



**Trinity College Dublin**  
Coláiste na Tríonóide, Baile Átha Cliath  
The University of Dublin

**School of Computer Science and Statistics**

## Assessment Submission Form

<b>Student Name</b>	MUKESH ARAMBAKAM
<b>Student ID Number</b>	19301497
<b>Course Title</b>	MSc. COMPUTER SCIENCE – DATA SCIENCE
<b>Module Title</b>	CS7IS4 - Text Analytics
<b>Lecturer(s)</b>	Dr Carl Vogel
<b>Assessment Title</b>	Final Essay
<b>Date Submitted</b>	11-04-2020
<b>Word Count</b>	3644

I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at: <http://www.tcd.ie/calendar>

I have also completed the Online Tutorial on avoiding plagiarism 'Ready, Steady, Write', located at <http://tcd-ie.libguides.com/plagiarism/ready-steady-write>

I declare that the assignment being submitted represents my own work and has not been taken from the work of others save where appropriately referenced in the body of the assignment.

Signed .....MUKESH ARAMBAKAM..... Date .....11-04-2020.....

## Author Declaration for Group Assignments

Group Number: 7

Module Number: CS7IS4

Title of Assignment: Final Essay

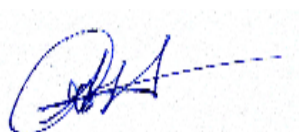
Student Number	Student Name	Nature of Contribution	Percentage contribution
19301497	Mukesh Arambakam	Worked on Data collection, Code Development and worked on the Design Methodology section. Helped define the process of data collection and processing.	21.5
19300364	Rajat Gogna	Worked on Data Collection and result visualization. Worked on Abstract, Introduction, Research Objective, Report Structure & Results sections. Helped define the process of data collection & processing.	21.5
19303333	Snehal Dey	Worked on the literature survey section. Helped in finding the hashtags used for election 2020.	21
17304308	Yang Liu	Worked on the Background Research and analysis of emotion in research papers.	15
19301303	Yeshwanth Rajareddy	Worked on the literature survey section. Helped in finding the candidate details. Worked on Data Collection.	21

We have read and we understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at <http://www.tcd.ie/calendar>.

We have also completed the Online Tutorial on avoiding plagiarism 'Ready, Steady, Write', located at <http://tcd-ie.libguides.com/plagiarism/ready-steady-write>.

We declare that this assignment, together with any supporting artefact is offered for assessment as our original and unaided work, except in so far as any advice and/or assistance from any other named person in preparing it and any reference material used are duly and appropriately acknowledged. We declare that the percentage contribution by each member as stated above has been agreed by all members of the group and reflects the actual contribution of the group members.

### Signed and dated:



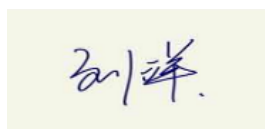
Mukesh Arambakam



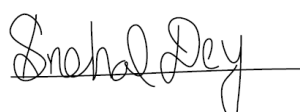
Rajat Gogna



Yeshwanth RajaReddy



Liu Yang



Snehal Dey

Signed on April 11th

# Creating Election Opinion poll using Twitter

Rajat Gogna  
gognar@tcd.ie

Mukesh Arambakam  
arambakm@tcd.ie

Snehal Dey  
sdey@tcd.ie

Yeshwanth RajaReddy  
rajareddy@tcd.ie

Yang Liu  
yliu8@tcd.ie

## Abstract

Analysing opinion trends from the micro-blogging site, twitter, and predicting the success of a product or a campaign has been the talk of the town for some time in the data analytics field. We have developed a method for analysing tweets suggesting the political sentiment of the users for the contenders in an election and correlating it with the actual results. Large amount of research has been done on this topic and we strive to grow our research on top of that. Our hypothesis is that **twitter popularity of political parties has a positive relation to the actual election results**. The null hypothesis is that there is no relation or a negative relation between a political party's twitter popularity and the election results. We are taking into consideration the recent General Elections in Ireland (2020) to validate our methodology. We have extracted tweets posted in Ireland for a period of two months before the elections. Tweets from known party accounts and candidate accounts have been dropped from our text corpus. The results of our research are in agreement with the actual election outcome.

**Keywords** - Text Analytics, Sentiment Analysis, Machine Learning, Opinion Poll, Twitter, Natural Language Processing, Social Media

## 1 Introduction

In this paper we discuss an application of text analytics on the micro-blogging website, twitter. In particular, we are analysing tweets of users during the time of elections in a particular region. We are using tweets from General Elections held in Ireland in 2020. Sentiment analysis was carried out on the collected tweets and user opinions were categorized to convey the political sentiment of the user.

Twitter is the most popular micro-blogging website. People express their opinion on all sorts of things on twitter. Since the time it was introduced, 21-Mar-2006, the user base of twitter has grown to 330 million active monthly users posting 500 million tweets a day <sup>1</sup>, making it one of the most valuable sources of data in today's time. To utilize data from twitter, one has to request a special account called the developer account. Upon approval, user gets access to twitter APIs for extracting tweets from Twitter's database based on various filters available.

Sentiment analysis is a field of study, where a medium of expressing opinion, for our cause, text, is analyzed and processed computationally to identify and categorise the user's attitude towards a topic. We have used this technique to identify twitter user's political opinion which in turn will be used to derive general trends from our tweet text corpus.

Our goal is to create an election opinion poll based on the tweets posted by users. To achieve this, we will derive and evaluate the trend of political sentiment of twitter users and predict which candidate or political party has a higher chance of winning the elections. The predictions of our model will be evaluated against the actual election results.

---

<sup>1</sup><https://ie.oberlo.com/blog/twitter-statistics>

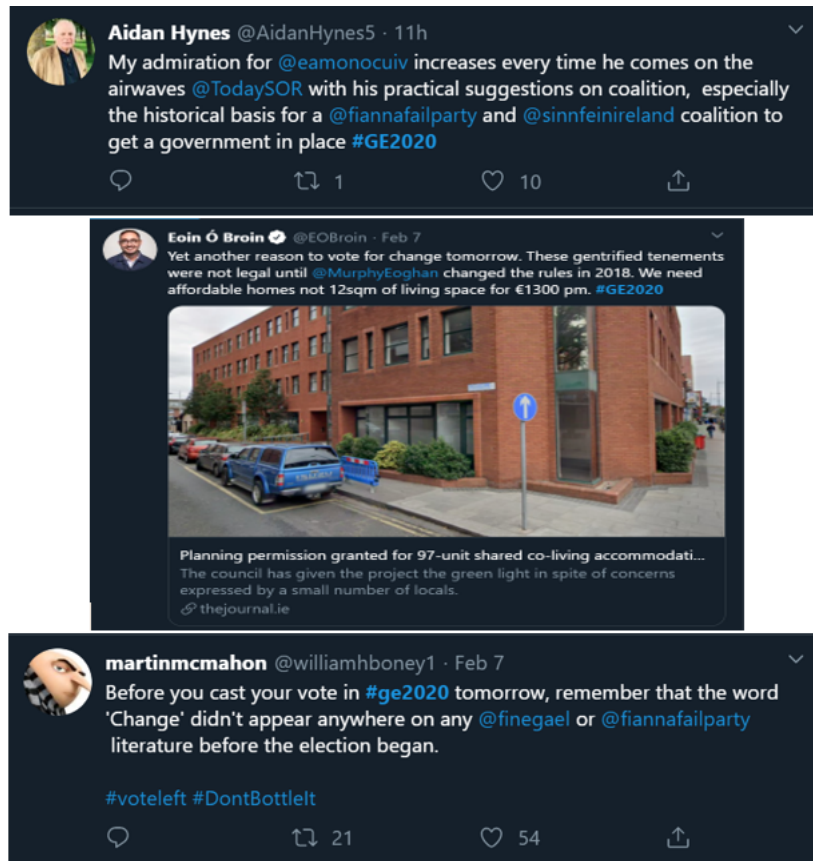


Fig 1.1: Examples of Tweets Showing Political Opinion of Users

The language in tweets is not formally written text. Tweets are generally informal in nature with common use of slang, images, emoticons etc., and hence pre-processing of data to create a usable text corpus is a major step in our research.

## 1.1 Research Question

In this research paper, we are trying to create an opinion poll based on tweets related to political parties and candidates. We will capture tweets posted in the period of three months before the elections and compare the observed trends with the actual election results. The question we are trying to answer through our research is:

**If an opinion poll created using public tweets could be considered as a strong indicator of the actual election results.**

Apart from our very evident main objective of validating the reliability of an election opinion poll created using twitter, our research entails various other objectives which need to be fulfilled in order to achieve our primary goal.

The following section and its sub-sections talk about these objectives in detail.

## 1.2 Research Objective

In our project<sup>2</sup> we are focusing on tweets depicting a user's political inclination for the elections under consideration. Irish General Elections were held on 8<sup>th</sup> of February, 2020 and we are in the process of extracting tweets posted between the period of 8-Nov-2019 to 8-Feb-2020. These tweets will be used as our data-set after some pre-processing. Following are our research objectives in detail:

<sup>2</sup><https://github.com/rgog/TwitterElectionTrends>

## 1. Data Collection

Our first objective is to extract relevant tweets from the twitter database. We have to collect tweets based on the following conditions:

- Geo-Location - Tweets posted from users in the Republic of Ireland.
- Language - Tweets posted in English language.
- Keywords - Learning from Bovet et al. (2018), we will filter tweets on keywords of leading political parties and their leaders.

## 2. Data Processing

The data we collect needs to be processed further so that it can be utilized for performing further analysis. The text corpus generated in the data collection step now needs to be categorized into different labels denoting each of the top election candidates. For our cause, the following categories will be used:

Category	Keywords
Fine Gael	FineGael, LeoVaradkar
Sinn Féin	SinnFein, MaryLouMcDonald
Fianna Fáil	FiannaFail, MichealMartin

Table 1.1: Keywords Used to Collect Tweets

Apart from this other text formatting techniques like Stemming, Case-Handling, etc also need to be performed.

## 3. Sentiment Analysis

After categorizing our data under the required labels, our next objective is to analyse the data to identify trends. Here we need to perform sentiment analysis on tweets of each category and calculate a custom parameter called "Positive Sentiment Score". This score will lie between 0-1 and will denote the ratio of tweets showing positive sentiment towards a political party with respect to the total number of tweets posted for it. Here, our goal is to analyse the overall sentiment of the people towards a political party, considering only those users who have an opinion on that party. We also find the percentage of positive tweets for each political party with respect to the total number of positive tweets.

## 4. Aggregation of Results

Another objective of this research is to compare the positive sentiment score to the actual election result to see if the opinion poll created with twitter data, is a reliable indicator of the actual election results and if it can be considered as a probable predictor of results in the future. We will also compare the contribution of each political party to the positive sentiment tweets to see if there is any correlation with the final results.

## 1.3 Report Structure

We have prepared this report in a structured format. Each section talks about a different aspect of our research. Following is a chapter-wise description of the report:

- **Chapter 1: Introduction**

This is the basic introduction to our research. It gives a brief discourse of our research question and objectives.

- **Chapter 2: Literature Review**

This chapter includes a review of various studies done in the past, which we have referred while doing our research.

- **Chapter 3: Design Methodology**

This chapter explains our basic design methodology of conducting this research. Here we describe the choices that we made in each phase of our research and why we took those choices.

- **Chapter 4: Conclusion**

Here we present the results of our research in a quantitative way with a final conclusion. In one section we summarize our results and in the other, future work and shortcomings.

## 2 Literature Review

In this section we will discuss some of the academic research that motivated our work and inspired some techniques we used in our analysis.

Lei et al. (2016) proposed analysis of social sentiments from restaurant reviews. This method uses Latent Dirichlet Allocation (LDA) – a Bayesian model to find correlation between reviews, topics and words. The sentiments of the items are calculated using the HowNet dictionary. The words used in tweets are categorised as sentiment dictionary, negation dictionary, and sentiment degree dictionary. The best part of this research is that the method filters stop words such as articles, prepositions, pronouns etc. It extracts feature based on the frequencies of words in corpus and considers similarity sentiments like influence of friend's reviews on others, influence of interpersonal sentiment and influence of the best and the worst ratings on other items, to identify the most reputed dish.

Belcastro et al. (2020) proposed Iterative Opinion Mining using Neural network (IOM-NN). The technique is used to predict the public opinion based on hashtags or words that exist in tweets. The method was implemented on USA presidential election and Italian general election and in both the case, the winning candidate was predicted with a good accuracy. Belcastro et al. (2020) showed IOM-NN based polarized algorithm, that filters data based on the number of parameters the user's tweets must satisfy, improved the accuracy to a great extent. Unlike this technique which uses neural network to build the model, we use the Naïve Bayes machine learning method for classifying the polarities.

Text mining and Natural Language processing techniques widely used for information analysis with small and Big data. Franco-Riquelme et al. (2019) proposed sentimental analysis on Spanish 2015 and 2016 general election to measure the support regarding the political parties. The research uses linguaKit to measure the polarity of the tweets, the analysis was made on the 250,000 tweets. Findings indicate consistent support towards party and optimists. The interesting part of this analysis is that the evaluations are made for 30 days before and 20 days after election data to measure the sentiments on the user tweets. Furthermore, in the past decade social media analysis has undergone a large development in response to tweets. Research carried out by Joyce and Deng (2017) showed a correlation between sentiment and hashtags. This research showed that the positive emotion on one candidate, might express negative emotion on the other candidate. The research is based on lexicon corpus and NLTK. The results showed that the hashtags can convey a strong correlation of sentiment towards the election candidate.

Jose and Chooralil (2015) discussed how sentiment analysis can be performed using the Lexicon based approach. The lexicon analysis mainly contains dictionary based and corpus based approaches. In order to assimilate semantics in their work, the authors used 'SentiWordNet' which is a collection of sentiment organised words along with Word Sense Disambiguation which relies on lexical resources of the WordNet as well as of the SentiWordNet. WordNet is an English language lexical database, used to distinguish between Nouns, Verbs, Adjectives, and Adverbs. It classifies words written in English into a set of synonyms. Jose and Chooralil (2015) extended his research to assign the numerical sentiment scores, these scores are categorized as positive, negative or neutral.

A similar kind of work was done in this field by Sharma and Moh (2016). They predicted the result of Indian Election of 2016 based on Hindi tweets. They used both supervised and unsupervised learning techniques. Their classifier was based on Dictionary Based, Naive Bayes and SVM algorithm. They classified the sentiments as positive, negative or neutral. A total of 42,235 tweets were mined and it was found out that SVM had predicted a 78.4% chance that the BJP (Bharatiya Janata Party) would win the election due to the positive sentiments they received in the tweets.

Research done by Choy et al. (2011) on sentiment analysis of Singapore Presidential Election 2011 using Twitter data predicted the share of votes for each electoral candidate. He suggested that with proper re-calibration with the help of census information, tweets can provide quite accurate information regarding the political scenario even though the twitter users are not as common.

### 3 Design Methodology

In this sections we will be talking about the design methodology steps involved in the project.

1. Data Collection.
2. Data Prepossessing.
3. Data Classification (Sentiment Analysis).
4. Aggregation of Results.

#### 3.1 Data Collection

The data was collected using the Twitter Search API<sup>3</sup> (Application Programming Interface) made available to users who have a Twitter Developer Account<sup>4</sup>. Twitter makes two kinds of API available to the public for free to access their data:

1. Search API - this allows you to find historical tweets based on various criteria.
2. Streaming API - this allows you to stream real-time tweets.

The tweets used in this project were collected using the Search API provided by twitter. A location filter was applied on this API so that only tweets posted from within the bounding area of Republic of Ireland were returned. A date filter was also applied to fetch historical data from within the dates of interest for the project i.e., from 8th November 2019 till 8th February 2020. A language filter was also placed into effect to avoid tweets from Gaelic or any other language. Tweets were retrieved based on keywords. A curated list of all the parties in Ireland and their presidents was compiled, along with the hashtags used for the elections and election campaigns. These were used as keywords. These tweets were stored as a CSV (Comma-Separated Values) file locally. So that they can be used in the future while training the model or for reference.

#### 3.2 Data Prepossessing

The set of gathered tweets was pre-processed before it was used for training the classifier. Taking ideas from Dickson (2019) the following pre-processing steps were applied on the data-set before training:

- **Lower-Case:**

All the text was converted into lower case by default so that words in different cases get recognised as the same while processing.

---

<sup>3</sup><https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets>

<sup>4</sup><https://developer.twitter.com/>

- **Punctuation:**

Special characters and punctuation's present in the tweets were removed using regular expressions. Some examples of these special characters are: . , ? ! ( ) & \* ; : .

- **Stemming:**

Stemming or Lemmatization in more specific was performed on the generated tokens using the Word Net lemmatizer from the Natural Language Toolkit (NLTK)<sup>5</sup>. Lemmatization is the process of grouping together words so they can be analysed as a single item. For example words like 'give', 'gave', 'giving', 'given' get lemmatized as one word 'give'. This reduces the number of words fed to the classifier.

- **Stopwords:**

Stopwords in the tweets were removed. All tweets (in general text) contain certain words like 'and', 'I', 'are', 'the', 'because', etc, which are extraneous and do not add any additional value to the sentiment of the tweet, these stopwords are removed when tokenizing the tweet.

- **Emojis:**

Emojis in the tweets were replaced with corresponding alias using Python's Emoji Library<sup>6</sup>. Ex - a smile emoji will be replaced with ':smile:' and then to 'smile' once punctuation are removed.

- **Case Change:**

Tweets contain words like 'GeneralElections', these words could also be split into 'General' and 'Elections'. This is something that is required because a lot of hashtags used in twitter has words of this type. And these words need to be processed to understand the context behind them.

- **Word-Digit Boundaries:**

Certain tweets also contain texts or hashtags like 'GeneralElections2020', these words need to be converted to 'General Elections 2020' to understand their contextual meaning.

### 3.3 Sentiment Analysis

We have obtained a corpus of 10,000 tweets, 5,000 positive and 5,000 negative, from the nltk library. These tweets were pre-processed using the above mentioned methods and the final cleaned and lemmatized tokens were obtained for each tweet. These tokenized tweets were used to train the classifier.

Here we have used the Gaussian Naive Bayes Classifier to train the model, which works by applying the Bayes Rule along with the assumption that the features of the dataset are independent.

The Bayes Rule is as follows:

$$P(H|E) = \frac{P(E|H) * P(H)}{P(E)} \quad (1)$$

Gaussian Naive Bayes Classifier assumes that the likelihood function is Gaussian:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} * \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right) \quad (2)$$

We have used Scikit-Learn's Gaussian Naive Bayes Classifier to train the model.

### 3.4 Aggregation of Results

Once we have the trained model, we pre-processed our data-set collected from twitter about the elections and we classified the model. The overall positive sentiment score for all the categories, i.e., the percentage of tweets with positive sentiment towards the party/candidate, considering the total number of tweets

---

<sup>5</sup><https://pypi.org/project/nltk/>

<sup>6</sup><https://pypi.org/project/emoji/>



in our text corpus was calculated. This result will be compared to the actual election results to see if the opinion poll created with twitter data, is a reliable indicator of the actual election results and if it can be considered as a probable predictor of results in the future.

## 4 Conclusion

In this section we will talk about the results of our research and what we believe can be good projects in the future to enhance our research.

### 4.1 Results

In this paper we had various objectives that we needed to fulfil in order to answer our research question. we will talk about them in the following points:

- **Data Collection** - Our objective to assess if data of good quality and quantity could be retrieved from the twitter database according to our requirements was successfully accomplished. We retrieved around 250,000 tweets for a period of two months, between 08-Dec-2019 and 08-Feb-2020.
- **Data Processing** - Our next objective, to filter tweets further and perform pre-processing on them to use with nltk sentiment libraries was also completed successfully. After this, we had 219,553 tweets left in our final dataset.
- **Sentiment Analysis** - Our goal of categorizing tweets into categories based on their sentiment, i.e., user opinion on the leading political parties or candidates has been successfully accomplished.

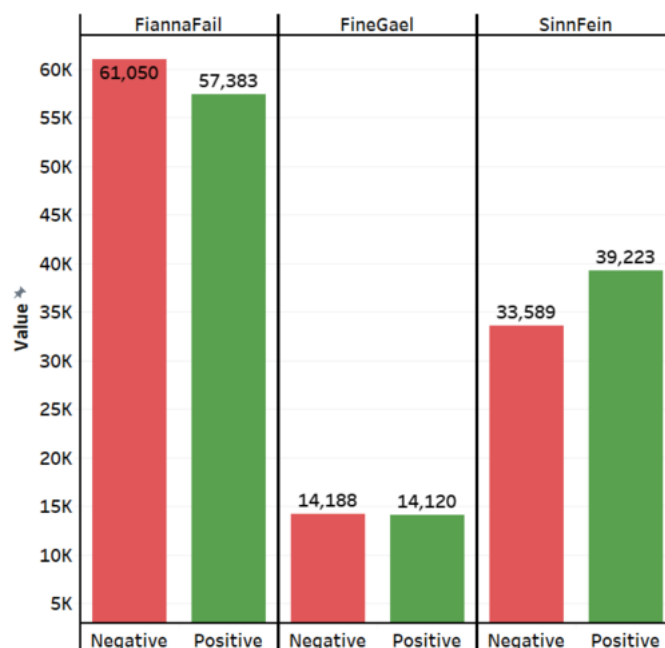


Fig 1.2: Number of Tweets with Positive and Negative Sentiment for each Political Party.

From Fig 1.2, we can calculate positive sentiment score (PSS) (Positive Tweets/Total Tweets) for each of the three parties -

Fianna Fail - 0.484

Fine Gael - 0.50

Sinn Fein - 0.54

From the PSS shown above, we can say that out of all the tweets posted for each of the contending political parties, more people had a positive opinion than negative for Sinn Fein, opinion of people on FineGael was more or less equally divided while more people had a negative opinion than positive, about Fianna Fail.

Further, we will see, share of each party in the total positive tweets posted..

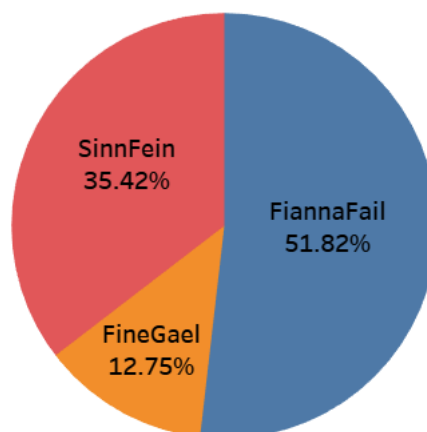


Fig 1.3: Percentage of positive tweets for each political party out of the total positive tweets.

Here, we can see that Fianna Fail has the maximum number of positive tweets, followed by Sinn Fein and then FineGael.

- **Aggregation of Results** - In this section, we will compare our results to the actual election outcome. The actual election outcome is as follows:

Fianna Fail - 38

Sinn Fein - 37

Fine Gael - 35

As per our PSS, the results of our research do not conform to the election outcome. Whereas, the results of overall positive sentiment contribution (Fig 1.3), are in agreement with the outcome.

Therefore, we can say that the percentage of positive opinion tweets for each of the contending parties parties, can be a good indicator of the election outcome. Some shortcomings of this research are talked about in the next section with some prospects of future work.

## 5 Shortcomings and Future Work

The research results don't correlate strongly with the real outcome. The seats won by the two leading parties were almost the same whereas the prediction shows Fianna Fail to be the clear winner. This can be considered as a shortcoming. One option that would probably improve the results is to collect tweets for all hashtags being used for candidates in all the constituencies.

Another shortcoming of this research is that there is a huge difference between the total tweets posted for Fine Gael and that for Sinn Fein and Fianna Fail. This makes us question the scalability of our results. One could question, if the number of tweets for Fine Gael were higher than the other parties, with its PSS same, our model would predict Fine Gael to be the winner.

Another direction could be to take this research and implement the same for some other elections to verify if it is globally applicable. Bovet et al. (2018) did a similar study and delivered promising results for the US Presidential Elections 2016 and the same can be done for other countries.

Finally, a different methodology could be used to fulfil the same objectives as ours.

## References

- Belcastro, L., R. Cantini, F. Marozzo, D. Talia, and P. Trunfio (2020). Learning political polarization on social media using neural networks. *IEEE Access* 8, 47177–47187.
- Bovet, A., F. Morone, and H. A. Makse (2018). Validation of twitter opinion trends with national polling aggregates: Hillary clinton vs donald trump. *Scientific reports* 8(1), 1–16.
- Choy, M., M. L. F. Cheong, N. L. Ma, and K. P. Shung (2011). A sentiment analysis of singapore presidential election 2011 using twitter data with census correction. *ArXiv abs/1108.5520*.
- Dickson, S. (2019, April). Twitter as an alternative review site.
- Franco-Riquelme, J. N., A. Bello-Garcia, and J. Ordieres-Meré (2019). Indicator proposal for measuring regional political support for the electoral process on twitter: The case of spain’s 2015 and 2016 general elections. *IEEE Access* 7, 62545–62560.
- Jose, R. and V. S. Chooralil (2015, Nov). Prediction of election result by enhanced sentiment analysis on twitter data using word sense disambiguation. In *2015 International Conference on Control Communication Computing India (ICCC)*, pp. 638–641.
- Joyce, B. and J. Deng (2017). Sentiment analysis of tweets for the 2016 us presidential election. In *2017 IEEE MIT Undergraduate Research Technology Conference (URTC)*, pp. 1–4.
- Lei, X., X. Qian, and G. Zhao (2016, Sep.). Rating prediction based on social sentiment from textual reviews. *IEEE Transactions on Multimedia* 18(9), 1910–1921.
- Sharma, P. and T. Moh (2016). Prediction of indian election using sentiment analysis on hindi twitter. In *2016 IEEE International Conference on Big Data (Big Data)*, pp. 1966–1971.