Sentiment Analysis: Opinion Mining of Positive, Negative or Neutral Twitter Data Using Hadoop

[1] Komal Sutar, [2] Snehal Kasab , [3] Sneha Kindare, [4] Pooja Dhule

[1,2,3,4] Computer Department, SPPU, India

Abstract – Social Networking Service (SNS), is a platform to provide social relations among individuals who share common interest. Twitter has become very popular. Millions of users post their comments on twitter; they specify their view on current affairs. Daily large amount of row data is available and which can be helpful for industrial or business purpose. Hence the twitter data can be analyzed and used for different businesses which will helpful for decision making. This paper gives a way of analysis of twitter data using AFFIN, EMOTICON **for natural language processing. To store, categories & process large sentiments we are using Hadoop an open source framework.**

Keywords - *Sentiment Analysis, Stanford NLP, AFFIN, EMOTICON, Twitter4j API.*

1. Introduction

Due to nature of microblogging sites evolved to become a source of various kind of information. The explosive growth of user generated comments, Twitter has become a social site where lots of people can exchange their judgment & opinion about any current issues. Sentiment analysis of twitter data is providing an effective way to expose user opinion which is necessary for decision making in various streams or fields. Sentiment analysis is opinion mining. It is an analysis of feelings that is attitude, emotions & judgment behind word using NLP.

Sentiment analysis of twitter data is providing an effective way to expose user opinion which is necessary for decision making in various streams. Twitter allows the user to post real time short messages called as tweets. These tweets are restricted to 140 characters in length. Also it allows the user to use hashtag which is used to mark topics. For each hashtag, there may be lots of tweets & new tweets are generated every minute. So in order to handle so many tweets, we are using hadoop framework. So using hadoop, we analyze twitter data where cluster of positive, negative & neutral sense will be formed. The Hadoop framework was designed to solve problems which had huge amount of data for processing. Hadoop is distributed processing framework, which is work on large unstructured data. It is also work on semi structured & structured data. Twitter data is relatively unstructured data. So it is store using hadoop. Hadoop is write ones & read many times architecture. It is not suitable for online transaction. It is divides large data into 128 MB data chunks.

In hadoop, code is distributed over the slave machine & result is send to master machine. It increases efficiency & decreases data failure or data lost. Data is distributed to nodes using HDFS. HDFS is Hadoop’s distributed file system. Hadoop provides MAP Reduce framework which contains name node, secondary name node, data node & HDFS file system which contains job tracker, task tracker. Map Reduce is used for processing & HDFS file system is used for storage. With the use of sentiment analysis any industry can know the feedback of people about their product and can improve their quality of product.

2. Literature Survey

Sentiment analysis is a rising area of Natural Language Processing with research. Starting from being a documentation level (Turney, Pang and Lee) [1][2], they both calculate the summation of polarity of the adjectives and adverbs contained within text. Given the character limitations on tweets, classifying the sentiment of Twitter messages is most similar to sentence-level sentiment analysis [3][4]; and more recently there is phrase level [5]they were senses the sentiment words may have some different senses [7][8][9] uses lexical resources and decode whether a sentence expresses a sentiment by the present of lexical items. thus word sense disambiguation can improve sentiment analysis systems [6].Another line of works aims at identifying a wide range of sentiment classes expressing various emotions such as happiness, sadness, boredom, fear, gratitude, regardless, positive or negative rating.

For this Mihalcea and Liu [10] derive lists of words and phrases, and some terminology with happiness factor from a collection of blog posts, where each post is illustrated by the blogger with a mood label. Mishne [11] used ontology of over 100 moods assigned to blog posts to classify those blog texts according to expressions. While [11] Mishne classifies a blog entry, Mihalcea and Liu [10] appoint a happiness factor to specific words and expressions. Some of the recent and modern results on sentiment analysis of Twitter data are by the Go et al. [12], and Pak and Paroubek [13]. Go et al. [12] use distant learning to cultivate sentiment data. They use tweets which are ending in positive emoticons as positive or smilies and negative or sad emoticons as negative. They customize models using Naive Bayes, with parts-of-speech (POS) features. Pak and Paroubek [13], they perform a different classification task through subjective contrast objective. For subjective data they collect the Tweets which are annulling with emoticons in the same like manner as Go et al. [12] derived. For objective data they crawl twitter accounts of popular newspapers like “New York Times”, “Washington Posts” etc.

In twitter the people write short messages in which sometimes they make some spelling mistakes, so for that we are using emoticons dictionary and affine dictionary to identify overall weight of word. Davidov et al. [14] analyze the use of the #sarcasm hashtag and its contribution to automatic recognition of sarcastic tweets. To the best of our knowledge, there are no works employing Twitter hashtags to learn a wide range of emotions.

The Apporv Agarwal [14] gives detail description about resource and pre-processing of data. Data preprocessing consists of three steps: 1) tokenization, 2) normalization, and 3) part-of-speech (POS) tagging. The Real time sentiment analysis of twitter data using hadoop(Sunil B. Mane, Yashwant Sawant, Saif Kazi, Vaibhav Shinde)[16], so the work in this area include using some mathematical approach in which they uses a formula for the sentiment values depending on the proximity of the words with adjectives, and also used the hadoop cluster for distributed processing.

Introduction, Tweets Extraction and Preprocessing (Shulong Tan, Yang Li, Huan Sun, Ziyu Guan)[17] Tracking the tweets that is it involved three steps for this First, extract tweets related to interested Targets (e.g., “Modi”, “Amitabh Bacchan” *etc*), and preprocessed the extracted tweets to make them more appropriate for sentiment analysis. Second, assign a sentiment label to each individual tweet, finally, based on the sentiment labels obtained for each tweet; track the sentiment variation regarding the corresponding target using some descriptive statistics.

3. Data Description

As we are doing sentiment analysis on Twitter Data, so we are using tweets as a data for analysis. Tweets are nothing but a real time messages posted by users on the Twitter. As micro-blogging service provides the way to post short and quick messages. Because of that the users use emoticons and make spelling mistakes and also use some special characters which will express the emotion of that user (sense of the sentence).

Following terminologies are used by users in tweets:

1) *Hashtags:* These hashtags are used to give importance to a particular topics or it is used to mark topics.

2) *Target*: In tweets “@” symbol is used to give refer to other user on the twitter (i.e. referring to).

3) *Emoticons*: This is the data in which users express their feelings. That is these are the facial expressions which are represented in pictorial form using punctuations and (i.e. they express the mood of the people).

The process of doing sentiment analysis is as follows:

1) Fetch the live data from Twitter.
2) Then form a text file on Hadoop.
3) Distribute the process on master and slave processors by Hadoop.
4) Analyze the data on slave processors by using data normalization, POS, emoticon dataset, AFFINE dictionary.
5) Gather the all data at master processor and generate the result in positive, negative, neural sense.

4. Architecture

This is the architecture of Twitter Sentiment Analysis. It contains following three components:

- Analyst
- Hadoop (big data)
- Training Data

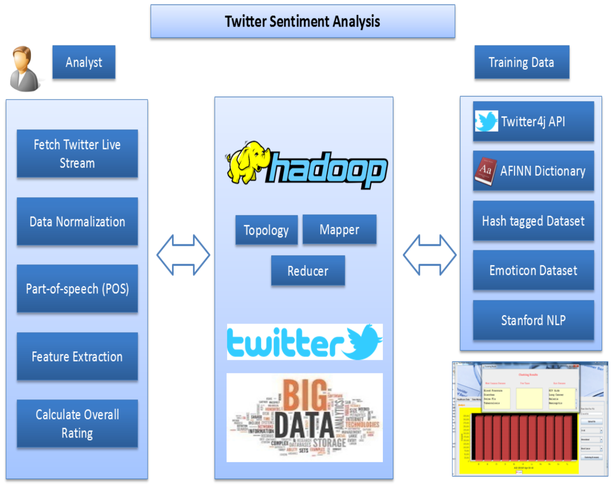

Fig.1 System Architecture

At the analyst we will be performing following operations:

First we are fetching the live twitter data by using the Twitter4j API. Twitter4j is an API which provides the way to access the tweets on twitter. It is an unofficial Java library. We can easily integrate our application in Java with the twitter service using Twitter4j which is present at training data in architecture.

After fetching the data, we will form a text file for a particular hash tag data, it means we will be giving input as a hash name of person or any another thing (e.g. #modi, #BSNL). Then after forming of the files, these files are given to Hadoop for analysis. Hadoop automatically distributes data on master and slave processors.

Then the actual process of analysis is started by using training data and database connectivity at the back-end. After fetching the data, the data normalization will performed on the data which involves the following steps to clean the data:

- *Tokenization:* All sentences will convert into tokens.
- *Remove @:* It will remove all the "@" symbol from all input file.
- *Remove URL:* It will remove the entire "URL" from all input file by scanning all input file
- *Remove stop words*: As we have to analyze the specific sense of sentences, then we have to remove all the stop words like is, the, and etc. And keeps only the words which having sense, noun, adjectives from the all files of input. To remove stop words we will use the "Stanford NLP" from training data.

After this POS (part-of-speech) will be apply on the data. In this POS detects that if the term is verb, adjective or a noun. Then we will apply Stanford NLP to figure out the POS from given preprocessed data.

This separated data will apply to TFIDF (term frequency-inverse document frequency) algorithm to calculate the term frequency of the words. Because to avoid the time wastage in analysis of same words again and again. It will give the term frequency of words.

After this, Porter Stemming Algorithm will be used to find the root of the words. After finding the root, we calculate the rating of words by using the "AFFINE" dictionary and compare the weight. If the sentences are having emoticons then we also compare the weight of emoticons by using emoticon dataset.

Then calculate the overall weight of AFFINE and Emoticon approach. After this sum up both the weights and generate the result in positive, negative, neural sense for given input hashtag dataset. To generate result in positive, negative, neural sense we will apply "K-Means" algorithm to output data, which will creates clusters of positive, negative, neural sense & finally result will be generated.

5. Conclusions

In this paper, we introduced a new technique to do sentiment analysis of twitter data. It will give us an effective output which is easy to understand. This application is very useful for decision making in various domains. And because of HADOOP it becomes easy to process the data in less time.

References

[1] Turney, P. (2002). Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews ACL.

[2] Pang, B. and Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity analysis using subjectivity summarization based on minimum cuts ACL.

[3] Yu, H. and Hatzivassiloglou, V. (2003). Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion

sentences. Conference on Empirical methods in natural language processing, 10:129–136.

[4] Kim, S. M. and Hovy, E. (2004) Determining the sentiment of opinions. Coling.

[5] Wilson, T., Wiebe, J., and Hoffman, P. (2005). Recognizing contextual polarity in phrase level sentiment analysis. ACL.

[6] Akkaya, Cem, Janyce Wiebe, and Rada Mihalcea 2009. Subjectivity word sense disambiguation. In EMNLP.

[7] Wiebe, Janyce M. 2000. Learning subjective adjectives from corpora. In AAAI.

[8] Riloff, Ellen. 2003. Learning extraction patterns for subjective expressions. In EMNLP.

[9] Whitelaw, Casey, Navendu Garg, and Shlomo Argamon. 2005. Using appraisal groups for sentiment analysis. In CIKM.

[10] Mihalcea, Rada and Hugo Liu. 2006. A corpusbased approach to finding happiness. In AAAI 2006 Symposium on Computational Approaches to Analysing Weblogs. AAAI Press.

[11] Mishne, Gilad. 2005. Experiments with mood classification in blog posts. In Proceedings of the 1st Workshop on Stylistic Analysis Of Text.

[12] Go, A., Bhayani, R., and Huang, L. (2009). Twitter sentiment classification using distant supervision. Technical report, Stanford.

[13] Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining.Proceedings of LREC.

[14] Davidov, D., O. Tsur, and A. Rappoport. 2010. Semi-supervised recognition of sarcastic sentences in twitter and amazon. In CoNLL

[15] Apoorv Agarwal Boyi Xie Ilia Vovsha Owen Rambow Rebecca Passonneau ”Sentiment Analysis of Twitter Data” Department of Computer Science Columbia University New York, NY 10027 USA fapoorv@cs, xie@cs, iv2121@, rambow@ccls, becky@csg.columbia.edu 2011

[16] Sunil B. Mane, YashwantSawant, SaifKazi, VaibhavShinde ”Real Time Sentiment Analysis of Twitter Data Using Hadoop” College of Engineering, Pune International Journal of Computer Science and Information Technologies 2014

[17] Shulong Tan,Yang Li, Huan Sun, Ziyu Guan, XifengYan,Member.IEEE,JiajunBu,Member,IEEE ChunChen ,Member, IEEE, and XiaofeiHe, Member” Interpreting the Public Sentiment Variations on Twitter” , IEEE 2014

[18] Ryan M. Eshleman and Hui Yang reshlema@mail.sfsu.edu,Hey 311, come clean my street! A Spatio-temporal Sentiment Analysis of Twitter Data and 311 Civil Complaints huiyang@sfsu.edu Department of Computer Science San Francisco State University 1600 Holloway Avenue, San Francisco, CA, USA, IEEE 2014

[19] Erik Cambria Temasek Laboratories ”An Introduction to Concept-Level Sentiment Analysis” , National University of Singapore Springer-Verlag Berlin Heidelberg 2013 cambria@nus.edu.sghttp://sentic.net

[20] S Anna Jurek 1,2, Yaxin Bi2, Maurice Mulvenna 2 1 RepKnight Limited, 37A Upper Dunmurry Lane, Belfast, BT17 0AJ ”Twitter Sentiment Analysis for Security-Related Information Gathering” , UK 2 School of Computing and Mathematics, Faculty of Computing and Engineering, University of Ulster, BT37 0QB, UK anna.jurek@repknight.com (jurek-a@email.ulster.ac.uk),y.bi@ulster.ac.uk,md.mulvenna @ulster.ac.uk ,IEEE 2014