Sentimental Analysis for Political Activities from Social Media Data Analytics

[1]Naresh Kumar, [2]Aditi Sharma

Member, IEEE & Computer Society of India (CSI)

[1]Department of Mathematics, Indian Institute of Technology, Roorkee (India)

[2]Department of Computer Science & Engineering, M.BM. Engineering College,
Jodhpur, Rajasthan (India)

ABSTRACT

Emotional content and intelligence quotient play an important role in communication and psychological development. Facial expression in video has been supported multimodal behavior analysis by last decades. To go for more clear idea about the impact of an individual opinion in outer world, we compute it as a facts and figures of human sentiments. This work is carried out by collecting tweets data from social media and change point detection algorithm is used for temporal measurement of sentiments of an entity. In this research a model has been proposed which performs sentiment analysis over a time period on various news entities and performs interplay study between those entities. A predictive model is the expected outcome of this research which can promote in decision making in social, political and human health care issues.

Keywords: Sentimental analysis, Name entity recognition, Point wise mutual information, Social media activity, Opinion mining, Tweeters, Time series data,

I. INTRODUCTION

1.1 Sentiment analysis

Human being is closely linked with environmental outcomes which directly or indirectly force to stimulate the brain and consequently, it isperceived before any reflection and comes out as a thoughts of expression. These expression may create a facial pose, body pose or any action to respond the outcome accordingly the majority of dominating signals in the brain. Sentiment is the foremost response to reflect the thoughts. The necessity of sentiment analysis create research due to high involvement in industrial, political and psycho-visual aspect. Research evidence proved that sentimental analysis is highly close to human health and habit. Today'speople are pouring their opinions about almost anything on social sites inform of Facebook posts, tweets, blogs and news etc. for decision easier. So, sentiment analysis using social media data is very interesting area of research due to itsusefulness and efficiency because a lots of text data can be analyzed in very less time with verygood accuracy which is not possible manually.

Very natural motivation behind any sentimental analysis model is that the things are linked to each other which attract the researchers to collect data from news, tweets and other social media platforms. After collecting data sentiment analysis is performed to get polarity score for eachentity according to timeline then it is tried to capture interplay between entities i.e. how changein the sentiment of one entity trigger a change in other entity/entities.

1.2 Applications of Sentiment Analysis

Sentiment analysis is very useful manner to know contents of others brain. The majority of vote obtained from public can be applicable for election commission or any recruitment for particular public or private sector. The Businessmen always use the sentiment analysis to take decision that justifies it makes decision making easier. Facebook posts, tweets, blogs and newswhich is easily available for the correlation of sentiment analysis to take fast decision. The experimental sentiment analysis on patient can enhance health care guide and dietician knowledge. Literature evidence proved by making easier decision from the massive data samples that SA is highly applicable in data mining and the problem big data analytics to forecast many industrial facts study of impact analysis.

1.3 Challenges in Sentiment Analysis

Sentiment analysis deals with the subjectivity i.e. emotions which is very easy for humans to deal with but computer science finds it challenging things and sometimes fails in accuracy due toseveral challenges that include Languages complexity, Domain Dependency, Entity Identification, Detecting Spam and Fake opinion, Subjectivity Detection and negation.

1.4 Organization of Research Paper

Rest of the work is organized as literature survey and related work described in second section. In this section, wepresent a brief summary of literature related to sentiment analysis followed by the research gaps in context of our research problem. Section 3 presents the proposed methodology based on the research gaps identified for processing this research work. Implementation of the proposed work of section 3 is carried out in section 4 which presents brief discussion on result and analysis. Finally, section 5 belongs to conclusion and future directions for sentimental analysis.

II. RELATED WORK

A.brief literature survey of sentiment analysis

The huge growth of social media blogs, networking, and reviews etc. on the web becomes a rich source to collect data from web pages and news in several multimedia form of the opinionated text, which can be mined and used in decision makingprocesses by the organizations. Topic based text classification has been very old and well settledresearch area but text classification based on sentiment is relatively young but trending researcharea.According to Bo pang et al in [1] most of the work has focused on topic basedclassification e.g., sports, politicaletc. But apart from the topic of an article, there is always anemotions or sentiments about an entities are also present in it and sentiments can be very crucialfactor.According to test [8], presented a non-topic based text categorization system, which classified documents according to their source with statistically detected stylistic variation serving as an important cue e.g., The New York Times vs The daily News. In 2004, Wiebe et al attempted to findfeatures which can indicate if some subjective language has been used in an article.

Thesefeatures can be used to categories a text on the basis of genre. In thisthey used the idea of a classifier which can classify subjective and objective data by usingsubjective nouns which were learned by using bootstrapping algorithms and they achieved 77%precision. They just separated text documents into to class objective and subjective without explicitly address the task of specific opinion.Pang et al applied supervised learning methods [10] e.g. naïve Bayesianclassification, Support vector machines (SVM). They used these techniques to classify thepolarities of movie reviewsand proposed that standard machine learning performs better than baseline produced byhuman. According to their Experimental results, they applied all thecombination of unigrams, bigrams,POS and adjectives with term presence and termfrequencies. They found SVM by using unigram, bigram and term presence give best result.Generally, knowledge based research focuses on classifying the semantic orientation of individual words orphrases, using linguistic heuristics or a preselected set of seed words. In 2002, Turney andLittman [11] presented an algorithm which determines semantic orientation for unsupervisedlearning from very big volume of corpora. In this method they areusing a web search engine and a pointwise mutual information (PMI) function to find thesemantic orientation. In the training corpus they used for evaluating the result of algorithmcontains nearly 100 billion words The ratio PMI in equation (2.1) tells about the statistical dependency between two terms and semantic orientation (SO) tells about the opinion of phrase and computed by considering the factthat how much associated with the some positive word as excellent and with somenegative word like poorSO (phrase) in equation (2.2).

$$PMI(term_1, term_2) := log_2\left(\frac{P_r(term_1 \cap term_2)}{P_r(term_1).P_r(term_2)}\right) \quad (2.1)$$

$$\mathrm{SO(phrase)} := \mathrm{PMI\ (phrase, ``excellent") - PMI\ (phrase, ``poor")} \quad (2.2)$$

In 2004, Bing et al [12] have explained how we can getsentiment analysis at aspect level and how to summarize customer reviews. They described athree step method for extracting aspect expressions which includes mining product features or aspects onwhich customers have commented, filtering opinion sentences in a review and findingpolarity of each opinion sentence whether it is positive or negative and making summary of all the resultsto give resultant opinion. According to this paper, all aspects are nouns, phrases andopinion words are adjectives. They have mainly used unsupervised learning approach forsentiment classification. To find frequent features or aspects they used POStagging to findfrequent product features.

Bing et al [13] in 2012, used lexicon and dictionary based approach for mining and summarizing customer reviews. Theyapproached bootstrapping algorithms which start with small set of seed words with known sentiments and growthis set with the help of online dictionary like WordNet or thesaurus. Words which are newly discovered are added to the seed list anditerative process stops when new words or phrases are not found.They left future scope for handling difficulties in domain and context specific words. More research work related to sentiment analysis is discussed in related work because they arehelpful and somewhat related to our proposed model.

2.1 Related To Proposed Work

In our research work, we use twitter data, news headlines for sentiment analysisand polarity score, by lexicon based methods. The authors extracted data related**[**2] to political domain from web logs for two year and used Naïve Bayes and SVM classifier to predicate thesentiment and opinion of the political blog and posts. They targeted only political domain without having the interplay between sentiments of entities related to their respective domain. The work presented in [3]concentrates only on financial data extracted from bloggers towards companies and their stocks from which they prepared a corpus of financial blogs and by using text extraction techniquesto generate topic specific subdocuments which is further classified by document based sentimental analysis algorithms. In newspapers and blogs [4], there is opinionated data about the news entities like people, places and things. In their model, they assigned scores which indicates positive or negative sentiment or opinion to each distinct entity in their text corpus. They performed in two phases, one of which is sentiment identification phasein which sentiments toward each entity, associate it with respective entity and second phase is sentiment aggregation and scoring phase in which they summarized the overall sentiment of all the entities. In [3] and [5], authors used Twitter the most popular microblogging features like emoticons, hashtags etc., in order to perform sentiment analysis.

A target dependent sentiment classification approach [6] is proposed usingtwitter which takes a target as a query from user. According to query, they collected tweets related to that entity and performed sentiment classification. This involves two steps, one is defining features based on target provided by the query and second is collecting tweets about target and performing sentiment classification. A method for extracting sentiment [7] from of text is proposed which is based on lexicon dictionary in which they have words tagged with polarity of sentiment and strength of sentiment. This method performed consistently better across various domains and for unseen data as words mainly adjectives are tagged with polarity score.As mentioned earlier most of the research work in area of sentiment analysis is focused on howto assign a polarity score to a piece of text. Research work mentioned in related work has been focused onfinding sentiments for individual entities in a single domain but focus on interplay ofsentiment between entities across multiple domains and the relation between entities according to change in their sentiments is missing.

III. PROPOSED WORK AND SYSTEM ARCHITECTURE

3.1 Problem Definitionand Overview

In our research work, we want to find how closely two real world entity are sentimentally correlated with each other which motivates us to focus Sentiment analysis and interplay study of entities using news headlines and tweets. Change in sentiments of one entity triggers a change in other entity consistently over a period of time indicates the correlation of sentiments interplay between them. When some event happens with any person, product, or any real world object then sentiment and opinion of people towards it also changes, then how this change in sentiments affects the sentiments of people towards all the related entities. If we have frequent sentiment interplay pattern known in advance then decision making process can be very effective e.g., when price of patrol increases, prices of many things increase in market and when crimes are increasing in some country and sentiment of people about that country is negative then tourism business and many other domains also suffers. When some political party wins, this event can trigger changes in the sentiment of many related entities in various domain. If sentiments toward Aam Aadmi Party (AAP) and Kejriwal are positive then it is very natural that people will have positive sentiment towards everything related to AAP or Kejriwal.

Interplay study of entities means when a considerable change in sentiments of an entity is found at time T in a time series then how it is going to affect other related entities i.e. is there any considerable change in the sentiment of related entity at its time series in T+t. In this context'ť' can be any time depending on requirement of experimental setup. Following graph shows the 60 days sentiments change of Modi and BJP in the months of March and April day wise.

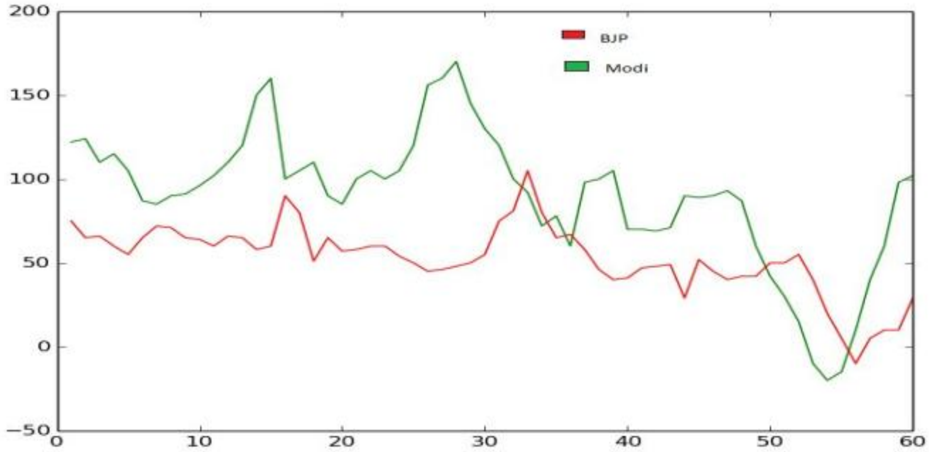

Fig. 3.1: Tweets Timeline Sentimental Score between Modi and BJP

The graph in fig. (3.1) shows change in the sentiments of Modi and BJP are correlated which is obtained by sentiment analysis of tweets in two months reports. We represents a model which can collect data from social networks site twitter and Facebook, candivide and perform sentiment analysis for many entities and monitor the interplay of sentimentsacross many related entities.

IV. EXPERIMENTS AND EVALUATIONS

In this research work, twitter data and news headlines are collected to determine the interplay and triggering of entities belonging to various domain.Overall process is divided in mainly two phases. In first phase, Data collection and preprocessing of the news headlines and tweets are performed for the entities appearing in these news headlines. Second phase corresponds to sentiment analysis and Interplay study, for each tweet collected and entity sentiment score is assigned depending on the kind of polarity it carries and after aggregating sentiment score to some level. The procedure is accomplished according to the fig 4.1 represented as an algorithmic steps.

Algorithm 4.1: Two Phase Algorithmic Steps for Sentiment Analysis

Algorithmic Steps:

1. Collecting news headlines classified into respective domain.
2. Finding entity/entities in each headline and selecting entities of interest among all.
3. Collecting tweets about each entity retrieved from the step 2.
4. Assigning sentiment polarity score according to the sentiment inside each tweet.
5. Aggregation of sentiment scores day wise or any granularity of user's choice and building a timeline graph i.e. time Vs sentiment for each entity
6. Performing interplay study between entities and finding relationship between entities pairwise i.e. how much correlated two entities are

4.1 Finding NER in Each News Headline

Named entity recognition (NER) is well known research area for information extraction (IR). NER is a process of labeling a word or phrases with names of persons, organizations, locations, and expressions of times. In our project we are using Stanford NER version 3.5.2. Stanford NER is a Java implementation of a Named Entity Recognizer. But when we apply this on tweets, due to informal structure of tweets sometime it fails to detect named entity. So overall accuracy it provides is around 80 %. We pass each news headline tweeted by a newsgroup to the Stanford NER and it appends named entity found in that news headline at the end of the headline. Table 2 is a sample of output produced in this step.

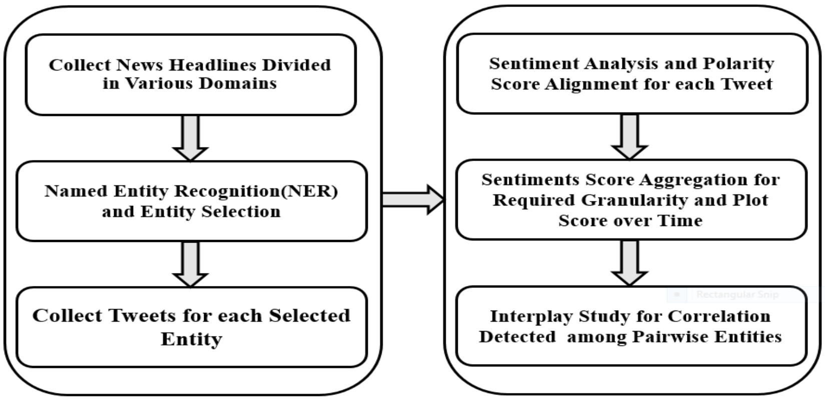
Collect News Headlines Divided in Various Domains
Named Entity Recognition(NER) and Entity Selection
Collect Tweets for each Selected Entity
Sentiment Analysis and Polarity Score Alignment for each Tweet
Sentiments Score Aggregation for Required Granularity and Plot Score over Time
Interplay Study for Correlation Detected among Pairwise Entities

Fig. 4.1:Implementation Design and System Architecture

4.2.Sentiment Polarity Score Assignment

The polarity score assignment to each tweet collected, is performed accordingly the flow algorithms in fig. 4.3, with the help of python library available for Sentiment Classifier 0.6 in which word sense disambiguation is performed using WordNet and word occurrence statistics from movie review corpus data. It classifies into positive and negative categories also assign a score according to polarity of tweet.

Due to poor performance of this classifier for negative classes wedesign a rule based approach represented by fig. 4.3, in which we used some NLP techniques and five type dictionaries for positive, negative, incremented, decremented andinverter to assign a score to each tweet.

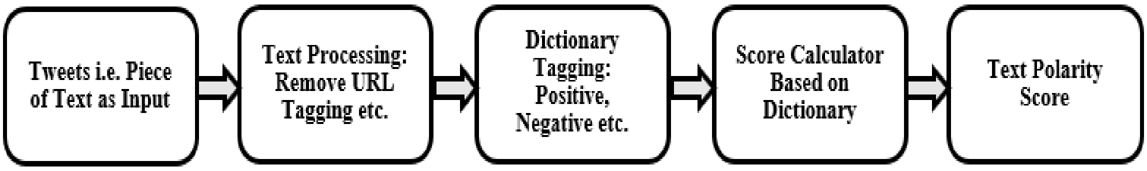
Tweets i.e. Piece of Text as Input
Text Processing: Remove URL Tagging etc.
Dictionary Tagging: Positive, Negative etc.
Score Calculator Based on Dictionary
Text Polarity Score

Fig.4.2: Text Polarity Score Algorithm

Positive dictionary which contains 2100 positive words, Negative contains 5000 negative words. We also used incremented dictionary which contains those words which increases the strength of sentiment by adjective specifications.

4.3 Performance Analysis of interplay Entity

Interplay study of sentiments analysis among various entities over a time series data is correlated by computing change point computation. Several statistical methods are available to carry out change point computations in time series of every entity.

4.3.1 Change Point in Time Series

Sentiment analysis in 2D sentiment score and time graph, where each data point is represented by <s, t > meaning s is sentiment score at time t.Let d_1, d_2, d_3 …... d_n are n data points in our time series.To find all such change point we have various methods. We apply cumulative summary based on algorithm 4.2 which shows that firstly mean of polarity scores is computed for all data points in the time series for each entity. Rest of the computations follows as per the steps figured out. This has been highly voted but it does not work for short term trends. It's preferable for whether forecasting and drug discovery like long terms goal sentiments. After getting cumulative summary for each data point, these cumulative sum (CUMSUM) values need to becompare with two threshold values of upper thresholds and lower thresholds. A data points is saidto be a change point if its CUMSUM value is either more than upperthreshold or less than lower threshold. Threshold values are decided by standard deviation computation.

Algorithm 4.2: Time Series Change Point Detection

```
Algorithmic Steps:
1.  Calculate mean X of all the data points.
2.  Initialize Cumulative sum CS=0. Collect tweets about each entity retrieved from
    the step 2.
3.  For finding the CS:   for all i=1 to n
        a. Do $CS_i = CS_{(i-1)} + d_i - X$
4.  $CS_i$ for each point is called as cumulative summary (CUMSUM) for each data
    point $d_i$which can be seen as $\sum_i(d_i-X)$ where i=1 to n.
```

Due to failure algorithms 4.2 in short term and critical sentiments a new algorithm is proposed. In our method, to incorporate the failure algorithm 4.2, we are using a sliding window scheme to calculate moving mean, SD on time series change points where dynamic threshold values is obtained for change point detection.Mean X_i and standard deviation SD_i are computed after fixing i_{th}window size. Computation for upper threshold (UT) and lower threshold (LT) for each values is given by equation (4.1) and (4.2).

$$UT_i = X_i + SD \qquad (4.1)$$

$$LT_i = X_i - SD \qquad (4.2)$$

V. RESULTS AND COMPARATIVE ANALYSIS

Graph between sentiment polarity score and time series is shown by figure 5.1 which represents Modi's sentiments time line with sliding window of mean, lower threshold and upper threshold over 90 days of tweeter data computations. Change points are shown by circles which are determined by deviation from mean and SD of that moving window and signified by the suddenchange in the sentiment score of an entity.Furthermore, Change point from tweets of BJP group is computed in the graph figured by 5.2 between sentiments score and 90 days periods of time series data from tweeters timeline of the group.

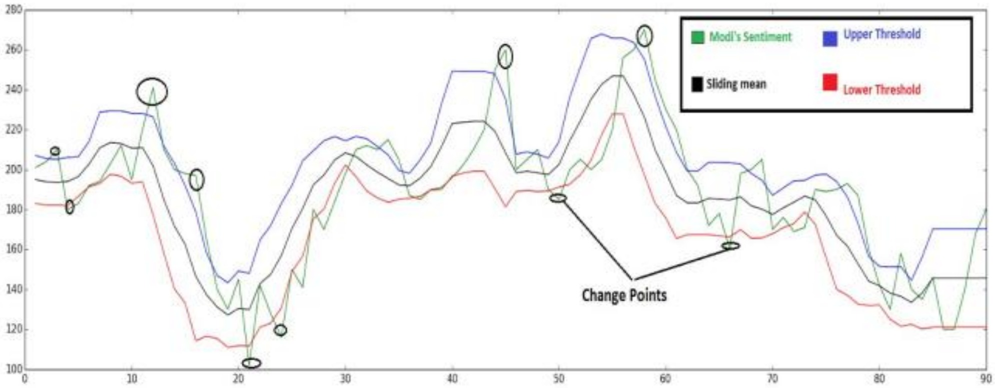

Fig 5.1:Modi's Sentiment Tweets with mean, LT and UT against Time Series Data

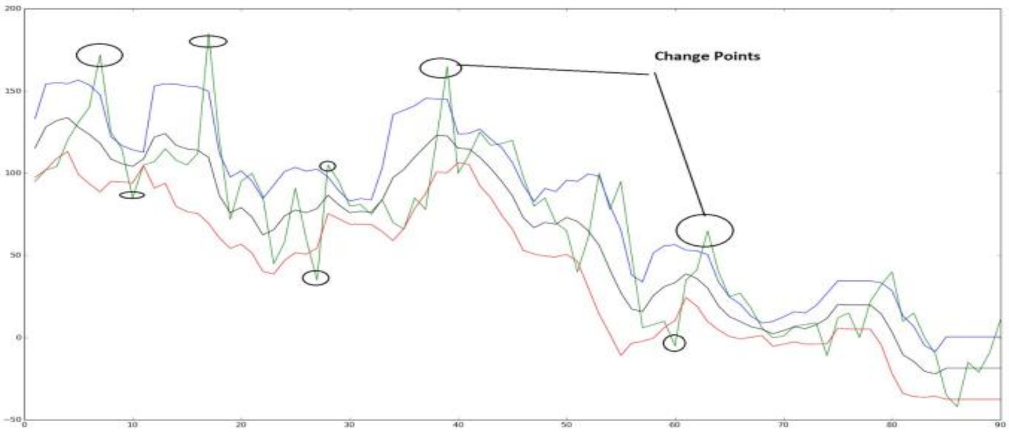

Fig5.2:BJP Sentiment Score with Change Points against Time Series Data

Table 5.1: Modi's Sentiments Timeline against Change-point

Change points	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Day	3	12	16	21	24	26	33	45	51	54	58	66	69	72	78	83
+ / -	+	+	+	-	-	-	+	+	-	-	+	-	+	-	+	+

Table 5.2: BJP Sentiments Timeline against Change-point

Change points	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Day	9	11	17	22	28	30	38	40	41	45	53	59	64	74	80	84
+/ -	+	-	+	+	-	+	-	+	-	+	+	-	+	-	+	+

Table 5.1 and 5.2 present the change points against polarity score and tweets time series data for Modi's, and BJP sentiments analysis. Here correlation is agreement and disagreement scores for two entity which is denoted by X → Y which find the effects of entity X on entity Y. Positive score is considered that change is X's timeline affects same way in Y's timeline.

Table 5.3: Interplay Score among the Sentiments from Tweets of Entities

Interplay Entity	Modi→ BJP
Total Change Points	16 (9+ &7-)
Agreement Points	7
Disagreement Points	4
Agreement Score	0.43
Disagreement Score	0.25

There are 16 change points in Modi’s timeline. Out of 16, 9 are positive and 7 are negative change points. Timeline of BJP also detected with same number of change points, positive and negative change points.Agreement score: In Modi’s timeline 7 Change points found a change point inBJP timeline within the window time of same nature. So agreement score is7/16=0.43 which is reflected in table 5.3.

VI. CONCLUSION AND FUTURE DIRECTIONS

Sentiment analysis is the process of classifying a given piece of text into three classes of polarity i.e. positive, negative and neutral. Freedom of expressing opinions andsentiments about almost everything on online social networks sites, increases huge amount of information every moment in the form of tweets, blogs, news, and forums. By using this data, we canmine and analyses this data for various decision making and prediction purposes.

Twitter is most popular microblogging social site where people pour their opinions about everything of the products and services, politics and religions. For generalization, twitter provides texts data from various social groups and people of different interests like celebrities, company representatives, politicians, and presidents of country. Expected features of news being fast, precise, timely and accurate attract everyone that results everyone have keen interest to know what are news headlines regarding the most popular and current events around the world. Real Time sentiment analysis is expected one of the future direction of this research while randomly data intrusion is considered. Furthermore, huge amount of tweets from various class of entity switch this problem to big data analytics.

REFERENCES

[1] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval, 2(1–2), 1-135.

[2] Soares, D., Gonçalves, L., & Ferreira, T. (2016). Political Sentiment Analysis.

[3] O'Hare, N., Davy, M., Bermingham, A., Ferguson, P., Sheridan, P., Gurrin, C., & Smeaton, A. F. (2009, November). Topic-dependent sentiment analysis of financial blogs. In Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion (pp. 9-16). ACM.

[4] Godbole, N., Srinivasaiah, M., & Skiena, S. (2007). Large-Scale Sentiment Analysis for News and Blogs. ICWSM, 7(21), 219-222.

[5] Java, A., Song, X., Finin, T., & Tseng, B. (2007, August). Why we twitter: understanding microblogging usage and communities. In Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis (pp. 56-65). ACM.

[6] Jiang, L., Yu, M., Zhou, M., Liu, X., & Zhao, T. (2011, June). Target-dependent twitter sentiment classification. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1 (pp. 151-160). Association for Computational Linguistics.

[7] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. Computational linguistics, 37(2), 267-307.

[8] Teahan, W. J., & Harper, D. J. (2003). Using compression-based language models for text categorization. In Language modeling for information retrieval (pp. 141-165). Springer Netherlands.

[9] Wiebe, J., Wilson, T., Bruce, R., Bell, M., & Martin, M. (2004). Learning subjective language. Computational linguistics, 30(3), 277-308.

[10] Pang, B., Lee, L., & Vaithyanathan, S. (2002, July). Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10 (pp. 79-86). Association for Computational Linguistics.

[11] Turney, P. D., & Littman, M. L. (2002). Unsupervised learning of semantic orientation from a hundred-billion-word corpus. arXiv preprint cs/0212012.

[12] Hu, M., & Liu, B. (2004, August). Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 168-177). ACM.

[13] Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167.