Sentiment Analysis of Twitter Data Using Hadoop

Milind B. Wardule[1], Vikram N. Hon[2], Tushar D. Changan[3], Vinay A. Adhav[4]
[1, 2, 3, 4] Zeal College of Engineering & Research,Pune-41

Abstract- *Twitter is most widely used for displaying various opinion of people, making it a valuable platform for tracking and analyzing public sentiment. Previous research mainly focused on modeling and tracking public sentiment. In this work, we move a step further to interpret sentiment variations. We observed that emerging topics within the sentiment variation sessions are highly related to the genuine reasons behind the variations .In the past few years, there is more growth in the use of platforms such as Twitter. Influenced by that growth, companies, online sites, media are ways to mine Twitter for information about what people think about their products .We download messages for a particular hash tag and perform sentiment analysis to find positive, negative or neutral sense of that tweet using hadoop . Each hash tag may have 1000 of comments and new comments are added in a minute, in order to handle so many tweets we use apache hadoop framework.*

We use a k-means and Porter Stemming Algorithm for twitter sentiment analysis. Hadoop is mostly used for data processing on data and analyzing. K-means algorithm is used for classification of twitter data. Porter stemming algorithm is used for removing suffixes by automatic means is an operation which is useful in the field of information retrieval.

Keywords- hadoop, hdfs, Public Sentiment, Tweets Extraction and Preprocessing, Sentiment Variation Tracking

I. INTRODUCTION

Today most of the things is totally based on Internet. Now a day's people can not imagine life without Internet and social media. Also, OSNs are also a part of modern life. From last few years people share their views, ideas with each other using social media sites . Social Media is the most significant Information exchange technology of the 21st century. All people ages use social media to share they views and opinions with friends or the wider social web. Social media, such as Twitter can share their views and opinions of users related to any topics. Sentiment analysis of social media data may be of interest to different public sector organizations, i.e. in the security and law enforcement sector.

Twitter data can be used to identify the correlation between public and market . This can be achieved through a flexible system that allows users to customize the filtering criteria to be applied to their walls, and a Machine Learning-based soft classifier automatically labeling messages in support of content-based filtering. Previous research like O'Connor et al. focused on tracking public sentiment on Twitter and studying its correlation with consumer confidence and presidential job approval polls. They reported that real time events have a significant effect on the public sentiment on Twitter.

Early works is to divide entire document as positive or negative polarity, or rating scores of reviews. Such systems were mainly based on supervised approaches relying on manually labeled samples, such as movie or product reviews where the opinion is overall positive or negative attitude was explicitly indicated. However, opinions and sentiments do not occur only at document-level, nor they are limited to a single valence or target. Contrary attitudes toward the same topic or multiple topics can be present across the span of a document. However, none of these studies performed further analysis to mine useful insights behind significant sentiment variation, called public sentiment variation. Important decision-making information can be done using sentiment analysis. For example, if negative sentiment towards Arvind Kejrival increases significantly, the White House Administration Office may be tries to know why people have changed their opinion and then react accordingly to change this trend. Another example is, if public sentiment changes on some products, companies may want to know why their products receive such feedback.

1.1 Related work

Many companies now a day working on sentiment analysis topic. Almost all areas need sentiment analysis od data ,e.g. political party need sentiment analysis for some event taken by party, company needs sentiment analysis for their product etc. User of twitter twits on some event or some issue or some problem happen. But now a day people uses emoticons for showing their views or opinion. The emoticons based sentiment analysis is not happen in previous work.

Also some words which is sense to be positive are calls at some negative sense, this is also difficult to find in sentiment analysis. Some people uses short form in messages this is also challenging task to find sentiment analysis of these words. Sentence which is written by user may be not follow grammatical rule this is also challenge .So there is tremendous scope for sentiment analysis.

II. PROPOSED SYSTEM

Proposed method uses an twitter4j API for downloading or fetching the live twits from twitter. After fetching the live twits the sentiment analysis is done on that messages. preprocessing part contains remove URL. Remove @, remove stop wards and finally sentiment analysis. After the sentiment analysis is done the result is stored in the form of graph and stored on the database.

System is mainly divided into two part first part is the analyst part and second part is the training data.

Analyst Part

Analyst part contains fetching live twits from twitter, data normalization, parts of speech(POS),feature extraction, calculating overall rating

1. Fetch Twitter Live Stream

We can fetch the live twits from twitter by using twitter4j API, and these twits are used for sentiment analysis. We does sentiment analysis for particular hash tag also.

2. Data Normalization

In data normalization we use hadoop for processing on data. So in pre-processing system does following operations

- Removing URL
- Removing @
- Removing stop wards

3. Calculating overall Rating

In overall rating the opinion of people is calculated i.e. we find that whether public opinion is positive, negative, or neutral.

Training Data

Training data set means data provided to the system by default. It contains Twitter4j API, AFINN Dictionary, Hash Tagged Dataset, Emoticons Dataset, Stanford NLP.

1. Twitter4j API

It is a application program interface used for fetching live twitter streams from twitter.

2. AFINN Dictionary

AFINN Dictionary is a dictionary where wards and their weight are stored i.e. positive or negative.

3. Hash Tagged Dataset

Hash tagged dataset contains downloaded twits which is related to some particular hash tag. To create the hash tagged data set, we first filter out duplicate tweets, non-English tweets, and tweets that do not contain hashtags. From the remaining set (about 4 million), we investigate the distribution of hashtags and identify what we hope will be sets of frequent hashtags that are indicative of positive, negative and neutral messages. These hashtags are used to select the tweets that will be used for development and training.

4. Emoticons Dataset

This contains emoticons and their weight. The Emoticon data set is created by collecting tweets with positive ':)' and negative ':(' emoticons. In this approach messages are classified based on positive and negative emoticons.

2.1. Algorithms

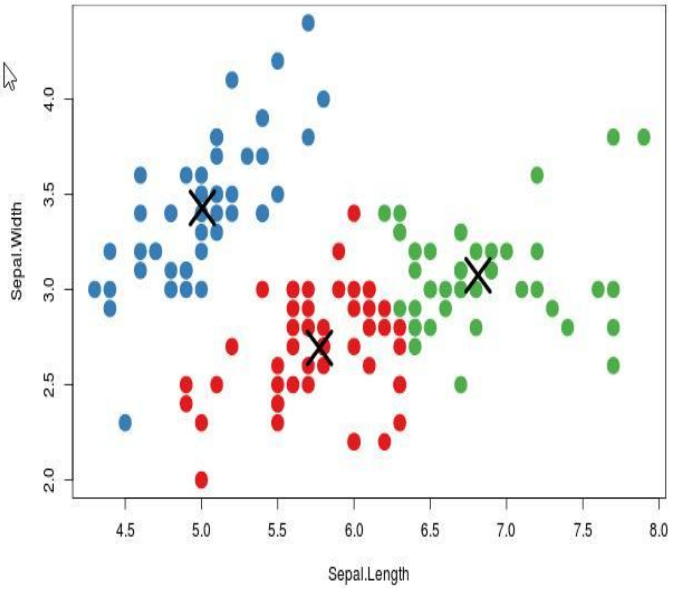
1. K-means Clustering

K-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. K-means clustering aims to partition observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells.

Given a set of observations (x_1, x_2, …, x_n), where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k sets ($k \leq n$) $S = \{S_1, S_2, \ldots, S_k\}$ so as to minimize the within-cluster sum of squares (WCSS):

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$

Where, μ_i is the mean of points in S_i.



2. Porter Stemmer Algorithm

Porter stemming algorithm (or '**Porter stemmer**') is a process for removing the commoner morphological and inflexional endings from words in English. Following are the steps of this algorithm:-

- Gets rid of plurals and -ed or -ing suffixes
- Turns terminal y to i when there is another vowel in the stem
- Maps double suffixes to single ones: -ization, -ational, etc.
- Deals with suffixes, -full, -ness etc.
- Takes off -ant, -ence, etc.

Removes a final –e

III. CONCLUSIONS

In this paper, new approach for sentiment analysis of twitter data is used . To avoid the drawbacks of absence of emoticons based sentiment analysis we propose the system which also support for emoticons based sentiment analysis. For processing on data we use new technology namely hadoop. For proper working of system we use two algorithm, first is the k-means clustering algorithm and second is the porter stemmer algorithm. By using this system we find the public opinion for some event or some issue i.e. it is positive, negative or neutral.

ACKNOWLEDGEMENT

Prof. Gore

REFERENCES

[1] Barbosa, L., and Feng, J. 2010. Robust sentiment detection on twitter from biased and noisy data. In Proc. of Coling.

[2] Bifet, A., and Frank, E. 2010. Sentiment knowledge discovery in twitter streaming data. In Proc. of 13th International Conference on Discovery Science.

[3] Davidov, D.; Tsur, O.; and Rappoport, A. 2010. Enhanced sentiment learning using twitter hashtags and smileys. In Proceedings of Coling.

[4] Esuli, A., and Sebastiani, F. 2006. SentiWordNet: A publicly available lexical resource for opinion mining. In Proceedings of LREC.

[5] Hatzivassiloglou, V., and McKeown, K. 1997. Predicting the semantic orientation of adjectives. In Proc. of ACL.

[6] Jansen, B. J.; Zhang, M.; Sobel, K.; and Chowdury, A. 2009. Twitter power: Tweets as electronic word of mouth. Journal of the American Society for Information Science and Technology 60(11):2169–2188.