MANTECH Publications

Recent Trends in Computer Science and Software Technology
Volume 3, Issue 1, January-April 2018

Political Comments Analysis Using Text Mining

Anurag Khadase[1], Diksha Jambhulkar[2], Pradip Ramteke[3], Shubham Borghare[4], Amit Pimpalkar[5]

Department of Computer Science & Engineering

R.T.M.N.U

***Corresponding Authors' email id**: a.khadase1994@gmail.com[1], pradipramteke18@gmail.com[2], borghare.sb@gmail.com[3]*

Abstract

Text prospecting refers to the use of data prospecting, tweet analysis, compotation calculations, and symmetric to systematically determine, cutting, characterize, and study feeling states and instinctive useful data. Text mining is widely applied to voice of the client materials such as reviews and survey analysis, online and social media, and wellness program for applications that range from marketing to customer courtesy to clinical medicine. Conventionally on twitter posting, text analysis aims to determine the prejudice of a twitter account holder, advertisement or other subject regarding some motive or the overall transient polarity or impulsive receptivity to archive, intercommunication, or event. The attitude may be a prudence or estimation, affective state (haddest say, the impulsive state of the author or speaker), or the steady impulsive articulation. We extracted recent 50 tweets as a sample to analyze the sentiment of a selective twitter handle or twitter account using the highest score generated for positive, negative & neutral file.

Keywords: *Text prospecting, Text Mining, Political Comments*

I. INTRODUCTION

Text prospecting refers to the use of data prospecting, tweet analysis, compotation calculations, and symmetric to systematically determine, cutting, characterize, and study feeling states and instinctive useful data. Text mining is widely applied to voice of the client materials such as reviews and survey analysis, online and social media, and

wellness program for applications that range from marketing to customer courtesy to clinical medicine. Conventionally on twitter posting, text analysis aims to determine the prejudice of a twitter account holder, advertisement or other subject regarding some motive or the overall transient polarity or impulsive receptivity to archive, intercommunication, or event.

The attitude on which user can automatically extract tweets i.e. its provide numbers of data that updated daily life information or tweets on twitter. These information will help for different purpose such as discovery of certain keywords, finding brand of certain product and services etc. hence twitter provides information in very short period of time.

This project explore the methods that automatically extract large numbers of information in the general public by detecting tweets that daily published on Twitter. Specifically, The extracted text are compared with words that present in dictionary that measure correlation also determine whether strong political sentiments performance that extracted from Twitter. After finding the sentiments its automatically gives the scores on may be a prudence or estimation, affective state (haddest say, the impulsive state of the author or speaker), or the steady impulsive articulation. We extracted recent 50 tweets as a sample to analyze the sentiment of a selective twitter handle or twitter account using the highest score generated for positive, negative & neutral file.

2. LITERATURE REVIEW

In general, sentiment analysis has been investigated primarily at three levels [4]. In document level, the major task is to classify whether an entire opinion document expresses, is a positive or negative sentiment. This level of study assumes that every document expresses opinions on one entity. In sentence level the basic task is to examine whether every sentence expressed a positive, negative, or neutral opinion.

This level of study is closely associated with sentiment extraction, sentiment classification, and subjectiveness classification, report of opinions or opinion spam detection, among others [6]. It aims to investigate people's sentiments, attitudes, opinions emotions, etc. towards components like, products, people, topics, organizations, and services [7]. Gold standards has been created by combining sense-tagged and positive, negative sense labels. Gold standard means that w has

sense label Ws, Ws has +/- effect label l. To model relations between sense groups and arguments for each +/- effect , LDA model was used. Two LDA based models to investigate tweets in a given variations period. The primary version, referred to FB-LDA, can filter out history topics and extract foreground subjects from tweets in the given variant length. They recommended generative model called RCB-LDA. RCB-LDA first extracts representative tweets for the foreground subjects.

3. METHODOLOGY

We've used several phases to describe our project that are:

3.1 Tweet retrieval

Twitter data is extracted using the existing Twitter APIs (twitter API) for data extraction. For extraction of a tweets would be selected based on a keyword (name of person) the purpose was, we are getting exact tweets of that person, i.e. product reviews movie reviews. We haveelected to use the Twitter API due to ease of data.

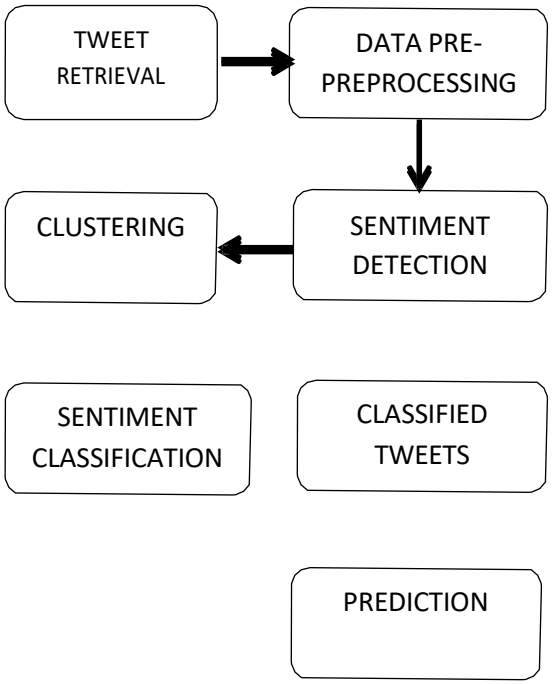

Fig Proposed System

3.2 Data Pre-processing

Text pre-processing the process of filtering the extracted tweets before analysis. It includes identifying and eliminating short texting massage's, non-textual content and content that is irrelevant to the area of study from the data.

3.3 Sentiment Detection

At this stage, each sentence of the review and opinion is examined for subjectivity. Sentences with subjective expressions are retained and that which conveys objective expressions are discarded. Sentiment analysis is done at different levels using common computational techniques like Unigrams, lemmas, negation and so on.

3.4 Clustering

To categories the tweet analysis into positive, negative and neutral clustering technique called hierarchical clustering will be used. Hierarchical clustering is used to build hierarchy of clusters to make the. according to TF-IDF algorithm. Clustering starts with the positive, negative and neutral tags, where tags refer to the particular group which contains sentiments of same type. For ex. Positive tag contains all positive sentiments. For further processing only positive and negative sentiments will be considered representation on that classified tweet. without the classification of tweet, we are not able represent in graphical form of tweet. For further processing only positive and negative sentiments will be considered.

3.5 Sentiment Classification

Sentiments can be broadly classified into two groups, positive and negative. At this stage of sentiment analysis methodology, each subjective sentence detected is classified into groups-positive, negative, good, bad, like, dislike.

TF-IDF (Term Frequency-Inverse Document Frequency) is a text mining technique used to categorize documents. Have you ever looked at blog posts on a web site, and wondered if it is possible to generate the tags automatically? Well, that's exactly the kind of problem TF-IDF is suited for.

3.6 Classified tweet

The classified tweet is those tweet which are classified form which mean that we can perform the task graphical representation on that classified tweet.

3.7 Prediction (Output)

The main aim of sentiment analysis is to convert unstructured tweet into meaningful information. The output of the analysis is

shows the tweet in which form either it is in positive, negative or a neutral form. Also we areused shown the graphical way representation of output analysis.

TF-IDF: Term Frequency-Inverse Document Frequency

Term Frequency-Inverse Document Frequency (TF-IDF) is a text mining strategy used to a classification documents. Have you ever looked at blog posts on a web site, and wondered if it is possible to generate the tags automatically? Well, that's exactly the kind of problem TF-IDF is suited for.

It is worth noting the differences between TF-IDF and sentiment analysis. Although both could be considered classification techniques for text, their goals are distinct. On the one hand, sentiment analysis aims to classify documents into opinions such as 'positive' and 'negative'. On the other hand, TF-IDF classifies documents into categories inside the documents themselves. This would give insight about what the reviews are about, rather than if the author was happy or unhappy. If we analyzed product review data from an e-commerce site selling computer parts like, laptop, mouse etc.

Example: Tagging Blog Posts

Step 1: Generate Scores for Each Document

Let's say you have a 100 word blog post with the word "JavaScript" in it 5 times. The calculation for the Term Frequency would be:

TF = 5/100 = 0.05

Next, assume your entire collection of blog posts has 10,000 documents and the word "JavaScript" appears at least once in 100 of these. The Inverse Document Frequency calculation would look like this:

IDF = log (10,000/100) = 2

To calculate the TF-IDF, we multiply the previous two values. This gives us the final score:

TF-IDF = 0.05 * 2 = 0.1

Step 2: Decide a Threshold to Tag

After running this algorithm against all 100 of the blog posts with the word "JavaScript", you end up with a score for each.

This is where you will have a chance to exercise the creative aspect of being a data scientist. Let's assume that you have a wide range of scores, ranging from 0.05 to 0.5.

Continuing a simple example from this collection, 0.05 would be a 100 word document with 1 instance of "JavaScript" and 0.5 would be a 100 word document with "JavaScript" appearing 25 times. To determine if the document will be tagged with "JavaScript", you need to decide on a threshold score.

The score you choose will vary depending on your data set. A document with only one instance of "JavaScript" (score 0.05) is unlikely to be focused on JavaScript, but obviously the high score of 0.5 is probably on topic.

4. RESULT AND DISCUSSION

4.1 **Data Description:** We have got the integrated data for a particular twitter handle, we got recent 50 tweets and that described in the separate page i.e., time, date that means at what time particular tweet is tweeted, and at what date particular tweet is tweeted etc.

4.2 **Sentiment Description:** We have created three program files as in PHP program as negative.php, positive.php&neutral.php. Each file represent the sentiment score and highlight the words matched with dictionaries and separated those words and made a separate list of those words. Based on highest score among those three files, we determined the sentiment for a particular twitter account holder.

Accuracy: We have 85% accuracy as of now, previously it was around 80% accuracy.

Table: 2

Twitter Handle	Narendra Modi	Devendra Fadnavis	Barack Obama	Total
No. of Tweets	50	50	50	150
Accuracy	85	83	89	85

CONCLUSION

We presented results for sentiment analysis on Twitter. We use previously proposed state-of-the-art unigram model as our baseline and report an overall gain of over 4% for two classification tasks: a binary, positive versus negative and a 3-way positive versus negative versus neutral. We presented a comprehensive set of experiments for both these tasks on manually annotated data that is a random sample of stream of tweets. We investigated two kinds of models: tree kernel and feature based models and demonstrate that both these models outperform the unigram baseline. For our feature-based approach, we do feature analysis which reveals that the most important features are those that combine the prior polarity of words and their parts-of-speech tags. We tentatively conclude that sentiment analysis for Twitter data is not that different from sentiment analysis for other genres. In future work, we will explore even richer linguistic analysis, for example, parsing, semantic analysis and topic modeling.

REFERENCES

I. Yoonjung Choi, Janyce Wiebe, and Rada Mihalcea. "Coarse-grained +/-Effect Word Sense Disambiguation for Implicit Sentiment Analysis ".Journal Of Latex Class Files, Vol. 14, No. 8, August 2015

II. Pallavi Tasvir Shende and AmitPimpalkar. "Opinion Analysis of Public Sentiment on Social Websites using Web Mining and Natural Language Processing ".Journal Of Latex Class Files, Vol. 14, No. 8, August 2015, E-ISSN: 2349-7610

III. Taimur Islam, Ataur Rahman Bappy, Tanzila Rahman, and Mohammad Shorif Uddin. "Filtering Political Sentiment in Social Media from Textual Information".

IV. Cristina Bosco, Viviana Patti and Andrea Bolioli, "Developing corpora for sentiment analysis and opinion mining: the case of irony and Senti-TUT", IEEE Intelligent Systems, 2013.

V. Amit Pimpalkar, Tejashree Wandhe, M. Swati Rao and MinalKene "Review of Online Product using Rule Based and Fuzzy Logic with Smiley's", International Journal of computing

and technology (IJCAT), Volume 1, Issue 1, February 2014.

VI. Rathawut Lertsuksakda, Ponrudee Netisopakul and Kitsuchart Pasupa "Thai Sentiment Terms Construction using the Hourglass of Emotions", 6th International Conference on Knowledgeand Smart Technology (KST),2014

VII. Aditi Gupta, Karthik Sondhi, NishitShivhre and Raunaq Kumar, "Sentiment Analysis for Social Media", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 7, July 2013.

VIII. A. Mudinas, D. Zhang, and M. Levene, "Combining lexicon and learning based approaches for concept-level sentiment analysis", Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining, no.5, pp. 1-8, 2012

IX. Hemalatha, Dr. G. P Saradhi Varma and Dr. A. Govardhan, "Sentiment Analysis Tool using Machine Learning Algorithms", International Journal of Emerging Trends &Technology in Computer Science (IJETTCS), Volume 2, Issue 2, March – April 2013