International Journal of Research
Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 04 Issue 14
November 2017

Enactment of Sentiment Analysis in Twitter Data Using Hadoop

K.SRAVANA KUMARI

Lecturer in Computer Applications,GDC Jammikunta,Satavahana University,Telangana, India

ABSTRACT: Sentiment analysis is a broad research area in academic as well as business field. The term sentiment refers to the feelings or opinion of the person towards some particular domain. Hence it is also known as opinion mining. In this work, a method which performs classification of tweet sentiment in Twitter is discussed. To improve its scalability and efficiency, it is proposed to implement the work on Hadoop Ecosystem, a widely-adopted distributed processing platform using the Map Reduce parallel processing paradigm. Finally, extensive experiments will be conducted on real-world data sets, with an expectation to achieve comparable or greater accuracy than the proposed techniques in literature.

KEYWORDS- Twitter, Sentiment Analysis, Hadoop, Map reduce, HDFS

I. INTRODUCTION

Nowadays, the age of Internet has changed the manner humans explicit their views, evaluations. It is now particularly finished through blog posts, online forums, product review web sites, social media ,etc. Nowadays, millions of people are using social network sites like Facebook, Twitter, Google Plus, and so on. To express their emotions, opinion and share views approximately their each day lives. Through the net groups, we get an interactive media in which consumers inform and have an effect on others via boards. Social media is generating a huge quantity of sentiment rich data within the shape of tweets, status updates, weblog posts, feedback, critiques, and many others. Moreover, social media offers an opportunity for corporations via giving a platform to connect to their clients for advertising and marketing. People in general depend on person generated content over online to a fantastic volume for decision making. For e.g. If someone wants to buy a product or desires to use any service, then they firstly appearance up its critiques on-line, speak approximately it on social media earlier than taking a decision. The quantity of content generated through customers is too sizable for a ordinary consumer to research. So there is a want to automate this, various sentiment analysis techniques are widely used.

Sentiment evaluation (SA)tells consumer whether the information about the product is great or now not before they purchase it. Marketers and companies use this evaluation data to recognize about their products or services in the sort of way that it could be presented as in line with the consumer‟s requirements. Textual Information retrieval techniques particularly recognition on processing, searching or reading the genuine statistics present.

Facts have an objective factor but,there are a few other textual contents which explicit subjective traits. These contents are particularly critiques, sentiments, appraisals, attitudes, and feelings, which form the center of Sentiment Analysis (SA). It gives many tough opportunities to develop new packages, specially due to the large boom of available facts on on-line resources like blogs and social networks. For instance, guidelines of gadgets proposed by way of a advice machine may be predicted via taking into account issues including high-quality or bad evaluations about the ones gadgets through utilizing SA.

The project focuses on using Twitter, the most popular micro blogging platform, for the task of sentiment analysis. The tweets are important for analysis because data arrive at a high frequency and algorithms that process them must do so under very strict constraints of storage and time. It will be shown how to automatically collect a corpus for sentiment analysis and opinion mining purposes and then perform linguistic analysis of the collected corpus. All public tweets posted on twitter are freely available through a set of APIs provided by Twitter.

Using the corpus, a sentiment classifier, is constructed that is able to determine positive, negative and neutral sentiments.

II. BACKGROUND WORKS

Pak and Paroubek (2010) [1] proposed a model to classify the tweets as objective, positive and negative. They created a twitter corpus by collecting tweets using Twitter API and automatically annotating those tweets using emoticons. Using that corpus, theydeveloped a sentiment classifier based on the multinomial Naive Bayes method that uses features like Ngram and POS-tags. The training set they used was less efficient since it contains only tweets having emoticons.

Parikh and Movassate(2009) [2] implemented two models, a Naive Bayes bigram model and a Maximum Entropy model to classify tweets. They found that the Naive Bayes classifiers worked much better than the Maximum Entropy model.

Go and L.Huang (2009) [3] proposed a solution for sentiment analysis for twitter data by using distant supervision, in which their training data consisted of tweets with emoticons which served as noisy labels. They build models using Naive Bayes, MaxEnt and Support Vector Machines (SVM). Their feature space consisted of unigrams, bigrams and POS. They concluded that SVM outperformed other models and that unigram were more effective as features.

Barbosa et al.(2010) [4] designed a two phase automatic sentiment analysis method for classifying tweets. They classified tweets as objective or subjective and then in second phase, the subjective tweets were classified as positive or negative. The feature space used included retweets, hashtags, link, punctuation and exclamation marks in conjunction with features like prior polarity of words and POS.

Bifet and Frank(2010) [5] used Twitter streaming data provided by Firehouse API , which gave all messages from every user which are publicly available in real-time. They experimented multinomial naive Bayes, stochastic gradient descent, and the Hoeffding tree. They arrived at a conclusion that SGD-based model, when used with an appropriate learning rate was the better than the rest used.

Agarwal et al. (2011) [6] developed a 3-way model for classifying sentiment into positive, negative and neutral classes. They experimented with models such as: unigram model, a feature based model and a tree kernel based model. For tree kernel based model they represented tweets as a tree.The feature based model uses 100 features and the unigram model uses over 10,000 features. They arrived on a conclusion that features which combine prior polarity of words with their parts-of-speech(pos) tags are most important and plays a major rolein the classification task. The tree kernel based model outperformed the other two models.

Davidov et al.,(2010) [7] proposed a approach to utilize Twitter user-defined hastags in tweets as a classification of sentiment type using punctuation, single words, n-grams and patterns as different feature types, which are then combined into a single feature vector for sentiment classification. They made use of K-Nearest Neighbor strategy to assign sentiment labels by constructing a feature vector for each example in the training and test set.

Po-Wei Liang et.al.(2014) [8] used Twitter API to collect twitter data. Their training data falls in three different categories (camera, movie , mobile). The data is labeled as positive, negative and non-opinions. Tweets containing opinions were filtered. Unigram Naive Bayes model was implemented and the Naive Bayes simplifying independence assumption was employed. They also eliminated useless features by using the Mutual Information and Chi square feature extraction method. Finally , the orientation of an tweet is predicted. i.e. positive or negative.

Pablo et. al. [9] presented variations of Naive Bayes classifiers for detecting polarity of English tweets. Two different variants of Naive Bayes classifiers were built namely Baseline (trained to classify tweets as positive, negative and neutral), and Binary (makes use of a polarity lexicon and classifies as positive and negative. Neutral tweets neglected). The features considered by classifiers were Lemmas (nouns, verbs, adjectives and adverbs), Polarity Lexicons, and Multiword from different sources and Valence Shifters.

In the first Map-Reduce pass, the mapper takes the labeled tweets from the training data and outputs category and word as key value pair. The Reducer then sums up all instances of the words for each category and outputs category and word-count pair as keyvalue. The Map-Reduce thus deals with formation of model for the classifier. The next Map-Reduce pass does the classification by calculating conditional probability of each word (i.e. feature) and outputs category and conditional probability of each word as keyvalue pair. Then final reducer calculates the final probability of each category to which the tweet may belong to and outputs the predicted category and its probability value as key-value pair.

III. SUGGESTED APPROACH

Collect unstructured data from Social Media sources.
2. Real-Time Processing with a sentiment analysis engine based on keyword search.
3. Store processed data (with sentiment) in NoSQL database.
4. Extract sentiments from NoSQL to visualization layer.
5. Visualize with a tool of choice.
The proposed system has the following modules;
1. Data streaming
2. Preprocessing
3. Sentiment polarity analysis
4. Visualization
5. Evaluation metrics

The details of the modules are presented below.
Data Streaming: Extracting real time tweets using Twitter Streaming API. For classification and training the classifier we need Twitter data. For this purpose we make use of API's twitter provides. Twitter provides two API's; Stream API1 and REST API2. The difference between Streaming API and REST APIs are: Streaming API supports long-lived connection and provides data in almost real -time. The REST APIs support short-lived connections and are rate-limited (one can download a certain amount of data [*150 tweets per hour] but not more per day).

Preprocessing: In this phase, the tweets are available as text data and each line contains a tweet. Initially we clean up or remove retweets as that will induce a bias in the classification process. We need to remove the punctuations and other symbols that doesn't make any sense as it may result in inefficiencies and may affect the accuracy of the overall process

Sentiment Polarity Analysis: MapReduce is a new parallel programming model, hence the classical Naive Bayes based sentiment analysis algorithm is adjusted to fit into Map Reduce model. we choose to employ a Naive Bayes classifier and empower it with an English lexical dictionary SentiWordNet

Visualization: Tweets are presented using several different visualization techniques. Each technique is designed to highlight different aspects of the tweets and their sentiment.

Heatmap: The heatmap visualizes the number of tweets within different sentiment regions. It highlights "hot" red regions with many tweets, and "cold" blue regions with only a few tweets.

Tag Cloud: The tag cloud visualizes the most frequently occurring terms in four emotional regions: upset in the upper-left, happy in the upper-right, relaxed in the lower-right, and unhappy in the lower-left. A term's size shows how often it occurs over all the tweets in the given emotional region. Larger terms occur more frequently.
Timeline: The timeline visualizes when tweets were posted. Pleasant tweets are shown in green above the horizontal axis, and unpleasant tweets in blue below the axis.

Map: The map shows where tweets were posted. Twitter uses an "opt-in" system for reporting location: users must explicitly choose to allow their location to be posted before their tweets are geotagged.

Affinity: The affinity graph visualizes frequent tweets, people, hashtags, and URLs, together with relationships or affinities between these elements.

Evaluation Metrics: We will evaluate our experiment results by using following Information Retrieval matrices .

- Precision = TP/(TP + FP)
- Recall = TP/(TP + FN)

- F-measure = 2*Precision*recall/(Precision + recall)
- Accuracy = TP + TN /(TP + TN + FP + FN)
-

IV. CONCLUSION

It is proposed to stream real time live tweets from twitter using Twitter API, and the large volume of data makes the application suitable for Big Data Analytics. A method to predict or deduct the location of a tweet based on the tweet's information and the user's information should be found in the future.

REFERENCES

[1] A.Pak and P. Paroubek. „Twitter as a Corpus for Sentiment Analysis and Opinion Mining". In Proceedings of the Seventh Conference on International Language Resources and Evaluation, 2010, pp.1320-1326

[2] R. Parikh and M. Movassate, "Sentiment Analysis of User- GeneratedTwitter Updates using Various Classification Techniques",CS224N Final Report, 2009

[3] Go, R. Bhayani, L.Huang. "Twitter Sentiment ClassificationUsing Distant Supervision". Stanford University, Technical Paper,2009

[4] L. Barbosa, J. Feng. "Robust Sentiment Detection on Twitterfrom Biased and Noisy Data". COLING 2010: Poster Volume,pp. 36-44.

[5] Bifet and E. Frank, "Sentiment Knowledge Discovery inTwitter Streaming Data", In Proceedings of the 13th InternationalConference on Discovery Science, Berlin, Germany: Springer,2010, pp. 1-15.

[6] Agarwal, B. Xie, I. Vovsha, O. Rambow, R. Passonneau, "Sentiment Analysis of Twitter Data", In Proceedings of the ACL 2011Workshop on Languages in Social Media,2011 , pp. 30-38.

[7] Dmitry Davidov, Ari Rappoport." Enhanced Sentiment Learning Using Twitter Hashtags and Smileys". Coling 2010: Poster Volumepages 241-249, Beijing, August 2010

[8] Po-Wei Liang, Bi-Ru Dai, "Opinion Mining on Social MediaData", IEEE 14th International Conference on Mobile Data Management,Milan, Italy, June 3 - 6, 2013, pp 91-96, ISBN: 978-1-494673-6068-5, http://doi.ieee computer society.org/10.1109/MDM.2013.

[9] Pablo Gamallo, Marcos Garcia, "Citius: A Naive-Bayes Strategyfor Sentiment Analysis on English Tweets", 8th InternationalWorkshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland,Aug 23-24 2014, pp 171-175.

[10] Neethu M,S and Rajashree R," Sentiment Analysis in Twitter using Machine Learning Techniques" 4th ICCCNT 2013,at Tiruchengode, India. IEEE – 31661.

[11] Balakrishnan Gokulakrishnan, Pavalanathan Priyanthan, Thiruchittampalam Ragavan, Nadarajah Prasath, AShehan Perera," Opinion Mining and Sentiment Analysis on a Twitter Data Stream", 2012, IEEE, ICTer : 182-188.

[12] P. Grandin and J. M. Adán," Piegas: A System for Sentiment Analysis of Tweets in Portuguese", 2016, ieee latin america transactions, vol. 14, no. 7

[13] Alexander Porshnev, Ilya Redkin, Alexey Shevchenko," Machine learning in prediction of stock market indicators based on historical data and data from Twitter sentiment analysis," 2013, IEEE, 879234-645-345

[14] Changbo Wang, Zhao Xiao, Yuhua Liu, Yanru Xu, Aoying Zhou, and Kang Zhang," SentiView: Sentiment Analysis and Visualization for Internet Popular Topics", 2013, ieee transactions on human-machine systems, vol. 43, no. 6

[15] Rincy Jose, Varghese S Chooralil," Prediction of Election Result by Enhanced Sentiment Analysis on Twitter Data using Word Sense Disambiguation", 2015, IEEE, 978-1-4673-7349-4.