https://dx.doi.org/10.24001/ijaems.3.3.26 *ISSN: 2454-1311*

Opinion Poll: Big Data Implementation of Unstructured Data Analytics of Social Network Reviews Using Sentiment Analysis & SVM

S. Sylvia Irish[1], M. Sicily Sherin[2], R. Surya[3], Y. Vidhya[4], Vidya Ramamoorthy[5]

[1]Assistant Professor, Panimalar Engineering college, Tamil Nadu, India
[2,3,4,5]Student, Panimalar Engineering college, Tamil Nadu, India

Abstract— *Recent systems developed are dependent on user feedbacks or opinions. These feedbacks or opinions are generated in volumes everyday which are difficult to filter and analyse. We propose Sentiment based analysis is the major key in categorizing the user's Feedback. In thispaper, we study the processing of all the reviews posted in an online shopping application and classify them using SVM. We use big data to analyze the vast amounts of data generated. User reviews are the input to the Big Data HDFS System. Data are stored in the Data Nodes. Index is maintained in the Name Node. Reviews are analyzed using Sentiment Analysis and Positive & Negative Tweets are classified. Also products are recommended based on the previous purchases and group notification is sent to all the customers in a group.*

Keywords—Sentiment analysis, SVM, Big data.

I. INTRODUCTION

The opinions, feedback or reviews given in various online sources like social media, online discussion groups, news, blogs, online reviews or any other large collection of text are an important yardstick for the success of a product or a government policy. For instance, a product with consistently good reviews is likely to sell well. Sentiment analysis is basically done to classify the document or a sentence using subjective methods i.e based on the context of the data or feature based methods which analyzes the sentiment expressed regarding the features or aspects of the entities.

The orientation of a sentence or a document is determined by analyzing the orientations of the individual words. Sentiment dictionaries are used to summarize the documents.There are various systems that given a sentiment lexicon, analyze the structure of a sentence/document to infer its orientation, the holder of an opinion, the sentiment of the opinion, etc.Many dictionaries have been created manually or (semi)-automatically, e.g., general inquirer (GI), opinion finder (OF), appraisal lexicon (AL), Senti WordNet (SWN) and Q-WordNet (QW). The synsets (senses) in WordNet areclassified according to their polarities by the lexical resources QW and SWN. We call them sentiment sense dictionaries (SSD). OF, GI and AL are called sentiment word dictionaries (SWD).

We notice that these sentiment dictionaries have numerous inaccuracies. So the opinion results could not be principally categorized. Then FSM & EEM Algorithm wereused for the Word processing process.Sentiment analysis tasks without feature engineering use sentiment embedding as word features. Previously word-level sentiment analysis, sentence level sentiment classification, and building sentiment lexicons used sentiment embeddings. These sentiment embeddings consistently outperform context-based embeddings on several benchmark datasets of these tasks. This work provides information on the design of neural networks for learningspecific word embeddings in other natural language processing tasks.

In this paper, learning sentiment-specificword embeddings dubbed sentiment embeddings for sentimentanalysis is proposed. We retain the effectiveness of word contextsand exploit sentiment analysis of texts for determining more powerfulcontinuous word representations. By capturing both context and sentiment level evidences, the nearest neighbours in theembedding space are not only semantically similar but alsofavour to have the same sentiment polarity, so that it is ableto separate good and bad to opposite ends of the spectrum.In order to learn sentiment embeddings effectively, wedevelop a number of neural networks to capture sentimentof texts (e.g. sentences and words) as well as contexts ofwords with dedicated loss functions. We learn sentimentembeddings from tweets, using positive and negative emoticons as pseudo sentiment labels of sentences withoutmanual notes.

https://dx.doi.org/10.24001/ijaems.3.3.26 *ISSN: 2454-1311*

II. UNSTRUCTURED DATA ANALYTICS

It is quite difficult to obtain a general opinion or know about the orientation of people towards a particular subject or product. if the opinion is polled manually as it can only be done for a small group of people in a particular area. People in large areas with different environments cannot be covered. Also if done online it is difficult to judge the average opinion efficiently due to the vast amount of data present in the social media or other websites. The current system fails to identify the sentiment of the texts correctly as they only consider the context of the texts. So we propose a system which will correctly identify the sentiment of the texts and also classify them as positive or negative. We apply sentiment embeddings for sentiment classificationof reviews to investigate its ability in discovering discriminativefeatures from different domains. We run supervisedlearning pipeline regarding word embeddings asfeatures.

A sentiment-specific word embeddings dubbed sentiment embeddings can be used instead. Existing wordembedding learning algorithms typically only consider the contexts and ignore the sentiment of texts. It is problematic for sentiment analysis because the words with similar contexts but opposite sentiment polarity, such as good and bad, are mapped to neighbouring word vectors. We address this issue by encoding sentiment information of texts (e.g., sentences and words) together with contexts ofwords in sentiment embeddings. By combining context and sentiment level evidences, the nearest neighbours in sentiment embedding space are semantically similar and it favours words with the same sentiment polarity. A number of neural networks with tailoring loss functions is developed, and texts are collected automatically with sentiment signals like emoticons as the training data to learn sentiment embeddings efficiently.

We propose a system where all the user reviews are gathered and categorized according to our needs using map reduce method as given in fig. 1. They are then further analysed and categorized into positive or negative feed using sentiment analysis and algorithms like SVM. The reviews are posted only after the user are authenticated through transaction ID and OTP. When a customer buys a product a transaction ID is generated and sent to their mail id. When the ID is submitted in the application further an OTP is generated which is also received by the customer though gmail. Once the customer is verified after entering the OTP only then can they post reviews. These reviews are then categorized as positive and negative and is displayed with the product details. For each and every review the count of either the positive or negative feed is increased and also updated in the review details. Further the users are grouped using hadoop and the products are recommended to the users based on the previous purchased made by all the customers. Also whenever a new purchase is made by a customer in a group, all the members of the group are notified through a group mail. We use a feature based selection process to obtain a better idea of the sentiments of the users. The appropriate algorithms like SVM and sentiment analysis is applied on them which determine its sentiment polarity. These opinions/tweets are then clustered using reduce technique into positive and negative comments and also the number of tweets in both the categories are displayed as output which helps us to determine the overall opinion of people.

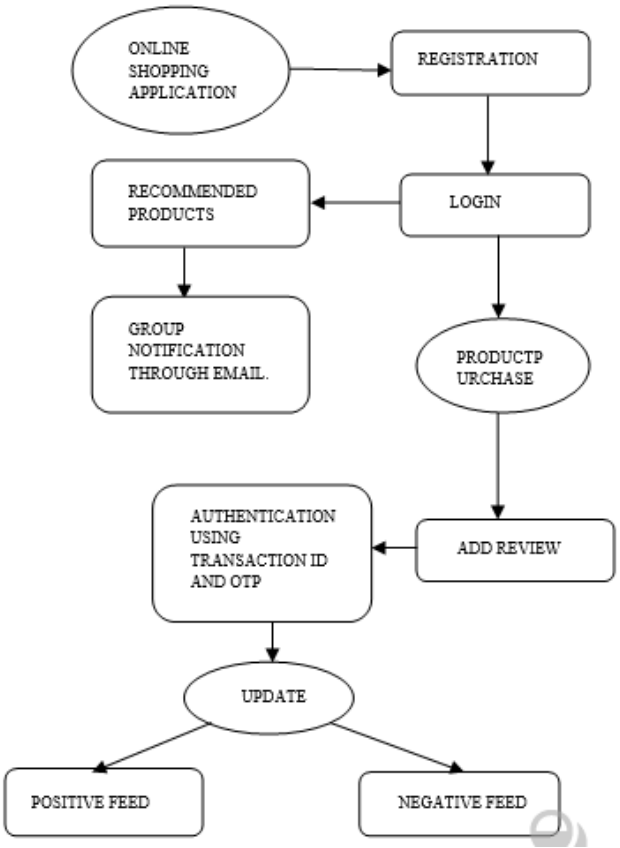

Fig.1: Opinion poll workflow

III. METHODOLOGY

3.1 User Registration

As detailed earlier this paper is on categorizing reviews or opinions in various social websites like twitter, various blogs, shopping sites like amazon, flipkart etc. For the purpose of experiment and demonstration we have created an online shopping application using Advanced Java Concepts like JSP and Servlets. This application will enable the user to login and record their comments similar to the way other websites works. The information entered in this application is secured with a user ID and password. Once logged in the user will be able to purchase products and post review for those products. These details are

https://dx.doi.org/10.24001/ijaems.3.3.26 *ISSN: 2454-1311*

captured in the back end MySQL DB and later on retrieved to serve as input to the sentiment analysis module.

3.2 Product purchase

The Server will monitor the entire User's information in their database and verify them if required. Also the Server will store the entire User's information in their database. Also the Server has to establish the connection to communicate with the users. The Server will update the each user's activities in its database. The Server will authenticate each user before they access the application. So that the Server will prevent the unauthorized user from accessing the application. The server also stores all the information about the products displayed, products purchased and reviews posted which can be retrieved easily when required. Once the customer is logged in the required product can be searched for and purchased. Upon which a transaction ID is generated and sent to the user's mail.

3.3 Customer feedback

After the product is purchased,the customer can add reviews to the products but only after verification and authentication which is done using the transaction ID previously generated and an OTP which are received by the customers through email. After the customer is verified, any review can be posted successfully which is then updated in the product details as an increased count in the positive or negative feed. The reviews can also be viewed if necessary.

3.4 OTP generation and verification

A one-time password (OTP) is a password that is valid for only one login session or transaction. OTPs avoid a number of shortcomings that are associated with traditional (static) passwords. The most important shortcoming that is addressed by OTPs is that, in contrast to static passwords, they are not vulnerable to replay attacks. This means that a potential intruder who manages to record an OTP that was already used to log into a service or to conduct a transaction will not be able to abuse it, since it will be no longer valid. On the downside, OTPs are difficult for human beings to memorize. Therefore they require additional technology to work. And for verification the code sent as email after that only feedback is accepted. For a customer to post a review, he must be validated which is why we are using this method. An OTP is generated whenever an user wants to post a review which is sent to the customer's mail. The user is allowed to post a review only when d OTP entered by the customer matches that which has been sent to him.

3.5 SVM and sentimental analysis

We introduce the review/comment analysis using SVM to categorize based on their nature and sensitivity and assess the sentiment to segregate positive and negative feedback. This will enable us to sense the public opinion on any product in the e-commerce application. This will also help in rating various products and forums. With minor modifications this can be used to evaluate user feedback/opinion posted in various otherwebsites.

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labelled training data (*supervised learning*), the algorithm outputs an optimal hyperplane which categorizes new examples. SVM is used for classification analysis, it determines in which category does a data fall in by depicting the examples in the training set as different categories using points in space separated by a gap. It maps the new data and predicts the category based on which side the fall on.

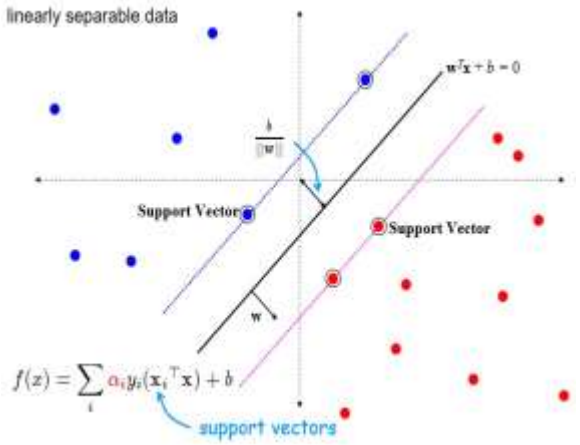

Fig.2: Support Vector Machine

Compute the convex hull of the positive points, and theconvex hull of the negative points

For each pair of points, one on positive hull and the otheron the negative hull, compute the margin

Choose the largest margin.

A Linear classifier has the form

$f(x) = w^T x+b.$

For a linear classifier, the training data is used to learn $\mathbf{w}$ *and then discarded, only* $\mathbf{w}$ *is needed for classifying new data.*

Since $w^T x + b = 0$ *and* $c(w^T x + b) = 0$ *define the same plane, we have the freedom to choose the normalization of w.*

Choose normalization such that $w^T x_+ + b = +1$ *and* $w^T x_- + b = -1$ *for the positive and negative support vectors respectively.*

Then the margin is given by

$w/||w||.(x_+ - x_-) = w^T (x_+ - x_-)/||w|| = 2/||w||$

https://dx.doi.org/10.24001/ijaems.3.3.26 *ISSN: 2454-1311*

In this algorithm number of labelled tweets/comments/data is given to the classifier as the training data which is then used to learn the line or hyperplane which separates the data into different categories i.e. as positive or negative. These training data are then discarded and only the learning is stored which is used for the classification of new data. Further all the tweets given as inputs are classified based on this learning w which is essentially the hyperplane which separates the positive and negative data.

3.6 Product recommendation

After various purchases are made and reviews are posted, products are recommended to the users based on these previous purchases. The hadoop system groups the users according to the products purchased by them. It groups all the users who have bought same items together. This is done so that the products may be recommended to the customers. For each and every customer different product is recommended based on the products they have already purchased and the group they belong to. Once these groups are established, the members of the group are also notified if any new purchase is made by any of the group members through email. We can also view the products recommended for us in the online shopping application also.

IV. EXPERIMENTAL RESULTS

A large data set with various reviews was given as training data to the classifier. It was trained with different data sets containing labeled positive and negative opinions/comments/reviews. Various kinds of training data were given to the classifier. We observed different results for each training set. The classifier worked best if the training set had equal positive and negative tweets. It proved to be more accurate than the existing system.

The classifier correctly identified the positive and negative reviews but was not as efficient for mixed or neutral tweets. Also it worked best when used in the same domain as it was trained, but when the same classifier was used in other domains its efficiency dropped considerably. It also correctly recommended the products for each customer based on the previous purchases made and also the group they are mapped into. We also received group notification each and every time a product was purchased by a fellow group member.

V. CONCLUSION

We study the problem of determining the sentiment orientation of the users by performing sentiment analysis and checking the polarity as positive and negative on their posted data to determine the overall opinion of people towards a particular product. This drastically reduced the time consumed for review analysis and provided more accurate and reliable results. Also the products were efficiently recommended to the customers. It is also noted that the classifier trained in a particular domain does not work as efficiently in other domains.

For future work, we plan to work towards various directions. We plan on implementing three way classification i.e. positive, negative and neutral. As including a category as neutral has shown to improve the accuracy of the analysis. This can be done either by determining the neutral sentiment first and then later categorize the positive and negative sentiments or by performing a three way classification altogether. We also plan to train a single classifier in various domains which would increase the efficiency by multiple times.

REFERENCE

[1] J. Wang and Y. Zhang, “Opportunity model for E-commerce recommendation: Right product; right time,” in Proc. 36th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2013, pp. 303–312.

[2] M. Giering, “Retail sales prediction and item recommendations using customer demographics at store level,” SIGKDD Explor. Newsl., vol. 10, no. 2, pp. 84-89, Dec. 2008.

[3] G. Linden, B. Smith, and J. York, “Amazon.com recommendations: Item-to-item collaborative filtering,” IEEE Internet Comput., vol. 7, no. 1, pp. 76–80, Jan./Feb. 2003.

[4] V. A. Zeithaml, “The new demographics and market fragmentation,” J. Marketing, vol. 49, pp. 64–75, 1985.

[5] W. X. Zhao, Y. Guo, Y. He, H. Jiang, Y. Wu, and X. Li, “We know what you want to buy: A demographic-based system for product recommendation on microblogs,” in Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2014, pp. 1935–1944.

[6] J. Wang, W. X. Zhao, Y. He, and X. Li, “Leveraging product adopter information from online reviews for product recommendation,” in Proc. 9th Int. AAAI Conf. Web Social Media, 2015, pp. 464–472.

[7] Y. Seroussi, F. Bohnert, and I. Zukerman, “Personalised rating prediction for new users using latent factor models,” in Proc. 22nd ACM Conf. Hypertext Hypermedia, 2011, pp. 47–56.

[8] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in Proc. Adv. Neural Inf. Process. Syst., 2013, pp. 3111–3119.