ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 12, December 2014

Decision Making Using Sentiment Analysis from Twitter

M.Vasuki[1], J.Arthi[2], K.Kayalvizhi[3]

Assistant Professor, Dept. of MCA, Sri Manakula Vinayagar Engineering College, Pondicherry, India[1]

MCA Student, Sri Manakula Vinayagar Engineering College, Pondicherry, India[2]

MCA Student, Sri Manakula Vinayagar Engineering College, Pondicherry, India[3]

ABSTRACT: More Social Networks used as a platform for millions of users to share their thoughts and opinions about different aspects so dataset become larger. For example Twitter is a rich source of information for decision making using Sentiment analysis. Sentiment analysis to provide better decision making given to particular person, product or any service**.** Sentiment analysis and Hadoop tool has been used to process the dataset. Sentiment analysis over the Twitter offers organizations a fast and effective way to monitor the publics' feelings towards their brands. It was focuses to predict the polarity of words and then classify them into positive and negative feelings with the aim of identifying attitude and opinions that are expressed in any form or language. The proposed method includes various pre-processing steps before feeding the text to the classifier and Map reduce algorithm used for getting accurate decision about product from different perception of feedback from users. Experimental results show that the proposed technique overcomes the previous limitations and achieves higher accuracy when compared to similar technique.

KEYWORDS: Big Data, Twitter, Hadoop Tool, Sentiment Analysis, Map/Reduce.

I. INTRODUCTION

1.1The emergence of big data

A recent McKinsey report has referred to big data as "data sets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyse" Such data come from everywhere: pictures and videos, online purchase records, and geo-location information from mobile phones. Big data are not just about sheer volume in terabytes though. Other important aspects have been emphasized in addition to volume, including variety, velocity and value big data may be unstructured too: examples are text with social sentiments, audio and video, click streams, and website log files. Such data may flow in real-time streams for analysis, which can enable a firm to maximize business value by supporting business decisions in near to real-time.

This new trend in decision support is evocative of what we saw in the 1990s with the emergence of data mining, and the new emphasis on data with a large number of dimensions and much higher complexity (e.g., spatial, multimedia, XML and Internet data). Most of the datasets were "one off" opportunities, rather than data that had become available due to systemic and technological advances. Considerable challenges are present in the quest to capture the full potential of big data.

1.2 Motivation

The explosive growth of the textual information on the web in the past few decades has brought radical change in human life. In the web, people share their opinions and sentiments. In many forms about products or services the services they are aware of. This creates a large collection of opinions and views in the form of texts, which needs to be analysed to know the efficacy of the product or service. Opinions are a usually subjective expression that describes person's sentiment, feelings towards the object or service.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 12, December 2014

The sentiment can be positive or negative. This survey is a summary of the work on sentiment analysis, covering the new challenges which appear in sentiment analysis as compared to traditional fact based analysis, covering the new challenges for sentiment analysis. Those subjectivity classification, word sentiment classification, document sentiment classification and decision extraction. There are several challenges in decision making using sentiment analysis. The first is an opinion word that is considered to be positive in one situation may be considered negative in another situation. Another challenge is that people don't always express their thoughts in same way. Most traditional text processing. People can be contradictory in their statements. Sentimental analysis over the Twitter data and other similar micro-blogs faces several new challenges due to the typical short length and irregular structure of such content,

- The named language processing method of extracting entities such as people, organizations and locations from Twitter corpus.
- Anaphora is resolving the problem a word or phrase to or stands for a later word or phrase.
- The process of identifying the subject and object of the sentence. The verb and adjective are referring to what?
- The verb actually stands for? Does 'bad' mean bad or good?
- Insufficient data or very few useful labels in the training set.
- Twitter abbreviations, poor spellings, poor punctuation, poor grammar, incomplete sentences.
- The accuracy of tweets classification as compared to human judgment.

1.3. Decision Making

When we search the meaning of Decision Making in dictionary we can find the following meaning:

- The action or process for making important decision.
- The thought process of selecting the best among the different alternatives choice from selective option. When trying to make words decision is a subjective belie good decision.

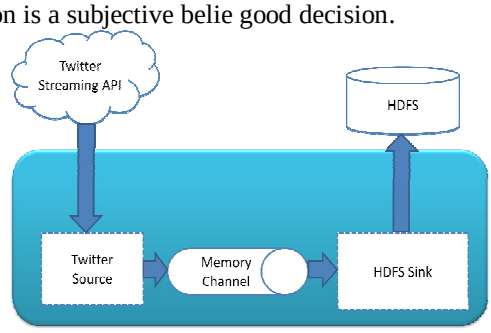

Fig. 1 Overall Process Diagram

II. RELATED WORK

There is multiple text mining technique used to mine the twitter feeds.

Ohbyung Kwon, Namyeon Leea.[1] Showed that sentiment analysis of tweets is a challenging task due to multilingual and informal messages. In this study, a research model is proposed to explain the acquisition intention of big data analytics mainly from the theoretical perspectives of data quality management and data usage experience. Our empirical investigation reveals that a firm's intention for big data analytics can be positively affected by its competence in maintaining the quality of corporate data.

William D. Abilhoa, Leandro N. de Castro. [2]This paper proposes a keyword extraction method for tweet collections that represents texts as graphs and applies centrality measures for finding the relevant vertices (keywords). To assess the performance of the proposed approach, three different sets of experiments are performed. The first experiment applies TKG to a text from the Time magazine and compares its performance with that of the literature. The second set of experiments takes tweets from three different TV shows, applies TKG and compares it with TFIDF and KEA, having human classifications as benchmarks. Finally, these three algorithms are applied to tweets sets of increasing size and their computational running time is measured and compared. Altogether, these experiments provide a general overview of how TKG can be used in practice, its performance when compared with other standard

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 12, December 2014

approaches, and how it scales to larger data instances. The results show that TKG is a novel and robust proposal to extract keywords from texts, particularly from short messages, such as tweets.

Nadia F. Silva, Eduardo R. Hruschka, Estevam R. Hruschka Jr. [3] in this paper, we introduce an approach that automatically classifies the sentiment of tweets by using classifier ensembles and lexicons. Tweets are classified as either positive or negative concerning a query term. This approach is useful for consumers who can use sentiment analysis to search for products, for companies that aim at monitoring the public sentiment of their brands, and for many other applications. Indeed, sentiment classification in micro blogging services (e.g., Twitter) through classifier ensembles and lexicons has not been well explored in the literature. Our experiments on a variety of public tweet sentiment datasets show that classifier ensembles formed by Multinomial Naïve Bayes, SVM, Random Forest, and Logistic Regression can improve classification accuracy.

Min-Chul Yang, Hae-Chang Rim [4] showed that identifying interesting and useful contents from large text-streams is a crucial issue in social media because many users struggle with information overload. Retweeting as a forwarding function plays an important role in information propagation where the retweet counts simply reflect a tweet's popularity. However, the main reason for retweets may be limited to personal interests and satisfactions. In this paper, we use topic identification as a proxy to understand a large number of tweets and to score the interestingness of an individual tweet based on its latent topics. Our assumption is that fascinating topics generate contents that may be of potential interest to a wide audience. We propose a novel topic model called Trend Sensitive-Latent Dirichlet Allocation (TS-LDA) that can efficiently extract latent topics from contents by modelling temporal trends on Twitter over time. The experimental results on real world data from Twitter demonstrate that our proposed method out performs several other baseline methods.

Arman Khadjeh Nassirtoussi. [5]The quality of the interpretation of the sentiment in the online buzz in the social media and the online news can determine the predictability of financial markets and cause huge gains or losses. That is why a number of researchers have turned their full attention to the different aspects of this problem lately. However, there is no well-rounded theoretical and technical framework for approaching the problem to the best of our knowledge. We believe the existing lack of such clarity on the topic is due to its interdisciplinary nature that involves at its core both behavioural-economic topics as well as artificial intelligence. We dive deeper into the interdisciplinary nature and contribute to the formation of a clear frame of discussion. We review the related works that are about market prediction based on online text- mining and produce a picture of the generic components that they all have. We, furthermore, compare each system with the rest and identify their main differentiating factors. Our comparative analysis of the systems expands onto the theoretical and technical foundations behind each. This work should help the research community to structure this emerging field and identify the exact aspects which require further research and are of special significance.

All the techniques discussed in this section have some advantages and limitations. Hence a comprehensive technique is still needed to overcome their limitations.

III. SENTIMENT ANALYSIS

Sentiment analysis is also known as opinion, use of language processing, text analysis and computational linguistic to identify the extract subjective information in source materials. Micro-blogging website Twitter has evolved to become a source of rich and varied information. It is due to nature of micro-blogs on which people post real time messages about their opinions on a variety of topics, discuss about current issues, complaints, etc,. A key problem in this area is sentimental analysis, where a document is labelled as good or bad that is positive or negative evaluation of product. The sentiment may be his or her judgment, mood or evaluation.

3.1. Tweets classifications

The main contribution of this paper as follows,

- Introduces and implements a hybrid approach for determining the sentiment of each tweet.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 12, December 2014

- Demonstrates the value of pre-processing data using detection and analyze language, slang, abbreviation, correction, grammatical mistakes and stop words removal.
- Tests the accuracy of sentiment identification on Tweets datasets. And produce average result.
- Compare with other technique to prove the effectiveness of the proposed hybrid approach.

3.2. Sentiment Classification

Sentiment analysis, also referred to Decision making system, implies extracting opinions, emotions, and sentiments in text. As we can imagine, one of the most common applications of sentiment analysis is to track attitudes and feelings on the web, especially for tracking products, services, brands. The main idea is to determine whether they are viewed positively or negatively by a given audience.

3.2.1. Classify emotion

There are multiple techniques used to classify the Twitter feeds. The emotion classification is used to analyze some text and classify it in different types of emotions, such as **anger, disgust, fear, joy, sadness, and surprise**.

3.2.2Classify decision

In contrast to classification of emotions, allows us to classify some text as **positive or negative**. In this case the classification is done by using **Map Reduce Algorithm.** The technical issue in proposed approach is, less accuracy of sentiment analysis, difficulty in tackling sentiment analysis of Twitter stream for longer time and weak emotion representation. The proposed algorithm focuses on classification of data streams and performs sentiment analysis in real time.

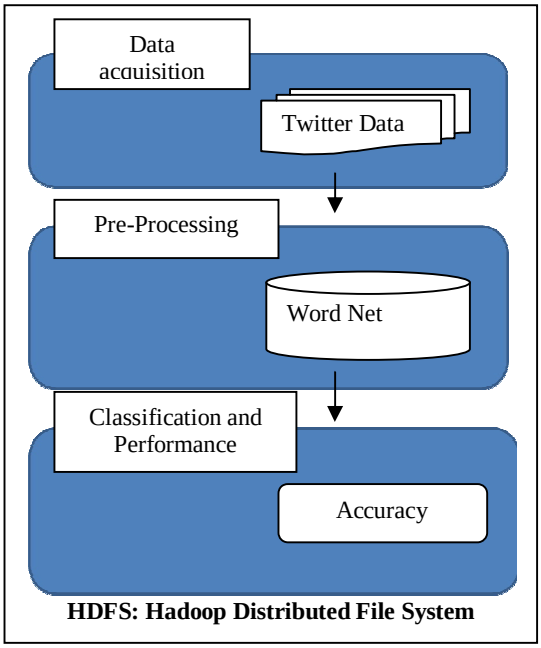

Fig. 2 Classification Framework

IV. MAP REDUCE ALGORITHM

Map Reduce is a programming model used by sentiment analysis large amount of data in a distributed computing environment using Hadoop. It is usually used to perform distributed computing on clusters of user opinion.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 12, December 2014

Computational processing of data stored on either a file system or a database usually occurs. Map Reduce takes the advantage of locality of data, processing data on or near the storage areas, thereby avoiding unnecessary data transmission like clumsy information, n-grammar.

```
Map (String key, String value)
// key: document name
// value: document contents
for each word w in value:
        Return (w, "1");
Reduce (String key, Iterator values)
// key: a word
// values: a list of counts
Int result = 0;for each v in values:
        Result += ParseInt(v);
    Return (AsString (result));
```

Fig. 3 Proposed Map Reduce for Sentiment Analysis

The simplicity of the programming model and the automatic handling of node failures hiding the complexity of fault tolerance make Map Reduce to be used for sentiment analysis. In order to achieve good performance a Map Reduce scheduler must avoid unnecessary data transmission. Hence different scheduling algorithms for Map Reduce are necessary to provide good performance. Decision making and sentiment analysis is a technique to detect and extract subjective information in text documents. In general, sentiment analysis tries to determine the sentiment of a writer about some aspect or the overall contextual decision of a document. The sentiment may be his or her judgment, mood or evaluation. A key problem in this area is sentiment classification, where a document is labeled as a positive or negative evaluation of a target product.

The binary classification task of labelling a document as expressing either an overall positive or negative opinion is called acquisition. Acquisition assumes that the opinionated document expresses opinions on a single target and the opinions belong to a single person. It is clear that this assumption is true for customer review of products documents which usually focus on one product and single reviewer writes it. A movie review, restaurant review, or product review consists of a document written by the review- might be more appropriate since SVMs must be exposed to a large set of data in order to build a UCI machine learning repository. Several techniques and features are used by researchers in learning process. One of the most important tasks in sentiment classification is selecting an appropriate set of features. The most commonly used features in sentiment classification are introduced below.

- Terms and their frequency: these features consist of single words or word n-grams and their frequency or presence. These features have been widely and successfully used in sentiment classification and shown quite effective for this task.
- Part of speech: Its information is a very important indicator of sentiment expression.
- For example adjectives carry a great deal of information regarding a document's sentiment.
- Decision words: Decision words (or sentiment words) and phrases are words and phrases that express positive or negative emotions. For in-stance, good, fantastic, amazing and brilliant are words with positive emotion and bad, boring, slow, worst and poor are words with negative emotion. Though almost opinion words are adjectives and adverb, nouns and verbs can also express an opinion. For example rubbish (noun), hate and like (verb) can indicate opinion in some documents.
- Negative opinion: Obviously, negation words are very important to evaluate the decision of a sentence because they can transform the sentiment orientation in a sentence. For instance, the sentence "I don't like this mobile" has negative orientation.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 12, December 2014

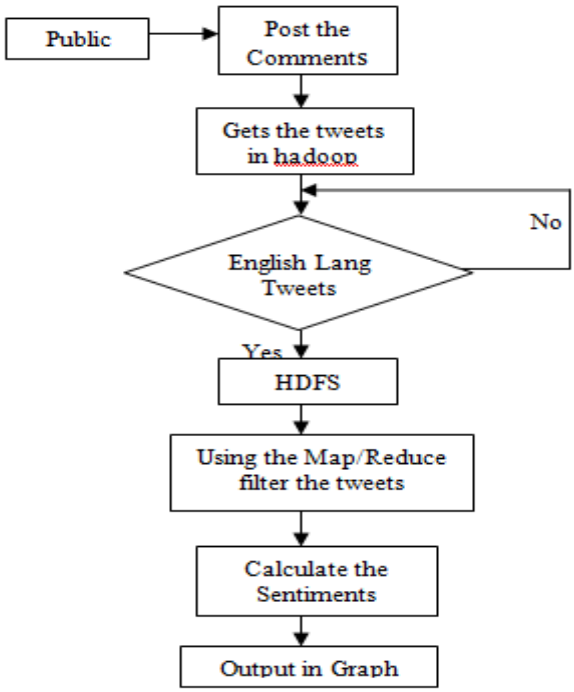

Fig. 4 Flow chart of Classification Framework

- Syntactic dependency: Several research work in this area used word dependency based features generated from dependency tree or parsing.

The main challenges are: classification accuracy, sarcasm and data sparsity problem, as they incorrectly classify most of the tweets. The reason behind these problems is use of slangs and other shorthand grammars due to the limit of tweet message (140 characters).The main goal of this research is to improve the accuracy of text classification and resolve the data sparsity issues. The core idea is to pre-process the raw data and perform different transformations to remove the slangs, grammatical mistakes, abbreviations and other noise and then feed it to the classifier. The tweets obtained from these data streams are used as input items. The proposed system is basically composed of three main modules. The first module is data acquisition, a process of obtaining twitter feeds from OSN; the second module performs pre-processing and transforms the tweets containing real valued features or arbitrary components and refines them into a stream pattern that can be easily used for subsequent analysis. The last component applies different classification techniques in a pipelined way which classifies the tweets into positive, negative or neutral. The proposed framework is shown in Fig. 2 and the flowchart of the proposed framework is given in Fig. 3.

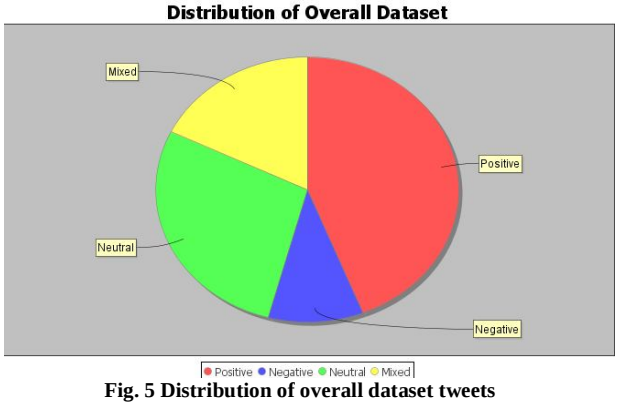

Fig. 5 Distribution of overall dataset tweets

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 12, December 2014

V. CONCLUSION AND FUTURE WORK

The research paper has proposed a new algorithm for twitter sentiment analysis and it is based on three way classification framework. Thus the product sentiment analysis provides a solution for obtaining feedback about any product and presenting the overall picture. Mostly it is the organization which is benefited through this analysis. Sentiment analysis can be further extended as document analysis and block analysis. Rather than its use since big data analysis is used its efficiency is enhanced. Large volume of data is obtained and its process is withholder by hadoop environment. Also analysing huge data with less time complexity proves its efficiency**.** The results of the proposed framework show great improvement when comparing with similar work. We have achieved an average accuracy. Future research directions include the development of a web application in order to compare the performance of our algorithm with other applications like Tweet Feel & Sentiment140 and the use of supervised learning algorithms to further increase the accuracy. So we can use Flume tool, Flume is used to obtain data from any social web site and store in HDFS and produce a dynamic result.

REFERENCES

[1] A. Cui, M. Zhang, Y. Liu, S. Ma, Emotion Tokens: Bridging the Gap among Multilingual Twitter Sentiment Analysis, Springer-Verlag, Berlin, Heidelberg, 2011, pp. 238–249.
[2] A. Bifet, E. Frank, Sentiment Knowledge Discovery in Twitter Streaming Data, Springer-Verlag, Berlin, Heidelberg, 2010, pp. 1–15.
[3] A. Bifet, G. Holmes, B. Pfahringer, MOA-Tweet Reader: real-time analysis in twitter streaming data, in: T. Elomaa, J. Hollm'en, H. Mannila (Eds.), DS 2011, LNCS 6926, Springer-Verlag, Berlin Heidelberg, 2011, pp. 46–60.
[4] http://www.edupristine.com/courses/big-data-hadoop-program/?jscfct=1
[5] http://searchbusinessanalytics.techtarget.com/definition/big-data-analytics
[6] http://efytimes.com/e1/fullnews.asp?edid=123367 Tom White, Hadoop The Definitive Guide , 1st edition, 2009.
[7] Pete Warden, Big Data Glossary ,International edition,2011.
[8] Eric Siegel and Thomas H. Davenport, Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die, Pearson edition, 2011.
[9] Willyan D. Abilhoa, Leandro N. de Castro, A keyword extraction method from twitter messages represented as graphs, 2014.
[10] Luca Cagliero, Tania Cerquitelli, Paolo Garza, Luigi Grimaudo, Twitter data analysis by means of Strong Flipping Generalized Itemsets,2014

BIOGRAPHY

Prof.Mrs.M.Vasuki is currently working as Assistant Professor in the Department of Master Of Computer Application at Sri.Manakula Vinayagar Engineering College, Madagadipet, Pondicherry, India. She received Master of Computer Application (MCA) degree in 1994 from Bharathidasan University, Karur, and completed M.Phil in 2005 from Manonmaniam Sundaranar University, Thirunelveli, and Completed M.Tech in 2012 from SRM University, Chennai, Tamil Nadu, India. Her research interests are Database Management Systems.

J.Arthi obtained her BCA degree from Pondicherry University Community College, Pondicherry, India. She is currently pursuing her MCA degree in at Sri.Manakula Vinayagar Engineering College, Madagadipet, India. Her areas of research interest accumulate in the areas of Database Management Systems and Data Mining.

K.Kayalvizhi obtained her BCA degree from Thiruvalluvar University, Tamil Nadu, India. She is currently pursuing her MCA degree in at Sri.Manakula Vinayagar Engineering College, Madagadipet, India. Her areas of research interest accumulate in the areas of Database Management Systems and Data Mining.