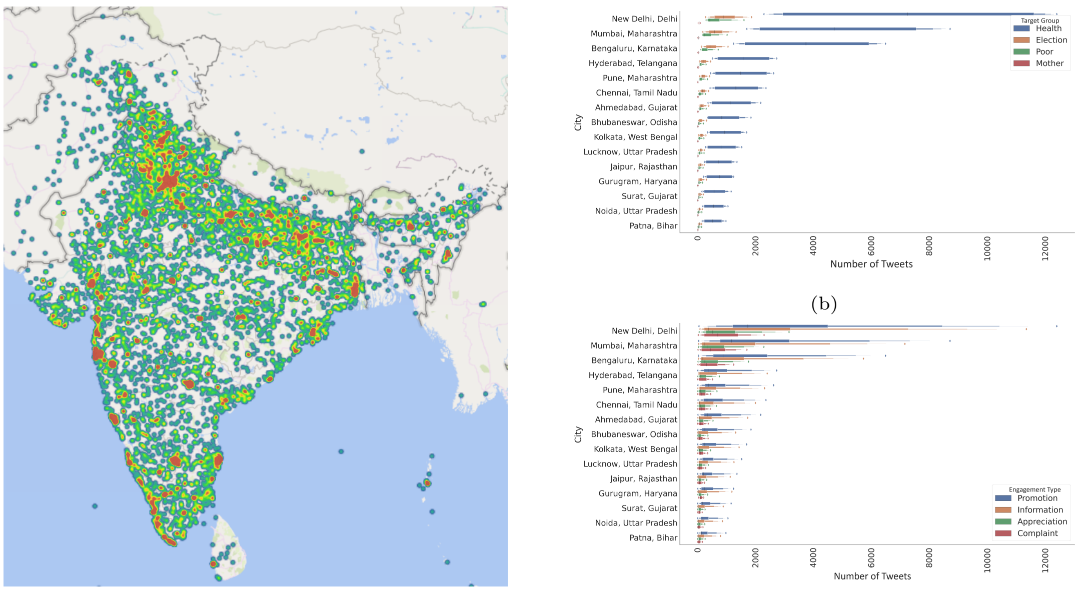


Figure 6: The box plot representation of number of tweets belonging to different types in different domains



(a) (c)

Figure 7: The geo-graphical analysis of the governement scheme related tweets all over India. (a) The heatmap plot representing the frequency of tweets on map of India. The red spots represent the areas with high tweet density. (b)The box plot representation for the target users in the most engaged cities all over India. (c) The box plot representation for the type of engagement in the most engaged cities all over India.

Table 6: A statistical description of the frequency of tweets for some time durations

	Total duration		Before Election		During Election		After Election	
	Hourly	Daily	Hourly	Daily	Hourly	Daily	Hourly	Daily
Average	121.886	2909.23	67.48	1606	206.60	4879	52.93	1399
Maximum	2778	19660	2690	7462	2778	19660	359	3128

the box plot representation for most active cities in Figure 7b and 7c respectively. The most popular cities, i.e., Delhi, Mumbai, Bangalore, Hyderabad, etc., are already the country's most populated metro cities. It can be seen from Figure 7b and 7c that the maximum proportion of tweets are related to general health and tend to have informational or promotional tweets. Next, we exploit the tweet classification and geo-parsing results to assess the relationship between the government scheme-related tweets and the elections.

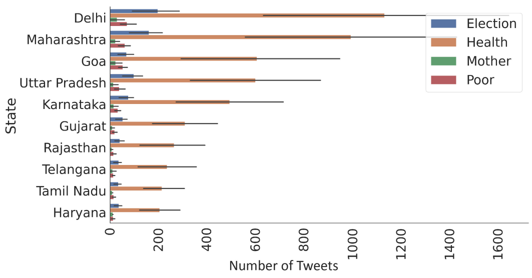
5.3. Government Scheme versus Electoral Dynamics

The overall geo-location and text classification analysis showed that the most highly populated cities generate the highest number of tweets. In this section, the effect of the voting phase on the government scheme-based Twitter data (Section 5.3.1) and the correlation with the electing parties (Section 5.3.3) have been discussed to get fine-grained insights on their relationship.

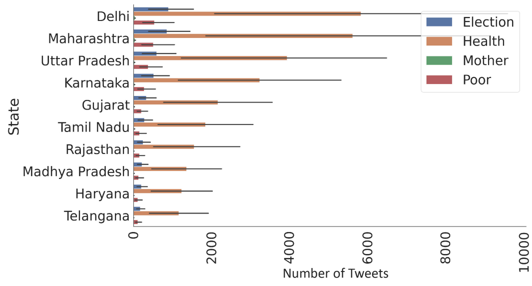
5.3.1. The Voting Phase Analysis

The government schemes Twitter data collection covered the period from December 2019 to November 2019, including the duration of elections (i.e., 1 Apr to 31 May 2019). The voting phase's effect on the government scheme tweets has been studied by splitting the Twitter data into three phases for three time windows. The value of the variables in Figure 3 are taken to be $d_e = 1$ Apr 2019, $\delta e = 2$ months (duration of election), $\delta t = 3$ months (duration of election with an additional month for analysis). Thus, these windows are *before elections* phase i.e., 1 Jan to 31 Mar (3 months), *during elections* phase i.e., 1 Apr to 29 May (2 months) and *after election* phase i.e., 1 Jun to 29 Aug (3 months). The time windows have been chosen to keep the data set size uniform.

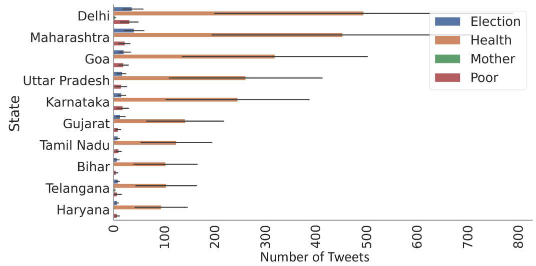
The elections were held in seven voting phases, i.e., 11 Apr, 18 Apr, 23 Apr, 29 Apr, 6 May, 12 May, and 19 May. The *during election* phase dataset has been further separated into seven chunks corresponding to the seven election phases for finer granularity. The time window has been chosen, such that it contains data from 2-3 days before and after the voting date. The seven chunks have been created for the following duration: 9 Apr-14 Apr, 15 Apr-20 Apr, 21 Apr-26 Apr, 27 Apr-2 May, 3 May-8 May, 9 May-16 May, and 17 May-22 May.



(a) Before election period



(b) During election period



(c) After election period

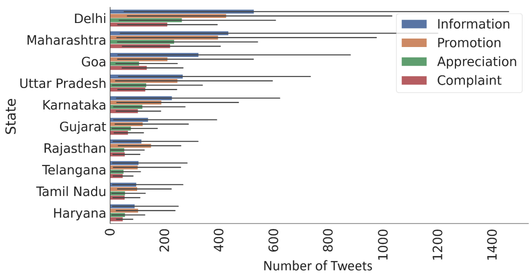
Figure 8: The number of tweets per target group type for the before election period, during election period and after election period.

5.3.2. Discussion

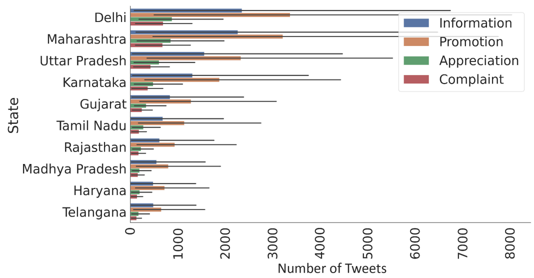
The average number of tweets shared hourly and daily was 121.886 and 2909.23, respectively, where the maximum reached up to 2778 tweets hourly and 19660 tweets daily. Table 6 provides hourly and daily tweets statistics for before election, during the election, and after the election phase. The table shows that the highest tweet frequency was in the during election phase, where the least was found in the after election phase. In this section, first the change in the concentration and spread of tweets with location in the *before, during* and *after* election phases (Figure 3) is studied. The increase and decrease in the number of tweets with election phases are analyzed in terms of Target groups and Engagement types.

As explained in Section 4.3.1, the tweet vector clustering helps to return the semantically similar tweets. The clusters with the highest population, i.e., clusters containing the highest number of tweets, represent the most shared similar content over the given duration. These clusters help to identify the most influential content in the given period in the form of bursts. The geo-parsing further helps to identify the locations which are most actively involved in sharing such content.

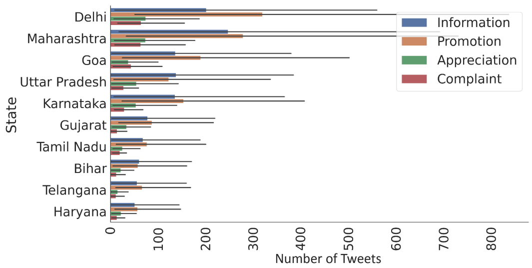
Figure 8 and 9 shows the most active states extracted from this study for *before, during* and *after* election phase for the different types of target groups and types of user-engagements, respectively. The states Delhi, Maharashtra, Uttar Pradesh, and Karnataka have remained on the top for all three phases. This implies



(a) Before election period



(b) During election period



(c) After election period

Figure 9: The number of tweets per user engagement type for the before election period, during election period and after election period.

that the metro cities and the highly populated cities generate the highest number of tweets. Most of the tweets are promotional and informational and focused on general health schemes. It can be seen from the subfigures a, b and c of Figure 8 and 9 that there is an increase in the average number of tweets from *before election* (maximum x-axis value 1600 in Figure 8a and 1400 in Figure 9a respectively) to *during election* (maximum x-axis value 10000 in Figure 8b and 8000 in Figure 9b respectively) phase and a decrease in the *after election* (maximum x-axis value 800 in Figure 8c and 800 in Figure 9c respectively) phase. Figure 8b shows that there is an increase (approx. 40%) in the tweets tagged *Election* in the *during election* phase but show a decline (approx. 200 %) in the *after election* phase (Figure 8c).

Figure 9a, shows that the *before election* phase has the highest average of *Informational* tweets, which then changes to the highest average of *Promotional* tweets in the *during election* phase. This depicts that during the election period, the government scheme-related information is promoted and distributed over the internet as the campaign process. The tweets are primarily focused on the schemes that focus the general public, e.g., general-health schemes such as Ayushman Yojna, and much less attention is being given to the smaller public domain, i.e., poor-public and mother & child healthcare. The study of change in tweeting pattern from before to during to after election phase implies that the government tends to focus on the schemes which have been the most popular and successful. This may lead to claiming the credits for the success during the elections.

The study is further concentrated on the *during election* phase to get a closer look at the different voting phases. Here, we study the change in the number/average of tweets just before and after different states' voting dates (phase). As the time duration between the first (11 Apr) and last (19 May) voting phase is approximately one month, a gap of one month has been considered to keep the time duration in *before and after* the voting date homogeneous for all the voting phases. For example, for Uttarakhand, which had its voting date on 11 Apr, the *just before phase* includes the tweets generated for/from Uttarakhand from 10 Mar to 11 Apr 2019. Similarly, the *just after phase* includes the tweets generated for/from Uttarakhand from 12 Apr to 11 May 2019. However, some states had polling/voting in a single-phase and others in multiple phases. For example, Uttar Pradesh had voted in seven phases, i.e., 11, 18, 23, 29 Apr, 6, 12, 19 May 2019 and Goa had voted in a single phase, i.e., 23 Apr 2019. So, this analysis has been separated into two parts, i.e., **(i)** analyzing the states having voted in a single phase, and **(ii)** analyzing the states having voted in multiple phases.

First, the tweeting pattern has been assessed for the states with only one voting phase. These states are Goa (23 April), Gujarat (23 April), Kerala (23 April), Andhra Pradesh (11 April), Uttarakhand (11 April), Telangana (11 April), Delhi (12 May), Haryana (12 May) and Himachal Pradesh (19 May). The north-eastern states have not been taken here because very few tweets were posted from these states, not enough to generalize the results. Figure 10 provides the bar plot representation of the number of tweets belonging to different schemes before and after the date of voting. The Figure depicts that there is a reduction of $\sim$ 63.10% (excluding Goa) and $\sim$ 37.23 % (including Goa) tweets on average for states Goa, Gujarat, and Kerala (except Goa), a reduction of $\sim$ 41.50 % tweets on average for states Andhra Pradesh and Telangana and a decrease of $\sim$ 58.40 % tweets on average for states Delhi, Haryana, Himachal, and Punjab. Only one state Goa has shown an unusual pattern showing an increase in the number of tweets after the polling. This implies that even though the propagation of government scheme-related tweets increases just before the voting date, it reduces drastically after the voting date. This result is homogeneous with the Center of Media Studies report- *Poll Expenditure, The 2019 Elections*, which states that "around 40 percent voters acknowledged receiving poll-related messages on their mobile phone just before the polling day." So, it can be said that the government scheme-related tweeting patterns show variation with the voting phases of the election.

Next, the states having polling in multiple voting phases have been investigated. The state-wise before and after voting phase analysis may result in erroneous outcomes because of the multiple voting phases. Thus, in this case, we study the most active states separately for each voting phase. Figure 11 shows the bar plot of the average number of tweets over the time duration of each voting phase $1 \leq i \leq 7$ i.e., $e_i - k_i$ to $e_{i+1} + k_{i+1}$ (refer to Section 4.3.1). This plot presents the most active states in these seven chunks, along with the type of user engagements. The error bars in Figure 11 indicate a 95% confidence interval around

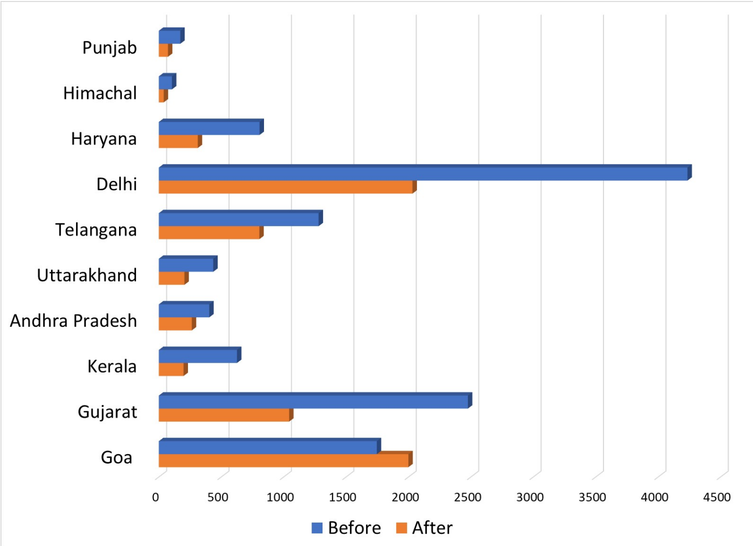
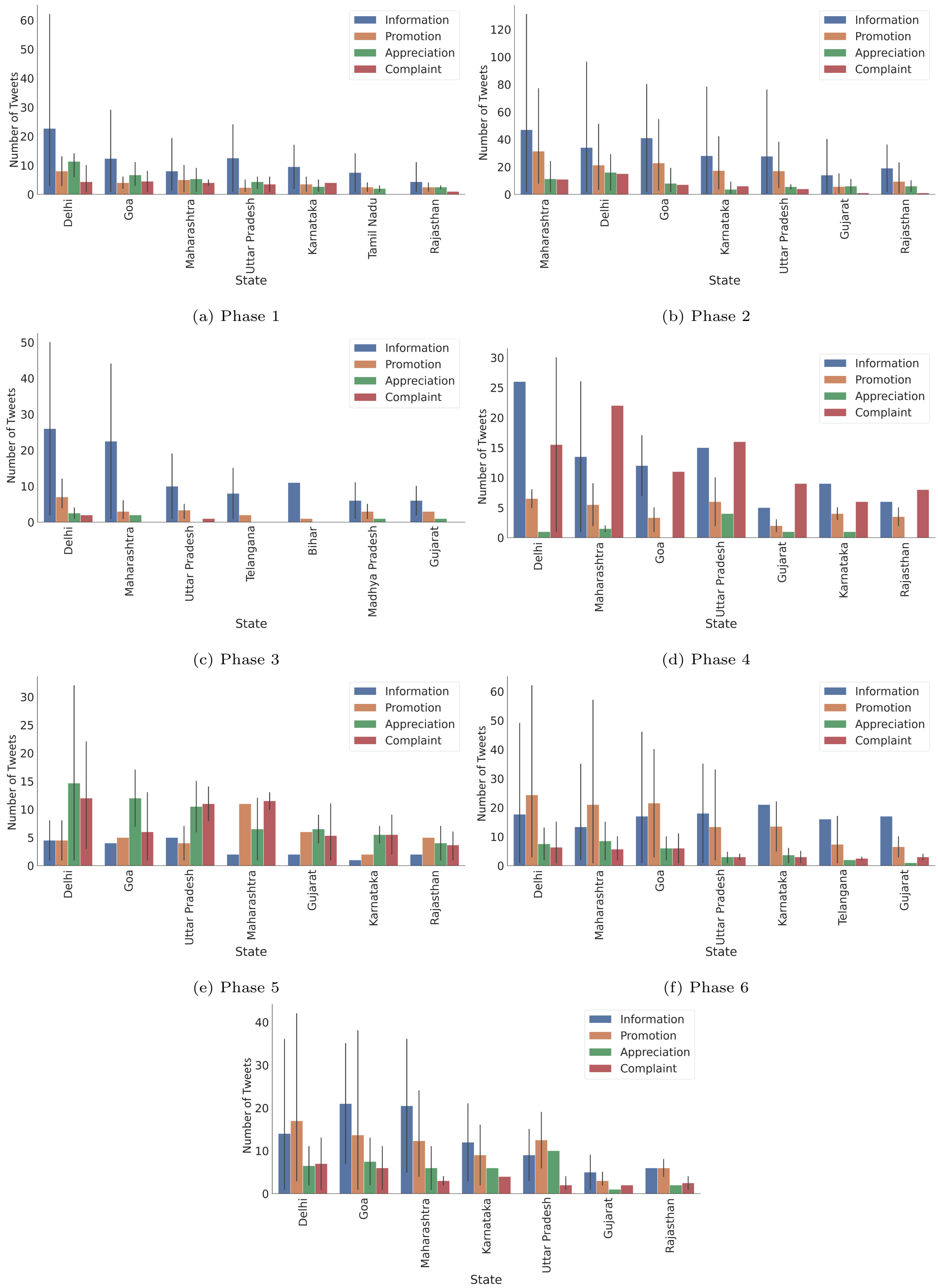


Figure 10: The bar plot representation of number of tweets before and after the date of election in the respective states having only one voting phase

the average number of tweets collected for a given location over the given duration. The error helps to reflect upon some level of inherent uncertainty in the range of data over the whole duration. A general hypothesis should be that the number of tweets should be affected by the voting phase's location, as visible in Figure 10. However, Figure 11 shows that the most active states in all the seven chunks generally remain the same, even though the polling may be happening in different locations.

Next, we explore the change in engagement type of the most active states in these seven phases. The user engagement levels show that in Phase 1 (Figure 11a), the tweets are more informational and appreciation based, and with time till Phase 3 (Figure 11c), the average of appreciation and complaint tweets reduce proportionally to the informational tweets. Phase 4 (Figure 11d) shows that there is an increment in the complaint tweets which is further overshadowed by the appreciation based tweets continuously in phase 5, 6 and 7 (Figure 11e, 11f and 11g). The promotional tweets are also increasing in the final phases. These results align with some previous studies that found that politicians sparsely talk about the policies but are more oriented towards promoting the campaign-related events and topics (Stier et al. (2018)).

It can be deduced from the above results that the states generating a higher number of tweets continue to generate higher amounts. On the other hand, the states generating fewer tweets continue to generate the lower amounts, only showing variations in peaks. However, the concentration of tweets does not change with the location of the voting phase. This shows that whatever be the source of the tweet generation, it spreads in equal proportions in all the states of India during the election duration. That is, the frequency of tweets changes synchronously in all states without regard to the voting location. This implies that the election date impacts the concentration of government scheme-related tweets, but it cannot be said that



(a) Phase 1 (b) Phase 2

(c) Phase 3 (d) Phase 4

(e) Phase 5 (f) Phase 6

(g) Phase 7

Figure 11: The user engagement in the seven voting phases of the most frequent states

Table 7: Correlation score between the state-wise positive, negative and neutral tweets and UPA and NDA seats won per state using the user-location of tweets

	NDA Tweets			UPA Tweets		
	Positive	Negative	Neutral	Positive	Negative	Neutral
NDA seats	**0.610201**	**0.305207**	**0.586651**	**-0.11386**	**0.412491**	**0.531352**
UPA seats	-0.28372	-0.3146	-0.41096	-0.21789	-0.21093	-0.22086
Total seats	0.465243	0.370769	0.515599	-0.12585	0.505683	0.515358
NDA %	0.465421	0.204371	0.364093	0.264336	0.003845	0.167563
UPA %	-0.41553	-0.54012	-0.59777	-0.25311	-0.42801	-0.36251

Table 8: Correlation score between the state-wise positive, negative and neutral tweets and UPA and NDA seats won per state using the location mention in the tweet text

	NDA Tweets			UPA Tweets		
	Positive	Negative	Neutral	Positive	Negative	Neutral
NDA seats	**0.251203**	**0.132715**	**0.455917**	-0.09321	0.202547	0.258623
UPA seats	0.094701	0.156766	-0.23426	**0.311827**	**0.09095**	**0.324703**
Total seats	0.181553	0.211143	0.295662	0.297461	0.02924	0.279676
NDA %	0.105955	0.085611	0.192474	-0.14103	0.215067	0.16078
UPA %	0.092275	0.115632	-0.28842	0.07049	-0.01137	-0.05702

this attention is specific to the polling location and shifts with the voting phase. Also, it can be stated that even if most of the tweets are generally informational or promotional, an increase in the complaint-related tweets induce an increase in the appreciation and promotional tweets in the later phases. This study can be extended further by studying the users who share such complaint-related and appreciation-related tweets to get deeper insights. However, it can be assumed that some misinformation could be spreading over the internet related to the elections or the related parties that were later rectified. This is supported by the fact that during 2019 general elections over 50,000 fake stories were shared over a 2 million times over the internet[18] [19].

5.3.3. Effects of the ruling and the electing parties

In this section, we study the correlation between the government schemes-related tweets and the election results in terms of the electing parties. The high correlation here would suggest that the government schemes-related tweets can also predict the election outcomes. In the 2019 Indian general elections, the

[18] `https://www.newindianexpress.com/states/karnataka/2019/oct/21/fake-news-shared-over-two-million-times-on\-social-media-during-lok-sabha-polls-2050676.html`

[19] `https://www.logically.ai/press/fighting-misinformation-in-the-biggest-election-on-planet-earth`

Table 9: Correlation score between the state-wise positive, negative and neutral tweets and UPA and NDA seats won per state using the combination of user-location and text mention of location

	NDA Tweets			UPA Tweets		
	Positive	Negative	Neutral	Positive	Negative	Neutral
NDA seats	**0.472765**	**0.390655**	**0.769929**	-0.161	-0.09277	-0.02056
UPA seats	-0.13125	0.077974	-0.15318	**0.421035**	**0.379985**	**0.502063**
Total seats	0.454807	0.397989	0.686728	0.269199	0.283455	0.570605
NDA %	0.259205	0.22191	0.307817	-0.22217	-0.12391	-0.1193
UPA %	-0.19181	-0.05749	-0.24572	0.428412	0.308787	0.118978

main alliances were NDA (BJP lead) and UPA (INC lead), with Narendra Modi and Rahul Gandhi as the party leaders, respectively. NDA won the election with 353 seats (303 seats by BJP) out of 545 total seats all over India. The studies showed that all the political parties spent approx 7.4 million dollars on the digital platforms for online ads where BJP[20] spent the most of it. The government schemes used in this study are also primarily launched by BJP. Thus, we want to study the correlation between the government schemes related tweets and the election results.

The results of Spearman correlation are given in Table 7, Table 8 and Table 9. The correlation results are also shown with the *percentage* of NDA and UPA seat winnings. Table 7 shows the results when only user-locations are used, Table 8 shows the results when locations from only text mentions are used, and Table 9 shows the results for the combination of both the locations. The tables show that the correlation score values improve with the use of a combination of user locations extracted from user-profile and text-mentions i.e., Table 9.

5.3.4. Discussion

Table 7, 8, and 9 show similar relationships among the NDA, UPA tweets and the corresponding election wins, where Table 9 provides values with higher polarities. In this section, the primary discussions are followed from Table 9. The primary hypothesis here is that the positive and neutral tweets for one electing party should be promoting that party, whereas the negative tweets may be demotivating this party. The negative tweets for one electing party, thus, may prove to be helpful in promoting the opposition party. This promotion and demotivation can be quantified in terms of the correlation scores. Here, we consider four possible combinations: NDA tweets-NDA wins, UPA tweets-UPA win, NDA tweets-UPA losses, and UPA tweets-NDA losses.

Table 9 shows that the NDA positive and NDA neutral tweets tend to show a high correlation with the NDA seats won per state. On the other hand, the correlation scores are weaker in the case of UPA

[20] `https://en.wikipedia.org/wiki/2019_Indian_general_election`

positive and neutral tweets with the UPA seat winnings. Now, we look at the inverse relationships, i.e., NDA tweets-UPA losses, and UPA tweets-NDA losses, i.e., the correlation between the seats won and the corresponding sentiments in the opposition parties. Table 7 and 8 show that UPA negative tweets have a positive correlation with the NDA seats won, and the UPA positive tweets have a negative correlation with the NDA seats won. Table 9 indicates that the NDA positive and neutral tweets have a much higher positive correlation with the NDA seats won and a negative correlation with the UPA seats won. Similarly, when we look at the UPA tweets, the correlation scores are very weak to suggest that they correlate with the NDA wins and losses. This implies that NDA government scheme tweets can have a relationship with the UPA and NDA wins and losses, but it is hard to establish the same for the UPA tweets and the UPA and NDA wins and losses.

As the current ruling party in 2019 was BJP and had launched the most popular government schemes mentioned in the study (e.g., AyushmanBharat Yojna), the government scheme-related tweets are more focused on the NDA alliance of BJP. It is also evident from the previous studies that BJP held a clear dominance over Twitter usage during the elections (Pal & Panda (2019)). So, it can be said that the government schemes related Twitter data can be used to assess the performance of the current ruling party, and the tweets may also create a positive influence for the current ruling party during the election duration. However, the same cannot be said regarding the competing parties, UPA, because of the weaker correlation.

6. Conclusion

Using social media data to assess information dissemination of government scheme-related user experiences promises better insights into citizens' perceptions and behaviors. Such schemes' success and failure can also affect the citizen stances towards the parties during election campaign promotions. This study employs deep learning embedding and classification tools on Twitter data to investigate the user types and patterns of information. Furthermore, the geolocation, cluster analysis, sentiment analysis and correlation analysis helped to identify the change in the scheme-related tweeting patterns during the election period. This study depicts that general health-based schemes are more trending than community targeting schemes such as maternal health-related schemes. The user engagements showed a changing trend with the start and end of election duration. It was seen that the tweeting patterns remained similar in all the portions of the country, and geolocation of voting phases did not affect the scheme-related tweeting patterns during these elections. We see that combining the location information from multiple sources, i.e., user-locations, text mentions, and geo-tags, provides the best correlation with the election results. The ruling party, i.e., BJP, showed dominance over scheme-related tweets more than any other electing party. The current study can be improved further by employing a more comprehensive range of emotions analysis in place of sentiment analysis of the tweets.

References

Addo, A., & Senyo, P. K. (2021). Advancing e-governance for development: Digital identification and its link to socioeconomic inclusion. *Government Information Quarterly*, *38*, 101568.

Agarwal, A., Toshniwal, D., & Bedi, J. (2019). Can twitter help to predict outcome of 2019 indian general election: A deep learning based study. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 38–53). Springer.

Aghababaei, S., & Makrehchi, M. (2018). Mining twitter data for crime trend prediction. *Intelligent Data Analysis*, *22*, 117–141.

Agostino, D. (2013). Using social media to engage citizens: A study of italian municipalities. *Public Relations Review*, *39*, 232–234.

Ahmed, S., Cho, J., & Jaidka, K. (2017). Leveling the playing field: The use of twitter by politicians during the 2014 indian general election campaign. *Telematics and Informatics*, *34*, 1377–1386.

Ahmed, S., Jaidka, K., & Cho, J. (2016). The 2014 indian elections on twitter: A comparison of campaign strategies of political parties. *Telematics and Informatics*, *33*, 1071–1087.

Alryalat, M. A. A., Rana, N. P., Sahu, G. P., Dwivedi, Y. K., & Tajvidi, M. (2017). Use of social media in citizen-centric electronic government services: A literature analysis. *International Journal of Electronic Government Research (IJEGR)*, *13*, 55–79.

Aneez, Z., T Neyazi, A., Kalogeropoulos, A., & Nielsen, R. K. (2019). *India Digital News Report*. Technical Report Reuters Institution.

Anstead, N., & O'Loughlin, B. (2015). Social media analysis and public opinion: The 2010 uk general election. *Journal of Computer-Mediated Communication*, *20*, 204–220.

Arthur, D., & Vassilvitskii, S. (2006). *k-means++: The advantages of careful seeding*. Technical Report Stanford.

Badawy, A., Ferrara, E., & Lerman, K. (2018). Analyzing the digital traces of political manipulation: the 2016 russian interference twitter campaign. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 258–265). IEEE.

Bekkers, V., Edwards, A., & de Kool, D. (2013). Social media monitoring: Responsive governance in the shadow of surveillance? *Government Information Quarterly*, *30*, 335–342.

Bertot, J. C., Jaeger, P. T., & Hansen, D. (2012). The impact of polices on government social media usage: Issues, challenges, and recommendations. *Government information quarterly*, *29*, 30–40.

Bertot, J. C., Jaeger, P. T., Munson, S., & Glaisyer, T. (2010). Social media technology and government transparency. *Computer*, *43*, 53–59.

Bjørnskov, C. (2010). How does social trust lead to better governance? an attempt to separate electoral and bureaucratic mechanisms. *Public Choice*, *144*, 323–346.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*, 993–1022.

Bloomberg (2019). Why india's election is among the world's most expensive. URL: `https://economictimes.indiatimes.com/news/elections/lok-sabha/india/why-indias-election-is-among-the-worlds-most-expensive/articleshow/68367262.cms`.

Bonsón, E., Royo, S., & Ratkai, M. (2015). Citizens' engagement on local governments' facebook sites. an empirical analysis: The impact of different media and content types in western europe. *Government information quarterly*, *32*, 52–62.

Bonsón, E., Royo, S., & Ratkai, M. (2017). Facebook practices in western european municipalities: An empirical analysis of activity and citizens' engagement. *Administration & Society*, *49*, 320–347.

Budiharto, W., & Meiliana, M. (2018). Prediction and analysis of indonesia presidential election from twitter using sentiment analysis. *Journal of Big Data*, *5*. doi:`10.1186/s40537-018-0164-1`.

Burnap, P., Gibson, R., Sloan, L., Southern, R., & Williams, M. (2016). 140 characters to victory?: Using twitter to predict the uk 2015 general election. *Electoral Studies*, *41*, 230–233.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C. et al. (2018). Universal sentence encoder for english. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 169–174).

Charalabidis, Y., Loukis, E. N., Androutsopoulou, A., Karkaletsis, V., & Triantafillou, A. (2014). Passive crowdsourcing in government using social media. *Transforming Government: People, Process and Policy*, *8*, 283–308.

Chauhan, P., Sharma, N., & Sikka, G. (2021). The emergence of social media data and sentiment analysis in election prediction. *Journal of Ambient Intelligence and Humanized Computing*, *12*, 2601–2627.

Criado, J. I., Sandoval-Almazan, R., & Gil-Garcia, J. R. (2013). Government innovation through social media.

De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. *Icwsm*, *13*, 1–10.

De Koster, W., Achterberg, P., & Van der Waal, J. (2013). The new right and the welfare state: The electoral relevance of welfare chauvinism and welfare populism in the netherlands. *International Political Science Review*, *34*, 3–20.

Derczynski, L., Albert-Lindqvist, T. O., Bendsen, M. V., Inie, N., Pedersen, V. D., & Pedersen, J. E. (2019). Misinformation on twitter during the danish national election: A case study. *Proceedings of the Conference for Truth and Trust Online 2019*, . doi:10.36370/tto.2019.16.

Deshpande, R., Tillin, L., & Kailash, K. (2019). The bjp's welfare schemes: Did they make a difference in the 2019 elections? *Studies in Indian Politics*, *7*, 219–233.

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

Dhiman, A., & Toshniwal, D. (2018). Do public and government think similar about indian cleanliness campaign? In *Annual International Symposium on Information Management and Big Data* (pp. 367–380). Springer.

Dhiman, A., & Toshniwal, D. (2020). An enhanced text classification to explore health based indian government policy tweets. In *ACM SIGKDD Workshop on Applied Data Science for Healthcare (DsHealth 2020*.

Driss, O. B., Mellouli, S., & Trabelsi, Z. (2019). From citizens to government policy-makers: Social media data analysis. *Government Information Quarterly*, *36*, 560–570.

Dwivedi, Y. K., Rana, N. P., Janssen, M., Lal, B., Williams, M. D., & Clement, M. (2017). An empirical validation of a unified model of electronic government adoption (umega). *Government Information Quarterly*, *34*, 211–230.

Ferrara, E., Chang, H., Chen, E., Muric, G., & Patel, J. (2020). Characterizing social media manipulation in the 2020 us presidential election. *First Monday*, .

Garimella, K., Morales, G. D. F., Gionis, A., & Mathioudakis, M. (2018). Quantifying controversy on social media. *ACM Transactions on Social Computing*, *1*, 3.

Garson, G. D. (2006). *Public information technology and e-governance: Managing the virtual state*. Jones & Bartlett Learning.

Gkotsis, G., Oellrich, A., Velupillai, S., Liakata, M., Hubbard, T. J., Dobson, R. J., & Dutta, R. (2017). Characterisation of mental health conditions in social media using informed deep learning. *Scientific reports*, *7*, 45141.

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on twitter during the 2016 us presidential election. *Science*, *363*, 374–378.

Grover, P., Kar, A. K., Dwivedi, Y. K., & Janssen, M. (2019). Polarization and acculturation in us election 2016 outcomes–can twitter analytics predict changes in voting preferences. *Technological Forecasting and Social Change*, *145*, 438–460.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*, 1735–1780.

Hu, Z., Yang, Z., Liang, X., Salakhutdinov, R., & Xing, E. P. (2017). Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70* (pp. 1587–1596). JMLR. org.

Hua, Y., Ristenpart, T., & Naaman, M. (2020). Towards measuring adversarial twitter interactions against candidates in the us midterm elections. In *Proceedings of the International AAAI Conference on Web and Social Media* (pp. 272–282). volume 14.

Huang, Q., & Wong, D. W. (2016). Activity patterns, socioeconomic status and urban spatial structure: what can social media data tell us? *International Journal of Geographical Information Science*, *30*, 1873–1898.

Hunter, W., & Power, T. J. (2007). Rewarding lula: Executive power, social policy, and the brazilian elections of 2006. *Latin American Politics and Society*, *49*, 1–30.

Imran, M., Ofli, F., Caragea, D., & Torralba, A. (2020). Using ai and social media multimodal content for disaster response and management: Opportunities, challenges, and future directions.

JACKSON, D. (2017). How obama used social media to win the 2012 elections. URL: `https://www.destinyjackson.org/blogs/articles-essays/how-did-the-obama-administration-use-social-media-to-win-the-2012-elections`.

Jia, R., & Liang, P. (2017). Adversarial examples for evaluating reading comprehension systems, .

Jiang, L., & Yang, C. C. (2017). User recommendation in healthcare social media by assessing user similarity in heterogeneous network. *Artificial intelligence in medicine*, *81*, 63–77.

Jungherr, A. (2016). Twitter use in election campaigns: A systematic literature review. *Journal of information technology & politics*, *13*, 72–91.

Kagan, V., Stevens, A., & Subrahmanian, V. (2015). Using twitter sentiment to forecast the 2013 pakistani election and the 2014 indian election. *IEEE Intelligent Systems*, *30*, 2–5.

Kamat, P. (2014). The obamafication of indian political campaigns. *South Asia@ LSE*, .

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, *30*, 81–93.

Kim, J., Bae, J., & Hastak, M. (2018). Emergency information diffusion on online social media during storm cindy in us. *International Journal of Information Management*, *40*, 153–165.

Kobayashi, S. (2018). Contextual augmentation: Data augmentation by words with paradigmatic relations, . (pp. 452–457).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Landsbergen, D. (2010). Government as part of the revolution: Using social media to achieve public goals. In *Proceedings of the European Conference on e-Government, ECEG* (pp. 243–250). volume 8.

Li, L., Zhang, Q., Tian, J., & Wang, H. (2018). Characterizing information propagation patterns in emergencies: A case study with yiliang earthquake. *International Journal of Information Management*, *38*, 34–41.

Li, Y., Kaneria, A., Zhao, X., & Manchaiah, V. (2019). Learning drivers' behavior using social networking service. In *International Conference on Applied Human Factors and Ergonomics* (pp. 341–350). Springer.

Liu, R., Xu, G., Jia, C., Ma, W., Wang, L., & Vosoughi, S. (2020). Data boost: Text data augmentation through reinforcement learning guided conditional generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 9031–9041). Online: Association for Computational Linguistics. URL: `https://aclanthology.org/2020.emnlp-main.726`. doi:`10.18653/v1/2020.emnlp-main.726`.

Lutz, M. (2009). The social pulpit: Barack obama's social media toolkit. *Retrieved February*, *6*, 2010.

Marzouki, A., Chouikh, A., Mellouli, S., & Haddad, R. (2021). From sustainable development goals to sustainable cities: A social media analysis for policy-making decision. *Sustainability*, *13*, 8136.

McCleary, R., Hay, R. A., Meidinger, E. E., & McDowall, D. (1980). *Applied time series analysis for the social sciences*. Sage Publications Beverly Hills, CA.

MeitY (2018). National e-governance plan: Ministry of electronics and information technology, government of india. URL: `https://www.meity.gov.in/divisions/national-e-governance-plan`.

MeitY (2020). Mission mode projects: Ministry of electronics and information technology, government of india. URL: `https:`

//www.meity.gov.in/content/mission-mode-projects.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, .

Miller, D. T. (2019). Topics and emotions in russian twitter propaganda. *First Monday*, . doi:10.5210/fm.v24i5.9638.

Mishra, J., & Attri, V. (2020). Governance, public service delivery and trust in government. *Studies in Indian Politics*, *8*, 186–202.

NDTV (2019). Pm modi receives global goalkeeper award for swachh bharat mission, dedicates it to 1.3 billion indians: News. URL: https://swachhindia.ndtv.com/pm-modi-receives-global-goalkeeper-award-for-swachh-bharat-mission-dedicates-it-to-1-3-billion-indians-38352/.

Nelimarkka, M., Laaksonen, S.-M., Tuokko, M., & Valkonen, T. (2020). Platformed interactions: how social media platforms relate to candidate–constituent interaction during finnish 2015 election campaigning. *Social Media+ Society*, *6*, 2056305120903856.

Nielsen, F. Å. (2011). Afinn. URL: http://www2.compute.dtu.dk/pubdb/pubs/6010-full.html.

Nielsen, F. Å. (2017). afinn project, .

Niu, T., & Bansal, M. (2018). Adversarial over-sensitivity and over-stability strategies for dialogue models. In *Proceedings of the 22nd Conference on Computational Natural Language Learning* (pp. 486–496). Brussels, Belgium: Association for Computational Linguistics.

Pal, J., & Panda, A. (2019). Twitter in the 2019 indian general elections: Trends of use across states and parties. *Econonmic and Political Weekly*, *54*.

Papadaki, M. (2017). Data augmentation techniques for legal text analytics. *Department of Computer Science, Athens University of Economics and Business, Athens*, .

Pedro-Carañana, J., Alvarado-Vivas, S., & López-López, J. S. (2020). Agenda-setting and power relations during the 2018 colombian election campaign on twitter. *The Journal of International Communication*, *26*, 260–280. doi:10.1080/13216597.2020.1806900.

Pennington, J. (). URL: https://nlp.stanford.edu/projects/glove/.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543).

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, .

Plasser, F., & Plasser, G. (2002). *Global political campaigning: A worldwide analysis of campaign professionals and their practices*. Greenwood Publishing Group.

Preotiuc-Pietro, D., Carpenter, J., Giorgi, S., & Ungar, L. (2016). Studying the dark triad of personality through twitter behavior. In *Proceedings of the 25th ACM international on conference on information and knowledge management* (pp. 761–770).

Rana, N. P., Dwivedi, Y. K., & Williams, M. D. (2015). A meta-analysis of existing research on citizen adoption of e-government. *Information Systems Frontiers*, *17*, 547–563.

Rao, A. (2019). How did social media impact india's 2019 general election? *Econonmic and Political Weekly*, *54*.

Sedgwick, P. (2012). Pearson's correlation coefficient. *Bmj*, *345*, e4483.

Seethaler, J., & Melischek, G. (2019). Twitter as a tool for agenda building in election campaigns? the case of austria. *Journalism*, *20*, 1087–1107. doi:10.1177/1464884919845460.

Severo, M., Feredj, A., & Romele, A. (2016). Soft data and public policy: Can social media offer alternatives to official statistics in urban policymaking? *Policy & Internet*, *8*, 354–372.

Shorten, C., Khoshgoftaar, T. M., & Furht, B. (2021). Text data augmentation for deep learning. *Journal of big Data*, *8*, 1–34.

Singh, P., Dwivedi, Y. K., Kahlon, K. S., Pathania, A., & Sawhney, R. S. (2020a). Can twitter analytics predict election outcome? an insight from 2017 punjab assembly elections. *Government Information Quarterly*, *37*, 101444. doi:10.1016/j.giq.2019.101444.

Singh, P., Dwivedi, Y. K., Kahlon, K. S., Sawhney, R. S., Alalwan, A. A., & Rana, N. P. (2020b). Smart monitoring and controlling of government policies using social media and cloud computing. *Information Systems Frontiers*, *22*, 315–337.

Socher, R., Bauer, J., Manning, C. D., & Ng, A. Y. (2013). Parsing with compositional vector grammars. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 455–465).

Srinivasan, S. M., Sangwan, R. S., Neill, C. J., & Zu, T. (2019). Twitter data for predicting election results: Insights from emotion classification. *IEEE Technology and Society Magazine*, *38*, 58–63.

Srivastava, R., Kumar, H., Bhatia, M. P. S., & Jain, S. (2015). Analyzing delhi assembly election 2015 using textual content of social network. In *Proceedings of the Sixth International Conference on Computer and Communication Technology 2015* ICCCT '15 (p. 78–85). New York, NY, USA: Association for Computing Machinery. URL: https://doi.org/10.1145/2818567.2818582. doi:10.1145/2818567.2818582.

Stamati, T., Papadopoulos, T., & Anagnostopoulos, D. (2015). Social media for openness and accountability in the public sector: Cases in the greek context. *Government Information Quarterly*, *32*, 12–29.

Stier, S., Bleier, A., Lietz, H., & Strohmaier, M. (2018). Election campaigning on social media: Politicians, audiences, and the mediation of political communication on facebook and twitter. *Political communication*, *35*, 50–74.

Sun, H.-L., Ch'Ng, E., & See, S. (2019). Influential spreaders in the political twitter sphere of the 2013 malaysian general election. *Industrial Management & Data Systems*, *119*, 54–68.

Tillin, L., & Pereira, A. W. (2017). Federalism, multi-level elections and social policy in brazil and india. *Commonwealth & Comparative Politics*, *55*, 328–352.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Fourth international AAAI conference on weblogs and social media*. Citeseer.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008).

Wei, J., Huang, C., Vosoughi, S., & Wei, J. (2020). What are people asking about covid-19? a question classification dataset. *arXiv preprint arXiv:2005.12522*, .

Wei, J. W., & Zou, K. (2019). Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, .

Weller, K., Bruns, A., Burgess, J., Mahrt, M., & Puschmann, C. (2014). *Twitter and society [Digital Formations, Volume 89]*. Peter Lang Publishing.

WHO (2018). India has achieved groundbreaking success in reducing maternal mortality. URL: https://www.who.int/southeastasia/news/detail/10-06-2018-india-has-achieved-groundbreaking-success-in-reducing-maternal-mortality.

Xie, Z., Wang, S. I., Li, J., Lévy, D., Nie, A., Jurafsky, D., & Ng, A. Y. (2017). Data noising as smoothing in neural network language models. *Proceedings of 2017 International Conference on Learning Representations (ICLR)*, .

Yang, X., Macdonald, C., & Ounis, I. (2018). Using word embeddings in twitter election classification. *Information Retrieval Journal*, *21*, 183–207.

Yaqub, U., Sharma, N., Pabreja, R., Chun, S. A., Atluri, V., & Vaidya, J. (2020). Location-based sentiment analyses and visualization of twitter election data. *Digital Government: Research and Practice*, *1*, 1–19.

Zar, J. H. (2005). Spearman rank correlation. *Encyclopedia of Biostatistics*, *7*.

Zhao, R., Chen, K., Norouzi, M., & Le, Q. V. (). Qanet: Combining local convolution with global self-attention for reading comprehension, .

Zhuravskaya, E., Petrova, M., & Enikolopov, R. (2020). Political effects of the internet and social media. *Annual Review of Economics*, *12*.

Zinnbauer, D. (2015). Crowdsourced corruption reporting: What petrified forests, street music, bath towels, and the taxman can tell us about the prospects for its future. *Policy & Internet*, *7*, 1–24.

Zucco Jr, C. (2013). When payouts pay off: Conditional cash transfers and voting behavior in brazil 2002–10. *American journal of political science*, *57*, 810–822.