Opinion Mining Using Machine Learning Algorithms

Tejas Zarekar[1] Kunal Shinde[2] Rohit Nair[3] Prof. Torana Kamble[4]
[1,2,3]B.E. Student [4]Professor
[1,2,3,4]Department of Computer Engineering
[1,2,3,4]Bharati Vidyapeeth College Of Engineering, Belpada, Navi Mumbai

Abstract— Feedback is one of the most important phase in a product development life cycle. But an honest opinion or a proper feedback isn't readily available or comes as a paid service. With the ever-growing online trend and social media use, millions of users share their views, opinions, ideas and reviews about almost everything they encounter in their daily life. The required information can be obtained from social media APIs and news APIs, thus enabling us to create a generalized opinion of the masses. Mining these tweets, comments, posts and other available data gives us a summarized view of the whole trend and how the users react to the products available in the market. This in turn can help an organization to get an idea of what the users expect from their products, what their shortcomings are and how they can improvise to improve their products.

Key words: Opinion Mining, Distributed Database, Machine Learning, HDFS, HBase, Hadoop

I. INTRODUCTION

Opinion mining is a growing field to identify the thoughts and sentiments of people, by making sense or inferring the meaning out of data that is available on the internet. In other words, it is an area of artificial intelligence which deals with evaluating the emotion or sentiment behind written text i.e. natural language processing. Today due to vast use of internet and social platforms, people are having a huge audience to publicly express their opinion. Social media provides them with a platform to voice their view about a product, political events, food from a restaurant and other product or service specific information. Users discuss and voice their opinion most of the time as it gives them the freedom to express their views without any hindrance.

Feedback is one phase which makes a lot of difference if implemented correctly. In every product life cycle, feedback is a vital phase. After launching a product, users may not react the way they are expected to or the current trend may not be what was expected while designing the product or scheme which leads to a gap between user expectation and what the current product provides them with. These minor issues that impact the success of the product in the market.

In our proposed system, we mine the data pertaining to the keyword provided by the user, sort them based on their polarity i.e. positive, negative or neutral and then create a generalized view for the same by using graphical representation. Also, the system provides individual posts that can be viewed by the user to further analyze the data and interpret the meaning out of the same.

II. REVIEW OF LITERATURE

Literature review provides us with the research papers and implemented system of a similar nature from the past few years. Our area of interest is data mining and automated feedback systems. The following are the papers published previously on these topics:

A. Harish Balaji, V.Govindasamy and V.Akila:

In the following paper [4] consists of an elaborate explanation of feedback measuring on different products using social media APIs and Python NLTK. Their system also summarizes the given input.

B. Aliza Sarlan, Chayanit Nadam and Shuib Basri:

In this paper [3], there is a detailed documentation of social media specific data mining. The system mines data using twitter API and classifies customer opinion via tweets into positive or negative review.

C. Santanu Mandal and Sumit Gupta:

In the following paper [7], the authors have mentioned the use of a lexicon based algorithms that help the machine segregate the mined data into positive, comparative and superlative degree words. This produced an F-score of 0.875 on test dataset.

D. V. K. Singh, P. Kumari, A. Singh and J. Thapa:

In this paper [8], the authors mention the use of opinion mining in order to rate a particular course in a department of a college. The collective review of multiple students is processed in order to generate the overall effectiveness of the course is decided. This helps in reviewing the course and how it can be improved to suit the students' need.

III. PROPOSED WORK

It is a known fact that a huge corpus is always gets an upper hand on enhancing the classification of the sentiment of the opinion mined. With the implementation of distributed file system, the problem of running out of space with ever growing background data can be solved and thereby creation of a self-improving system.

This is to propose the implementation of HDFS on Hadoop cluster which would have Hbase (Distributed Database) to store the mined opinions and the sentiment predicted by the algorithm. The opinions would be mined from the chosen sources selected by the user on the web interface. The sources that are social networking websites, blogging websites and technical forums, they would be mined for a particular keyword or phrase supplied by the user via online API's provided by the sources. HDFS will be holding the background corpus of known sentiment and also the newly mined data for the user entered corpus that would be rechecked manually by a volunteer for the accuracy of the prediction stored in the corpus. This practice directly includes a sentient in the process of sentiment prediction thereby justifying the adjective "self-improving" given to the proposed system.

Machine learning algorithms would be used for prediction. The algorithms include Naive Bayes, K-means,

ANN that would run on the freshly mined data and depend on the background corpus stored on HDFS for learning the sentiment. Each result would then be stored for future reference and use.

User is made presented with a report of the keyword fed to the tool that includes graphical visualizations like bar plots, aesthetically pleasing word-cloud and histograms of the sentiment words appeared in the mined texts that make the report presentable.

A. Methodology:

1) The user gives the input for which the data has to be gathered.
2) Gather the content related to the user given input and build the desired result using the appropriate APIs.
3) Clean the content by removing reposts, hashtags, links and words that don't contribute to the sentiment of the sentence.
4) Analyze the sentence for sentiment by rating the sentiment of each word.
5) After deciding the sentiment, it needs to be categorized by referring to the existing data set.
6) Visualize the result by plotting graphs and pie charts.

B. Architechture of the proposed system:

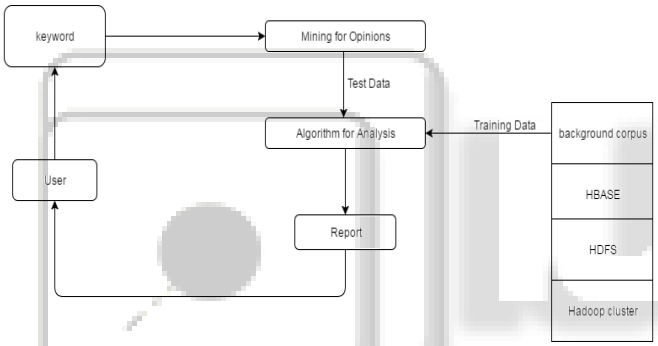

Fig. 1: Proposed System

C. Advantages of Proposed System:

1) User interface is easy to use.
2) The system is deployed on a web application which makes it easier for cross platform compatibility.
3) The system can be used on any platform as long as it has internet connection and a web browser.
4) The users need not be trained in any manner to use the system.
5) It uses data from multiple web sites so that it can maintain a large data set and hence the result is more accurate.
6) Manual intervention can be made to change erroneous predictions, which the system learns for future reference.

IV. ALGORITHMS USED

1) Classification Based Algorithm: Data mining uses association rules to analyze relationship between data that is included in the huge corpus of data or training set of data. Classification is mapping the instances of data into multiple distinct classes. Classification and association rules in mining create an integrated framework known as Associative Classification which is widely known as Class Association Rule. Class Association Rule is a predictive data mining task. On can build classifiers for large databases using Class Association Rule.
2) Naive Bayes Classification Algorithm: In machine learning interpreting the given sentiment of the words is the key to concluding the sentiment of the given text. Naive Bayes Classification Algorithm uses a training data set in machine learning. It loads a corpus of trained data whose polarities are already defined. This process is when the machine learns the meaning of each word and later uses this learnt knowledge to interpret the polarity of mined data.
3) Artificial Neural Networks: Neural networks are better understood when they are looked at as multiple nodes acting like neurons in a human brain working together in order to achieve a common goal. Each neuron can communicate with other neurons in the system. A general threshold value can be set to limit the number of connections each neuron makes before it can propagate to other neurons. This provides the system with flexibility it needs because it is self-leaning in nature. The main goal is to solve a problem like any human brain would, it is a distributed system working together in unison to process the given input. These types of systems are scalable as additional systems can be added later on if need be.

V. ERROR ANALYSIS

There are times when the system may not perform as expected and may label a statement incorrectly. This can happen due to underlying sense of sarcasm or humor which the machine cannot understand. These can be mitigated by human intervention and this would help the machine to learn and thereby reduce error propagation. Also, as the size of the training set grows, such errors can be minimized as a larger training corpus is always better for correct prediction.

VI. CONCLUSION

Conducting elaborate surveys for a given product or service needs extensive resources and manual work. These are documented and later on take a considerable amount of time to interpret the view of the customers of the product or service. These results may not be accurate as not all users would be fully aware about the product. These drawbacks are overcome by our proposed system as it takes into consideration a huge amount of data and is available free of cost. The system provides the user with a generalized report and gives the user a view of the current trend. The data is freshly mined using the APIs provided by the websites and hence is highly reliable.

ACKNOWLEDGEMENT

This research on "Application to improve Blood donation process using Data Mining techniques" was successfully represented with the help of our project guide Prof. Torana Kamble. We would like to thank her for providing us appropriate guidance whenever necessary.

We express our gratitude to our Head of Department Dr. D. R. Ingle for providing timely assistant to our query and guidance that he gave owing to his experience in the field for past many years.

We express our gratitude to our respected Principal Dr. M. Z. Shaikh for his coordination and obliging us with his great knowledge.

REFERENCES

[1] ZHU Nanli, ZOU Ping, LI Weiguo, CHENG Meng, "Sentiment analysis: A Literature Review"
[2] H. Binali, Chen Wu and V. Potdar, "A new significant area: Emotion detection in E-learning using opinion mining techniques", 2009 3rd IEEE International Conference on Digital Ecosystems and Technologies, 2009
[3] Aliza Sarlan, Chayanit Nadam and Shuib Basri, "Twitter sentiment analysis", 2014 IEEE International Conference on Information Technology and Multimedia
[4] Harish Balaji, V.Govindssamy and V.Akila, "Social opinion mining and consise rendition", 2016 IEEE international conference on Advanced Communication Control and Computing Technologies
[5] Apoorv Agarwal, Vivek Sharma, Getta Sikka, Renu Dhir, "Opinion Mining of news headlines using SentiWordNet", 2016 IEEE
[6] Sentiment analysis: A literature review by Zhu Nanli, Zou Ping, Li Weiguo and Cheng Meng, Published in Management of Technology (ISMOT), 2012 International Symposium, INSPEC Accession Number: 13952551
[7] Santanu Mandal and Sumit Gupta, "A novel dictionary-based classification algorithm for opinion mining", 2016 Second International Conference on Research in Computational Intelligence and Communication Networks
[8] V. K. Singh, P. Kumari, A. Singh and J. Thapa, "An automated course feedback system using opinion mining", 2011 World Congress on Information and Communication Technologies

