Analysis of Opinion Mining on Social Media Data Streams using Hadoop

Padala S. Venkata Durga Gayatri
M.Tech(CST)
Department of CSE
University College of Engineering
Adikavi Nannayya University

Archana Raghuvamshi
Assistant Professor
Department of CSE
University College of Engineering
Adikavi Nannayya University

ABSTRACT

Twitter is a social networking site in which the data to be processed is in rich amounts and which can be structured, semi-structured and unstructured data streams. Opinion mining over the Twitter offers organizations a fast and effective way to monitor the feelings of public towards their services. It focuses on predicting the polarity of words and then classifies them into positive and negative feelings with the aim of identifying attitude and opinions that are expressed in any form or language. Bian et al.’s method (2012) annotated the twitter corpus which was focused on Adverse Drug Reaction (ADR) which includes the broad pharmacological coverage. Bingwei et al.’s method (2013) evaluates the scalability of Naive Bayes classifier (NBC) in large datasets instead of using the standard library. Skuza et al.’s method (2015) estimated the future stock prices by calculating in distributed environment according to Map Reduce programming model. Mohit et al.’s method, (2014) explains how the Map – Reduce paradigm can be applied to existing Naïve Bayes algorithm to handle a large number of tweets. All these approaches say about the real-world data sets at its accuracy level by using Hadoop File System. This paper analyses all the above methods comparatively.

Keywords

Twitter, social networking sites, Navie Bayes Classifier (NBC), Map-Reduce, Hadoop File System (HDFS).

1. INTRODUCTION

In this fast growing environment and advancement of science and technology many software products, services are widely increasing. But they may either fail miserably since they haven’t satisfied the customers or satisfied well w.r.t product. The customer's reviews are Structured, Unstructured and semi-structured data which are growing rapidly on the Internet and many companies are trying to use this data to extract people’s views towards their products. Social networking sites are the ones that contain huge repositories of user-centered content which provides unique opportunities to gain reviews or opinions about the product or about the movie or about a service. Hence these are great sources of unstructured data but, such amounts of data cannot be analyzed manually.

Opinion mining is also known as sentiment analysis which is the best approach to get rid of this problem. It is fast growing topic in which most of the researchers and organizations are conducting researches in this area. Twitter has become a source of rich and varied information, whereas people posts real-time messages as their opinion on a variety of topic, whether it is good or bad i.e., positive or negative evaluation of that product. It is also related to text mining, computational linguistics and language processing in technical views. Automatic means of sentimental analysis leads to the concept of polarity [1], which is marked with each word of a sentence according to semantics.

The data which was collected from the Twitter API is parsed using the parser known as Stanford which includes a total of 215,54 phrases that are uniquely judged by 3 human judges. This new dataset obtained allows examining human sentiments and records to actual emotion.

In section 2, the list of notations used in this paper has given. Section 3 analyzes Bian et al.’s method at their accuracy level. In section4, deals with the probe of Bingwei et al.’s Navie Bayes Classifier method. Section5 describes the analysis of Skuza et al.’s method of opinion mining. Section 6 discusses the improved method of opinion mining proposed by Mohit et al. Section7 the performance analysis of all the methods are given. Finally, section 8 concludes the paper.

2. NOTATIONS

The notations used this paper is in the following Table1:

Table 1: List of Notations used

S.No	Notation	Specification of that notation
1	P(c|d)	It denotes the probability of a document d being in class c
2	P(d)	Is the constant for data set size
3	w_k	Is an individual word in the set of word W in document D
4	t_k	It is the frequency of each word w_k
5	ε	It has been termed classification result for predicting stock
6	Y_i	It denotes past value of stock at a given time x_i
7	ε_i	sentiment value calculated for a given amount of time of x_i , i=1,2,…..,n
8	β , α	The linear regression coefficients
9	p and t	mean values of stock prices over an amount of time period t

10	H and E	hypothesis and evidence
11	$P(W_k\|c)$	It is the posterior probability of word which is belonging to category c of the given trained data set
12	$P(c\|t)$, $P(c)$	The probability of tweet t being in category c and the probability prior to category c

3. ANALYSIS OF BIAN ET AL. METHOD

Bian et al. have proposed an approach to describe drug users and their potential adverse events [4] by analyzing the twitter messages using Natural Language Processing (NLP) which in turn builds Support Vector Machine (SVM) classifiers. Due to the nature of volume in the dataset (huge i.e., nearly 2 billion Tweets), the process is being conducted on High-Performance Computing(HPC) platform using Map Reduce [2], which shows the trend of big data analytics.

Methodology

Here in this method, initially, collected tweets dataset within a certain time period are organized by a time line. The raw twitter messages are dragged using this ‘Twitter’s user timeline API’ from which the specific tweet and the user information can be obtained. In this method, to evaluate the process the following few fields of a tweet are used.

1) Tweet ID- uniquely identifies each tweet

2) User Identifier

3) Timestamp of tweet

4) The tweet text

Steps involved in Analysis of Bian et al. Method of “Towards Large-Scale for Drug Related Adverse Events” [3]:

1. Extraction of tweets
2. From the derived tweets

a. Now, compare the brand and generic names etc., with the ADR (which 76 drugs related information).

b. Next, create a manual corpus.

The overall process is explained in Figure 1.

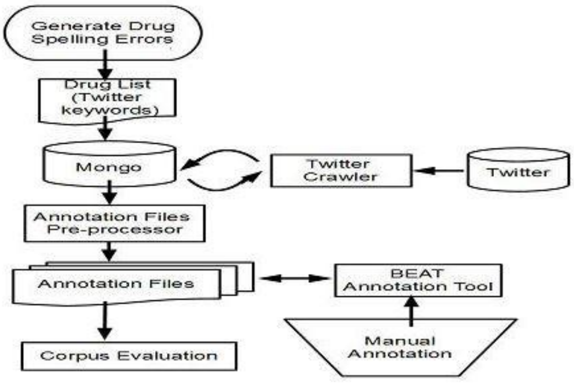

Figure 1: Overall process of data collection and annotation

To demonstrate the classification for classifiers the author has used binary classification task which consists of the following two classes for identifying and representing the tweet instances.

i) hasADR

ii) noADR

To evaluate the performance of classifiers, they have computed some of the metrics for evaluation like precision, recall, f-measure and accuracy using 10- fold cross validation.

$$\boldsymbol{precision} = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FP})}$$

$$\boldsymbol{recall} = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FN})}$$

$$\boldsymbol{Fmeasure} = \frac{2 * \mathrm{precision} \ * \mathrm{recall}}{(precision \ + recall \,)}$$

$$\boldsymbol{Accuracy} = \frac{\text{Total correct predictions}}{\text{Number of test instances}}$$

Equations 1:Formulaic equations for the classifier evaluation metrics.

As the processing is taking much memory, the author has utilized Amazon Elastic Compute Cloud (EC2) to run the Twitter indexers, simultaneously which also parses and index all the billions of tweets within two days.

Textual features such as bag-of-words (BoWs) model are derived based on actual twitter messages. Semantic features are derived from Unified Medical Language System (UMLS) which was developed at National Library of Medicine (NLM). So, in this approach author made use of this Two-class Support Vector Machine (SVM) for his way of classification.

4. ANALYSIS OF LIU ET AL. METHOD

Liu et al. have described how to predict or support decision making with high accuracy from the training datasets by using some Machine Learning technologies for sentiment classification [5]. As the datasets are very large all the algorithms [6] might not be well scaled up. So, the author evaluates the scalability of Navie Bayes classifier [7]with Hadoop framework. Instead of using Mahout Library, they have used NBC to achieve fine-grained control of analysis.

To analyze the sentiment of millions of movie reviews [8] with high throughput, the raw data is collected from research communities. In this approach, they used two large existing datasets:

a. The Cornell University movie reviews dataset3

b. Stanford SNAP Amazon movie review dataset4

The classification is scheduled into three sequential jobs as follows:

1) Training job: - All the training reviews are fed to this job to give a uniquely identified model with their own frequency in positive and negative reviews respectively.

2) Combining job: - Tested reviews are all combined to with all necessary information for the final classification.

3) Classify job: - Classify all reviews simultaneously and writes the classification result to HDFS.

Methodology

Suppose there are ‘m’ possible classes for the document, and set of unique words which occur at least once in the document in D. The complete steps involved in Bingwei et al. Method of

“Scalable Sentiment Classification for Big Data Analysis using Navie Bayes Classifier” [9] are:

1) The probability of d being in class c is computed as follows using Bayes rule: $\mathrm{P(c|d)} = \frac{\mathrm{P(c)P(d|c)}}{\mathrm{P(d)}}$ (1)

2) Now, Calculate the probability of each word occurrence in a document independently then Bayes equation is as follows : $\mathrm{P(c|d)} \propto \mathrm{P(c\,)}\prod_{k=1}^{\mathrm{n_d}}[\mathrm{P(w_k|c)}]_k^t$

3) Next, reduce the floating point underflow. $\log P(c|d) \propto \log \mathrm{P(c)}\ \sum_{\mathrm{k=1}}^{\mathrm{n_d}}[\mathrm{t_k}\ \log \mathrm{P(w_k|c)}]$ (2)

4) Maximize log P(c|d) in equation(2)

   $c^*= argmax_{c\in C}\{\log P(c) + \sum_{k=1}^{n_d}[\,t_k\ \log P(w_k|c)]\}$(3)

5) Applying Naive Bayes classifier (NBC), it can estimate P(c) and P($\mathrm{w_k}$|c) as:

   $\hat{\mathrm{P}}(\mathrm{c}) =\frac{N_c}{N}$ and $\hat{\mathrm{P}}(w_k|c)= \frac{N_{w_k}}{\sum_{w_i\in W} N_{w_i}}$

Here, N is the total number of documents, Nc is the number of documents in class c and Nwi is the frequency of a word wi in class c.

With these estimations, the calculation of the right-hand side of equation (3) is essentially a counting problem. This makes Map Reduce [2] a suitable framework for the implementation of NBC [7] in large-scale datasets.

5. ANALYSIS OF SKUZA ET AL. METHOD

Skuza et al. have described stock market based on the classification of data coming from twitter microblogging sites.

Twitter tweets are retrieved in real time using Twitter Streaming API [10]. Tweets were collected periodically which contains a specific company name or its hashtags.

Only the tweets which are in English are processed in this research work. Reposted messages are removed to avoid the redundancy in classification. Each and every word belonging to these tweets is stored in a bag of words model which is a standard technique.

In this approach, the system is designed with four components:

1) Retrieving twitter data, pre-processing and saving to database

2) Stock data retrieval

3) Model building,

4) Predicting future stock prices.

Methodology

The complete steps involved in Skuza et al. method of “Sentiment analysis of Twitter Data within Big Data Distributed Environment for Stock Prediction” [11] are:

1) Classify the methods for forecasting quantities prediction of future stock prices

2) Next, combining results of sentiment classification of tweets

3) Now, Predict stock prices from past interval and comparing with combined results. Sentiment value (**ε**) is derived as:

   $\varepsilon = \log_{10} \frac{number\ \_of\_positive\ \_tweets}{number\ \_of\_negative\ \_tweets}$ (1)

The relation for stocks taken in past analysis is given as:

$$y_i - \alpha + (\beta + \varepsilon_i)x_i \qquad (2)$$

Linear regression coefficient defined as: $\beta - \frac{\sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i}{\sum x_i^2 - \left(\frac{1}{n}\sum x_i\right)^2}$ (3)

And the α coefficient as

$$\alpha= \bar{\mathrm{p}} - \beta\,\bar{\mathrm{t}} \qquad (4)$$

Opinion mining consists of polarity mining in which input is classified either positive or negative. It is automated by using SentiWordNet. Future prediction of stocks is done by combining the results of sentiment classification of tweets as well as stock prices from a past interval.

6. ANALYSIS OF MOHIT ET AL. METHOD

Mohit et al. used Apache Hadoop framework which is open source java framework depend on Map-Reduce paradigm and Hadoop Distributed File System (HDFS) [12] to store and process the data of multi- class tweets [13].

Methodology

The complete procedure of Mohit et al. method- “Multi-Class Tweet Categorization Using Map Reduce Paradigm” [14] are:

According to Bayes’s rule, the author says that if one has hypothesis H and evidence E which bears on the hypothesis,then:

$$Pr[H|E] = \frac{Pr[E|H]*Pr[H]}{Pr[E]}$$

In the case of text classification, the classifier will model a document based on presence or absence of words in that document. Bayes classifier which considers the frequency of words. In this case, it is denoted as

$$P(c|t) \propto P(c) * \prod_{1<k<n_d} P(W_k|c)$$

Once the classifier dataset has obtained, to get the accurate results of opinion mining this classified dataset has to be further given to two Map-Reduce passes.

First Pass:

1) The mapper takes the labeled tweets from trained data and outputs as key value pair.

2) Now, the reducer groups all the instances of the words for each category and outputs a *key-valu*e pair as a category and its word-count.

Second Pass :

1) Now, the next pass returns the probability of each featured word and its category as a *key-value* pair.

2) Then the final reducer calculates final category probability for each tweet and its category as a *key-value* pair.

7. PERFORMANCE ANALYSIS:

Table 2: List of Methods and their performances

	METHOD	TECHNIQUE USED IN DEVELOPING	REMARKS
1.	Bian et al.'s method, Towards large –scale twitter mining for drug-related adverse events,2012	• It describes the events by analyzing the content of twitter messages. • Utilizes Natural Language Processing for building Support Vector Machine(SVM) classifiers.	The prediction accuracy on average over the 1000 iterations was evaluated .
2.	Liu Bingwei et al.'s method, Sentiment Classification for Big Data Analysis Using Navie Bayes Classifier,2013	• Implemented NBC to achieve fine-grain control of analysis procedure for Hadoop implementation. • Cornell University Movie review dataset3	Resulted in a 80.85% average accuracy.
3.	Skuza et al. 's method	• Discusses Stock Market Prediction. • Tweets having the name of the company or hashtag of that company name. • Naive Bayes method was chosen to employ SentiWordNet . Prediction of future stock prices.	Considered large volumes of data resulted in the decision to apply a map-reduce version of Navie Bayes algorithm.
4.	Tare Mohit et al.'s method	• Map-Reduce strategy for classification of tweets using Naive Bayes classifier.	The final reducer calculates the final probability of each category to which the tweet may belong to and outputs the predicted category and its probability .

8. CONCLUSION

Opinion mining has become a very popular field of research. There are many issues as it processes unstructured data. In this paper, we have analyzed some of the approaches of opinion mining at their accuracy level. Dictionary based approach takes less time than supervised learning approach. It can be concluded that extensive approaches will be done on real-world data sets, with an expectation to achieve comparable or greater accuracy than the existing techniques.

In future , we can plan to extend and improve this by implementing a novel method for opinion mining in MapReduce framework and the algorithm which uses the hashtags inside a tweet, as sentiment labels and proceeds to a classification procedure in distributed and parallel manner. And then compare the performance with present approaches.

9. 9. REFERENCES

[1] T. Wilson, J. Wiebe, and P. Hoffmann, “Recognizing contextual polarity in phrase-level sentiment analysis,” in Proceedings of HLT and EMNLP. ACL, **(2005)**, pp. 347–354

[2] C. C. Tao, S. K. Kim, Y. A. Lin, Y. Y. Yu, G. Bradski, A. Y. Ng and Kunle Olukotun, “Map-reduce for machine learning on multicore”, In NIPS, vol. 6, **(2006)**, pp. 281-288.

[3] B. Jiang, U. Topaloglu and F. Yu, “Towards large-scale twitter mining for drug-related adverse events”, In Proceedings of the 2012 international workshop on Smart health and wellbeing, ACM, **(2012)**, pp. 25-32.

[4] Jiang, K., & Zheng, Y. (2013). Mining Twitter Data for Potential Drug Effects. In *Advanced Data Mining and Applications* (pp. 434–443). Springer.

[5] M. Gamon, A. Aue, S. Corston-Oliver, and E. Ringger, “Pulse: Mining customer opinions from free text,” in Advances in Intelligent Data Analysis VI. Springer, 2005, pp. 121–132.

[6] U. Kang, D. H. Chau, and C. Faloutsos, “Mining large graphs: Algorithms, inference, and discoveries,” in Data Engineering (ICDE), 2011 IEEE 27th International Conference on, 2011, pp. 243–254.

[7] D. Pessemier and Martens “MovieTweetings: A Movie Reviews Dataset Collected From Twitter”, Ghent University, Ghent, Belgium, **(2013)**.

[8] M. Thomas, B. Pang, and L. Lee, “Get out the vote: Determining support or opposition from congressional floor-debate transcripts,” in Proceedings of the 2006 conference on empirical methods in natural language processing. Association for Computational Linguistics, 2006, pp. 327–335.

[9] L. Bingwei, E. Blasch, Y. Chen, D. Shen and G. Chen, “Scalable Sentiment Classification for Big Data Analysis Using Naive Bayes Classifier”, In Big Data, 2013 IEEE International Conference on, IEEE, **(2013)**, pp. 99-104.

[10] Twitter. Twitter Search API, available at https://dev.twitter.com/rest/public/search.

[11] S. Michal and A. Romanowski, “Sentiment analysis of Twitter data within big data distributed environment for stock prediction”, In Computer Science and Information Systems (FedCSIS), 2015 Federated Conference on, IEEE, **(2015)**, pp. 1349-1354

[12] T. White, “Hadoop: The Definitive Guide”, Third Edition, O'Reilley

[13] Malkani, Zahan, and Evelyn Gillie. "Supervised Multi-Class Classification of Tweets." (2012).

[14] T. Mohit, I. Gohokar, J. Sable, D. Paratwar and R. Wajgi, “Multi-Class Tweet Categorization Using Map Reduce Paradigm”, In International Journal of Computer Trends and Technology. **(2014)**, pp. 78-81.