



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

School of Computer Science and Statistics

Assessment Submission Form

STUDENT NAME	SIDDHARTHA BHATTACHARYYA
STUDENT ID NUMBER	19301936
COURSE TITLE	MSC COMPUTER SCIENCE – DATA SCIENCE
MODULE TITLE	TEXT ANALYTICS
LECTURER(S)	CARL VOGEL
ASSESSMENT TITLE	MIDTERM ESSAY
DATE SUBMITTED	8 MARCH 2020
WORD COUNT	3872

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 Siddhartha Bhattacharyya

DATE 8 MARCH 2020

AUTHOR DECLARATION FOR GROUP ASSIGNMENTS

GROUP NUMBER: GROUP 17

MODULE TITLE: TEXT ANALYTICS

MODULE CODE: CS7IS4

ASSIGNMENT TITLE: MIDTERM ESSAY

WORD COUNT: 231

STUDENT NUMBER	STUDENT NAME	NATURE OF CONTRIBUTION	% CONTRIBUTION
19300933	JAGADISH RAMAMURTHY	Providing an overview of the research with documenting the abstract and keywords, also responsible for documenting the different methods attempted during data collection.	25%
19301936	SIDDHARTHA BHATTACHARYYA	Summarizing the processing strategies to extract non-textual entities - emotions from the tweets, also contributed to explaining the results section with graphical visualizations obtained from the experimental data	25%
19301913	MRINAL JHAMB	Reviewing the previous work related to the research and summarizing the future work of the research, also consolidated the individual sections to provide the final draft version.	25%
19317919	SOUMEN GHOSH	Providing the introduction describing the reason and benefits of the research, also responsible for summarizing the conclusion of the document.	25%

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.

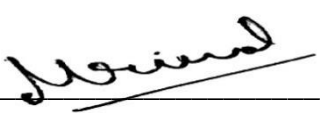
We declare that the percentage contribution by each member as stated above has been agreed by all members of the group and reflects the actual contribution of the group members.

SIGNED AND DATED









Sentiment Analysis of Political Opinions in Tweets: Before & After Elections

Jagadish Ramamurthy¹, Siddhartha Bhattacharyya²,
Mrinal Jhamb³, and Soumen Ghosh⁴

^{1,2,3,4}Department of Computer Science and Statistics, Trinity College Dublin, Ireland
^{1,2,3,4}{ramamurj, bhattasi, jhambm, ghoshso}@tcd.ie

Abstract. The daily usage of social media applications has taken a sharp increase among the people giving rise to the possibility of investigating and analysing social media data to detect and track political opinions of its users [21] [22]. In the paper, we perform sentimental classification analysis on a set of roughly 2 lakhs tweets produced during the run-up up to and after the Irish General Elections in February 2020. The tweets were retrieved via an open source package called GetOldTweets¹ and pre-processed using the Natural Language Toolkit² package. The tweets were analysed to find the preferences and emotions of the people to individual parties and/or candidates and equate the found analysis to the actual election results³ using the NRC Emotion Lexicon⁴ package. The results of the analysis currently derived are only partial and show the correlation between the tweets and the election results.

Keywords. *Analysis of public opinion, political sentiments, sentimental analysis, social media, elections, text analytics, NRC*

1. Introduction

The saying “*The pen is mightier than the sword*” suggests the fact that the written language is a more effective medium for problem solving than any other method in most cases [1]. Apropos to this, sentiment analysis can be defined as a methodology to extract non-textual information such as emotions and feelings from textual entities [2][3]. The social media is a platform where people cast their views, opinions about various entities and has adverse influential effects on the users’ point of view. Sentiment analysis and opinion mining are often used interchangeably in most platforms and areas of study, but we believe it is more appropriate to interpret these sentiments as thoughts that are emotionally charged. Sentimental analysis is merely used for predicting the emotions of the group of audience. This analysis is done mostly in social media chats or the tweets, thereby categorising the urgency of matter or the brand involved. Sentimental analysis triggers emotional aspect of the audience by inputting the data from social media and plotting the corresponding outcomes which can be used as deliverables.

Nowadays, if one wishes to purchase a consumer product, they are no longer required to ask for feedback personally, as there are customer reviews and debates about the product on public online forums. We have seen in recent years that opinionated social media posts have helped reshape industries and have shaped collective feelings and emotions that have had a profound impact on our social and political processes. Politics and elections involve a large amount of the population which leads to an outbreak of comments, debates and discussions amongst the people. Hence it is essential to analyse and correlate the textual comments and the election outcomes. Since social media platforms have a humungous audience, one such social media platform being Twitter, has more than 48 million active users, has become the most suitable prospect for sentimental analysis. The focus of the research is to get an understanding of the users’ emotions and opinions towards the candidates and/or parties before and after the elections by analysing the various tweets on political candidates and parties and equate the responses or emotions of the people towards the elections results.

The rest of paper is structured as follows. Section 2 describes the previous works related to the research. Section 3 speaks about the design methodology mentioning the description of the dataset and natural language processing tools. Section 4 provides the initial results achieved from the experimental

¹ GetOldTweets3. <https://github.com/Mottl/GetOldTweets3>

² NLTK Package. <https://github.com/nltk/nltk>

³ The RTE News. <https://www.rte.ie/news/election-2020/results/#/national>

⁴ NRC Emotion Lexicon. <https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm>

setup. The conclusion and directions for future work are mentioned in Section 6 and Section 7 respectively.

2. Related work

This can be further put under two categories, General sentiment analysis and sentiment analysis specific to political domain.

2.1. General Sentiment Analysis

There has been good amount of work done in the field of Sentiment Analysis. The research done at the early stages in the field was more centred around long documents like movie reviews and blogs. The expressiveness of blogs as being positive or negative is being established by Chesley in his work [6]. Godbole in his work on news articles tried to analyse the sentiments of people, places and things on the criteria that sentiment of each separate entity in the lexicon is represented by the assigned scores [7].

Post the era of analysis of large documents, when trend gradually shifted towards more shortened and summarized writings due to constraints imposed on the character count on posts on social media platforms, the research on the prediction of the sentiment on the basis of documents with less word count became the need of the hour. That's when Bermingham's research on analysis of microblogs for sentiment, which showed that opinions in small scale documents are more easily classified than those in large scale based on the hypothesis examined during the research. Microblogs from Twitter, blog posts etc. aided them in their research to establish the findings [8].

With the growing popularity of Twitter, there has been a great amount of work in the field of sentiment analysis based on Twitter data. Paroubek based on his work suggested that sentiments of a given text can be judged with the aid of their sentiment classifier [9]. Davidov in his work on sentiment analysis, was supervised, performed on fifteen different sentiment labels derived out of smileys and fifty Twitter tags which in turn aided to derive the sentiment type of the short texts [10]. Bakliwal in his work focused on the use of unigram and bigram features by using a corpus of pre-annotated tweets, which in turn boosted the classification accuracy [11].

2.2. Political Sentiment Analysis

With the growing penetration of internet and more people being active on it, it has become a main source of analysis to predict the popularity of a candidate or a party and sentiment surrounding them and even for the predication of results.

Tumasjan in his research in 2009 on the federal elections held in Germany used LIWC2007 (text analytics tool developed with the aim to reveal the thoughts, emotions etc. using text samples) examined more than a lakh tweets on politicians and political parties and concluded that proportion of tweets belonging to a particular party is proportional to their chance of winning [12][13].

Choy in his work presented the potential way to predict the vote share for each candidate in the presidential election of Singapore which happened in 2011 using online sentiment analysis [14]. During the US presidential election of 2012, a real time sentiment predictor based on the tweets extracted from Twitter was proposed by Wang [15].

Ringsquandl and Petković in their research deduced that there is a semantic relationship between the topics coined by political leaders, and the frequency of usage of noun phrases. This was derived while they were working on the campaign of politicians of Republican Party in the USA [16].

Sharma and Moh during the during the general state elections in 2016 worked to predict the Indian state elections using Hindi Twitter. They used SVM, dictionary-based and Naïve Bayes algorithms on a set of 42,235 Hindi tweets extracted using some tool to classify them as one of three emotions - positive, negative or neutral [17].

3. Design Methodology

This section describes the set up and methodology of the research. The section can be simplified into three steps –

1. Data Collection
2. Data Pre-processing
3. NRC Emotion Lexicon Package

3.1. Data Collection

The initial plan to retrieve the tweets required for the research was via the official Twitter API which can help search through the history of tweets posted on the website. However, the standard free API service provided for the non-paying users retrieves the tweets only from the past 7 days. The *GetOldTweets3* python package overcomes this limitation by taking advantage of the json loader in the current web browsers to retrieve tweets as far as from the start of Twitter.

This package allows to provide query search parameters along with a date range to retrieve the tweets. Though the feature of the tool is as same as the *Advanced Search* option available in Twitter, it provides further flexibility with numerous options such as to cap the number of tweets for each request, retrieve emoticons/emoji, tweets produced from a certain location, etc. and returns the set of tweets as a .CSV file for the analysis.

For the purpose of the research, we retrieved the tweets for one month exactly before and after the day of the election, 8th February 2020. Certain constraints and filters were in check to make sure there were no duplicates or unreadable symbols, etc. to bias the outcome of the results. The *language filter* was used to set the language to English to avoid conversion of other language characters to English, the *max tweets filter* to cap the maximum number of tweets to a certain limit to balance the tweets retrieved before and after elections. These constraints were set to simplify the data retrieval. The most important constraint was the *query search* parameter which includes keywords, hashtags (*words appended after #*) and mentions (*words appended after @*). Fig 1. shows a resulting tweet of a user's opinion on a candidate on Twitter for the keyword “*Mary Lou McDonald*” and hashtag “*GE2020*”.

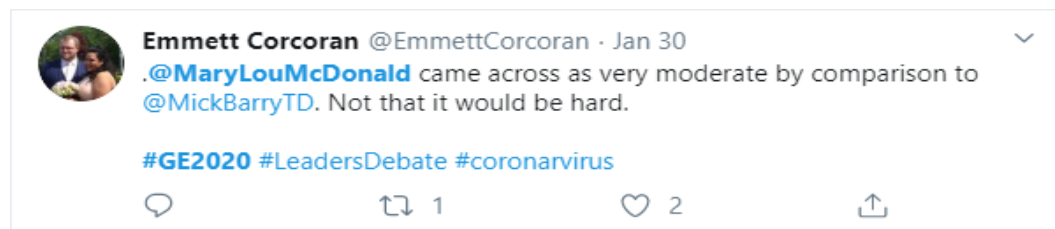


Fig 1. Example of a user's opinion on a candidate.

To make sure all the tweets retrieved were related only to political opinions of users and not from the actual candidates/parties, a list of possible hashtags, the candidates' names and their twitter handles, and the parties' names and their twitter handles were compiled and categorized based on the constituency. A few keywords used to retrieve related tweets are mentioned below.

KEYWORDS	
#sinnféin	Mary Lou McDonald
#finegael	Leo Varadkar
@fiannafailparty	Richard Bruton
@greenparty_ie	@neasa_neasa

The resulting tweets were split into individual datasets pertaining to parties and/or candidates. Table 1. shows the count of tweets collected after removing duplicates pertaining to each party before and after the elections.

PARTIES	TWEETS COUNT		TOTAL
Fianna Fáil	before	5717	12082
	after	6385	
Fine Gael	before	5889	13084
	after	7195	
Green Party	before	5526	10905
	after	5379	
Labour	before	9741	19540
	after	9799	
Sinn Féin	before	9449	17518
	after	8069	
Social Democrats	before	4997	9976
	after	4979	
Solidarity – People before Profit	before	1515	6515
	after	5000	

Table 1. Keywords based data count for each party

Table 2. shows the count of tweets collected post removing duplicates of the overall opinion of the users' regarding the Irish General Elections.

ENTITY	TWEETS COUNT		TOTAL
General Hash Tags	before	21109	32587
	after	11478	

Table 2. Hash Tag based data set count

3.2. Data Pre-processing

As previously stated, the datasets collected for each entity (viz. party, candidate and hashtag) were divided into 'before' and 'after' based on the election date. Data clean-up and pre-processing was required to transform the raw data collected from twitter to an acceptable format, so that it can be experimented on. Pre-processing in our initial experiment included the following steps:

- *Removing duplicates:* It can be considered with utmost certainty that the dataset obtained would contain duplicate tweets, therefore it was essential to keep only a single copy of each tweet in our corpus in order to prevent biased results. Duplicates were removed using python lists and dictionary data structures. As python dictionaries can contain only unique keys. This feature was exploited to remove duplicate tweet texts.
- *Removing noise:* There were certain words which did not correspond to any emotion. Such words like pronouns ('she', 'that', etc.), determiner words or stop words such as 'a', 'the' etc. were removed from the tweet texts. A list of stop words are provided in the NLTK package as well, which were used to check if our tweets contained stop words and then removed.

- *Lemmatizing*: It was necessary to lemmatize the tweet texts, so as to remove words having the same form. Lemmatizing is the process of transforming similar word forms into their root form. For example, the word ‘run’, ‘ran’, ‘running’ mean the same thing. The lemmatizer transforms each word to its equivalent form. If an equivalent form cannot be found, it is transformed to its root form. The *WordNetLemmatizer* function in NLTK (Natural Language Tool Kit) python package is used to achieve this.

Following these steps, we now have our clean corpora of twitter text for each of the datasets. The next section introduces the NRC Emotion Lexicon which is used to build a dictionary with the emotional correspondences of a collection of words found in the cleaned twitter corpora.

3.3. NRC Emotion Lexicon Package

This extensive lexicon package also known as ‘*EmoLex*’ has been developed using crowdsourcing with the help of the Amazon Mechanical Turk [18][19]. The ‘*EmoLex*’ package is essentially a dictionary with words and their corresponding emotions. This publicly contributed dictionary assigns 1 or 0 binary values to each sentiment that a word is assumed to have. The emotions considered for creating this dataset have been limited to the basic quintessential 8 emotions – ‘joy’, ‘anticipation’, ‘sadness’, ‘anger’, ‘fear’, ‘trust’, ‘disgust’, ‘surprise’, and ‘anticipation’ [20]. The lexicon also includes a ‘positive’ and ‘negative’ emotion in its lexicon as well. However, for our project, we have altered them to ‘for’ and ‘against’ respectively, as our project relates to election results. Therefore, a word with a positive connotation will mean that the tweet speaks ‘for’ a party or candidate, and vice versa.

abandon	anger	0	
abandon	anticipation		0
abandon	disgust	0	
abandon	fear	1	
abandon	joy	0	
abandon	negative		1
abandon	positive		0
abandon	sadness	1	
abandon	surprise		0
abandon	trust	0	

Fig 2.

An example is provided in the Fig 2. The word ‘*abandon*’ is given an emotion of ‘fear’, ‘negative’ and ‘sadness’. The other sentiment values are 0, which indicates the word does not correspond to those sentiments. Therefore, the word *abandon* elicits the emotions of fear and sadness and is a negative emotion overall. Using this lexicon, a dictionary is created with the emotions (viz. anger, anticipation, etc.,) as the keys and the values are the list of words the emotion elicits. This dictionary is used to analyse the tweets and check for words corresponding to the sentiment keys. The lexicon dictionary structure is as follows: {‘anger’: [list of anger words], ‘anticipation’: [list of anticipation words]}.

This dictionary is saved globally as a json file and reused on every dataset iteration. Once this setup is complete, the tweet texts are tokenized into individual words and looked up in our created lexicon dictionary above. If the words are found in the list, its sentiment key is fetched. Since a tweet has multiple tokens, each token might correspond to a separate emotion. The number of emotions calculated from each tweet is taken and the maximum among them is finalized as the overall emotion of the tweet. A global dictionary is maintained where the count for each of the emotions extracted through this method from the tweets is iteratively incremented. A layout of the setup is shown in Fig 3.

To explain our process through an example, consider the tweet in Fig 1. After tokenizing and clean-up of the tweet, the tokens collected are as follows: {‘*came*’, ‘*across*’, ‘*very*’, ‘*moderate*’, ‘*comparison*’, ‘*not*’, ‘*hard*’}. After looking up these tokens in our NRC lexicon dictionary, the emotion counts are collected. The word ‘*moderate*’ has a ‘for’ connotation in our lexicon, while the remaining words do

not elicit any noteworthy emotion. Therefore, the overall emotion for this tweet is considered as ‘for’. The final emotion counts dictionary is shown in Fig 3. It contains the number of tweets classified under each emotion.

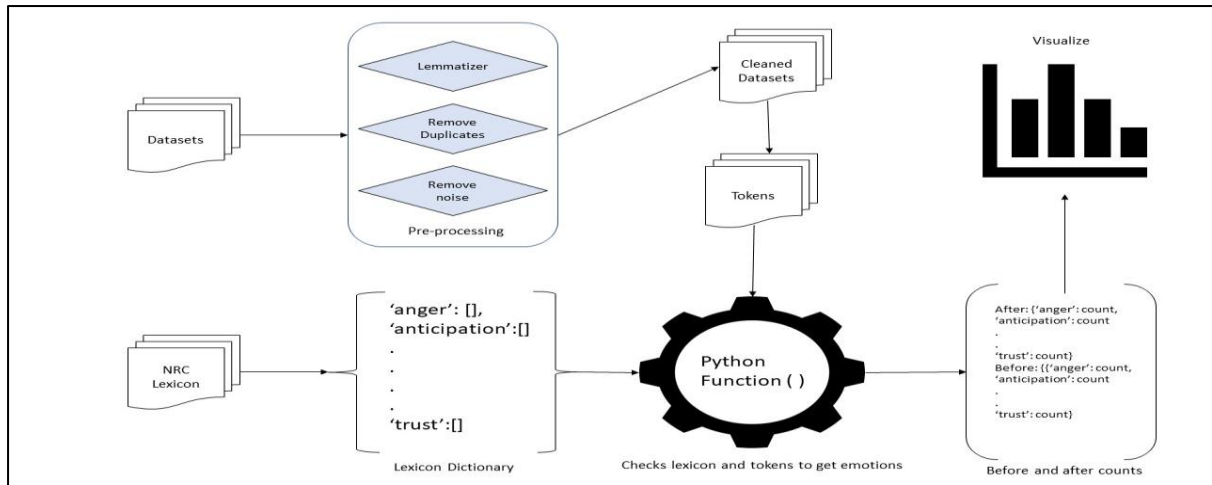


Fig 3.

4. Results

The results from our initial experiments are best described in visual form. The final emotion counts dictionary are used to create bar plots. We have considered the hashtag dataset and plotted the before and after emotions extracted from the tweets. From Fig 4.a, it is noteworthy that most tweets indicate an angry sentiment. There is not much change between the before and after sentiments, however there are significant differences between the emotions themselves. Most tweets seem angry, and on the first glance of a few collected tweets, it seems so too to the human eye.

We have also analysed the overall favourability of the 7 political parties in the Republic of Ireland from the datasets collected. This can be seen from Fig 4.b. This has been visualized by considering the number of ‘for’/‘positive’ emotion count for each of the party tweets. It seems that there are a lot of people who favour the Labour Party, followed by Sinn Féin. One of our future works would be to compare our findings with legitimate election results and check our method’s accuracy.

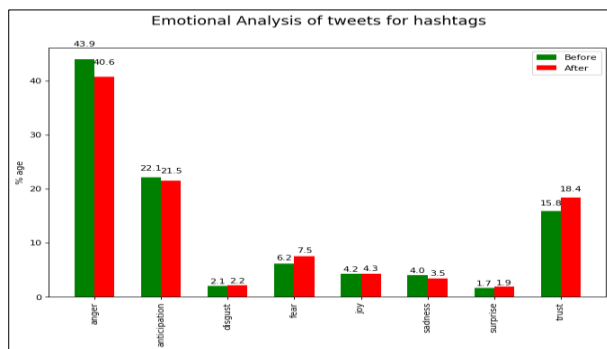


Fig 4.a

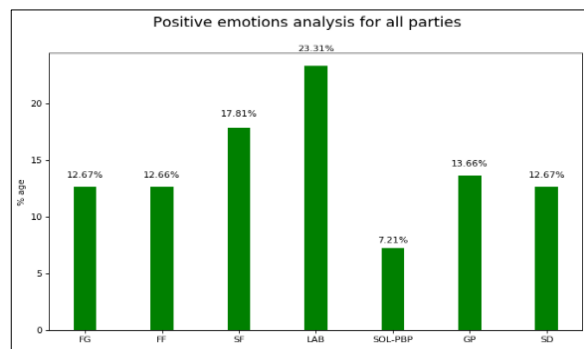


Fig 4.b

Furthermore, we have also plotted and visualized the positive sentiments data for individual political parties before and after the election date. They are displayed in Fig 5. The positive sentiments considered are ‘joy’, ‘anticipation’, ‘surprise’ and ‘trust’. Along with these four positive emotions, the number of ‘for’ and ‘against’ tweets have also been plotted for reference as percentages of total tweets.

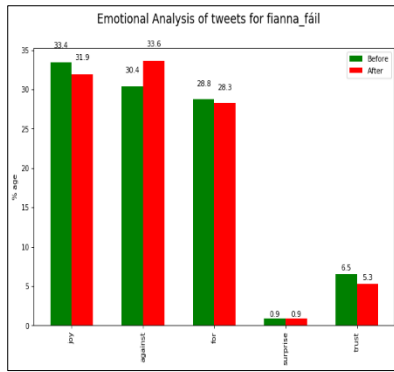


Fig 5.a

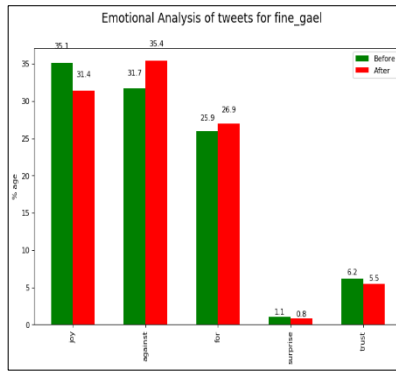


Fig 5.b

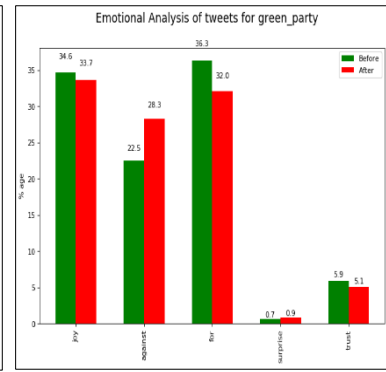


Fig 5.c

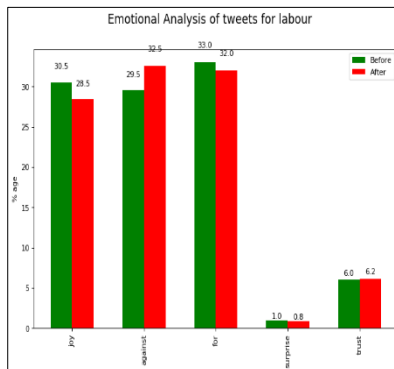


Fig 5.d

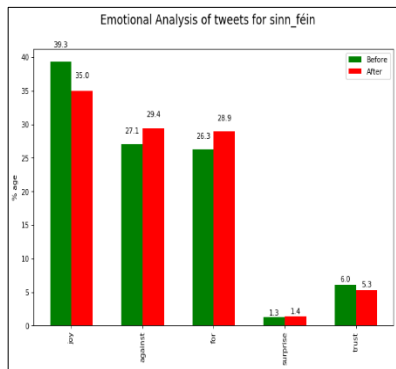


Fig 5.e

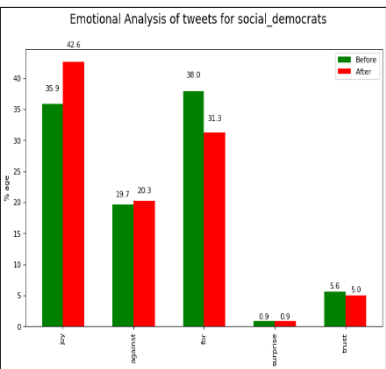


Fig 5.f

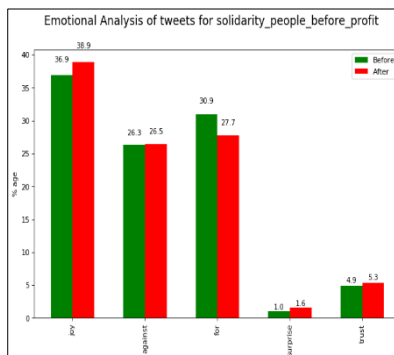


Fig 5.g

On initial analysis we found similar sentiments for each of the parties with certain noticeable difference between many of them. Twitter users do not seem to be ‘surprised’ about the events during the election campaigns. People favour the Labour Party followed by Sinn Féin. There is a significant difference between the number of tweets that are for the Social Democrats party as compared to the tweets against them. However, this favourability diminished after the election results by almost 7%. People also seem to have lost favourability for the Solidarity – People before Profit party by almost 3% after the elections. Increasing the data set size and selecting tweets that cover a range of emotions i.e. positive and negative, might improve the results. This is considered as a task for future work.

5. Conclusion

In this work, we brought in a whole new dataset of political tweets that were retrieved using the *GetOldTweets* python package. The tweets, cleansed and pre-processed using the *NLTK* package and the *NRC Emotion Lexicon* package, show that the people are angry and not surprised by the election results. The results show that is a lot of anger on the overall elections as compared to the primitive emotions while at the same time the voters do have a lot of more than the average support for their

favourite parties. The analysis and accuracy of the tweets could still very much differ by increasing the timeline for the tweets retrieved after the elections as to the current timeline of one month.

6. Future Work

Tweets normally are a combination of text and emoticons (which are a great depicter of sentiment). This research can further be extended by considering the emoticons used in tweets along with the textual information used in the paper with the aid of NRC package. This paper focuses on the sentiment for a subject before and after the election. Thus, in future work it can be worked upon to add the entire timeline of favourability.

The amount of data collected per subject can be increased which in turn will help to predict with even more certainty. This paper provides with the opportunity for future work by comparing with the information obtained from news and media to check the accuracy of sentiment estimation.

References.

- [1] The pen is mightier than the sword, Wikipedia, 22-Nov-2016. [Online]. https://en.wikipedia.org/wiki/The_pen_is_mightier_than_the_sword
- [2] B. Liu, Handbook Chapter: Sentiment Analysis and Subjectivity. Handbook of Natural Language Processing, Handbook of Natural Language Processing, Marcel Dekker, Inc. New York, NY, USA, 2009.
- [3] K. Dave, S. Lawrence, D.M. Pennock, Mining the peanut gallery: Opinion extraction and semantic classification of product reviews, in Proceedings of the 12th international conference on World Wide Web, 2003, pp. 519–528.
- [4] Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), e1253. doi:10.1002/widm.1253
- [5] B. Liu. Handbook of Natural Language Processing, chapter Sentiment Analysis and Subjectivity. Second edition, 2010.
- [6] Chesley, P.: Using Verbs and Adjectives to Automatically Classify Blog Sentiment. In Proc. of AAAI-CAAW-06. The Spring Symp. on Comp. Appro., 1 – 3 (2006)
- [7] Godbole, N., Srinivasaiah, M. Skiena, S.: Large-Scale Sentiment Analysis for News and Blogs. In Proc. of the Int. Conf. on Web. and Soc. Med. (ICWSM), Colorado, USA, 1 – 6 (2006).
- [8] Bermingham, A., Smeaton, A. F.: Classifying sentiment in microblogs: is brevity an advantage? In Proc. of the 19th ACM int. conf. on Inform. and Know. Manag. 1833 – 1836 (2010).
- [9] Pak, A., Paroubek, P.: Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In Proc. of the Seventh Int. Conf. on Lang. Res. and Eval. (LREC'10). 1320 – 1326 (2010).
- [10] Davidov, D., Tsur, O., Rappoport, A: Enhanced Sentiment Learning Using Twitter Hashtags and Smileys. In Proc. of the 23rd Int. Conf. on Comp. Ling.: Posters, 241-249 (2010).
- [11] Bakliwal, A., Arora, P., Madhappan, S., Kapre, N., Singh, M. Varma, V.: Mining Sentiments from Tweets. In Proc. of the WASSA'12 in conj. with ACL'12. 11 – 18 (2012).
- [12] Tumasjan, A., Sprenger, T. O., Sandner, P. G., Welpe, I. G.: Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In Proc. of the Int. Conf. on Web. and Soc. Med. 178 – 185 (2010).
- [13] Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., Booth, R. J.: The Development and Psychometric Properties of LIWC2007. Tech. Report, Austin, Texas, 1 – 22 (2007).
- [14] Choy, M., Cheong, M. L. F., Laik, M. N., Shung, K. P.: A Sentiment Analysis of Singapore Presidential Election 2011 using Twitter data with Census Correction. Research Collection Sch. of Info. Sys. 1 – 12 (2012).
- [15] Wang, H., Can, D., Kazemzadeh, A., Bar, F., Narayanan, S.: A System for Realtime Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle. ACL '12 Proc. of the ACL 2012 Sys. Demos. 115 – 120 (2012)
- [16] Ringsquandl, M., Petković, D.: Analyzing Political Sentiment on Twitter. Association for the Adv. of Artif. Intell. 40 – 47 (2013).
- [17] Sharma, P., Moh, T-S., Prediction of Indian Election Using Sentiment Analysis on Hindi Twitter. 2016 IEEE Int. Conf. on Big Data (Big Data). 1966 – 1971 (2016).

- [18] Crowdsourcing a Word-Emotion Association Lexicon, Saif Mohammad and Peter Turney, *Computational Intelligence*, 29(3), 436-465, 2013.
- [19] Amazon Mechanical Turk - <https://www.mturk.com/>
- [20] Plutchik, R. (1980). A general Psychoevolutionary theory of emotion. *Emotion: Theory, research, and experience*, 1(3), 3-33.
- [21] Woody D (2007). New competencies in democratic communication. *Blogs, agenda setting and political participation. Public Choice* 134(1-2): 109-123.
- [22] Madge C, Meek J, Wellens J, et al. (2009). Facebook, social integration and informal learning at university. *Learning, Media and Technology* 34(2): 141-155.