International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

A Review on Automatic News Classification using the Probabilistic Classification Algorithms

Mandeep Kaur[1], Pravneet Kaur[2]

[1]CGCTC, Jhanjeri, India

[2]Assitant Professor of Department CSE, CGCTC, Jhanjeri, India

Abstract: ***Reviews are unbiased information obtained from the sources outside an organization, which makes them more reliable in the eyes of customers. Online shoppers are very much concerned about product reviews before making any decision regarding buying the product. Product reviews plays an important role in determining what kind of product is. Such reviews provide useful information about customer concern and their experience with the product. Consequently, these reviews will be helpful for a business making products for the purpose of product recommendation, better customer understanding and attracting more loyal customers. As ecommerce has become so popular, numbers of reviews are increasing day by day. It is difficult for a customer to read all the reviews manually. In this paper, an approach is developed which is used to obtain the summary from thousands or hundreds of online reviews. This approach uses extraction summarization for summarizing the reviews thereby selecting the original sentences and putting it together into a new shorter text explaining the overall opinion about the product. Although previous studies of deriving useful information from customer reviews focus on categorical or numerical data and textual data has been ignored. But textual data are of equal importance so it should not be ignored. So, this approach includes every aspect of the review in the summary so that a customer would be able to make a right decision regarding product.***

Keywords: Sentiment analysis, product classification, unique intensity words, term frequency, polarity evaluation

1. Introduction

Product reviews play a vital role in a selection of a particular product. Customer reviews about a product are considered as sales drivers and are something that majority of the customers will want to know before making a decision to buy a product. It is a fact that online customer reviews are trusted nearly 12 times more than the description provided by the manufacturers. Inventors of ecommerce like Amazon and eBay have been using product review since 1997, they lead people to write their opinion and share their experience about the products they have used.

Gathering reviews from customers act as an asset for an organisation selling products as this will help manufacturers to aware about strength and weakness of their product and help them to improve it. When going to buy product online, customers usually look at ratings of the product, read out reviews given by other customers and then compare the product with other products of same category. Quite simply, customer reviews increase conversions. Customer reviews help in improving online business. Organizations look out the reviews given by customers to know what improvements they can make in their product. Different organizations use several different ways of getting product reviews from customers. For an instance, Amazon has a excellent model for collecting product reviews through email. Social media like Facebook, Twitter and many other are considered as reliable sources of getting reviews. Using customer service or suggestion cards, customers are suggested to leave their thoughts and opinions about products. But it is difficult for a customer to go through hundreds or thousands of reviews to make a decision whether to buy a product or not. In response, this paper has proposed a technique for the summarization of customer reviews.

There are various reasons that show the importance of customer reviews for an organization selling products online:

1) Whenever a company introduces a new product then customer feedback is very important for deterring customer needs and tastes.
2) Companies can better understand that how their products are better than other products by analysing the customer ratings of product and their reasons for selection.
3) Companies can determine whether their customers are getting satisfactory level of service by their employees.
4) Customer reviews help in deterring why consumers are no longer interested in buying products from them, if any. This will help in building up strategies that would help lose customers back into business.
5) Customer reviews are also important in determining technological trends in the market.

2. Related Work

Prollochs, Nicolas et. al. [2] has worked on the improvement of sentiment analysis of monetary news by detective work negation scopes. To predict the corresponding negation scope, connected literature ordinarily utilizes 2 approaches, namely, rule-based algorithms and machine learning. however, an intensive comparison is missing, particularly for the sentiment analysis of monetary news. to shut this gap, this paper uses German impromptu announcements as a standard example of monetary news so as to pursue a two-sided analysis. First, we have a tendency to compare the prognosticative performance employing a manually-labeled dataset. Second, we have a tendency to examine however detective work negation scopes will improve the accuracy of sentiment analysis. Cui, Limeng et. al. [3] has developed a hierarchy technique supported lda and svm for news classification. during this paper the authors have targeted on

news text classification, that is meaningful for data supplier to prepare and show the news however conjointly for the users to succeed in the dear data simply. A hierarchy technique supported LDA and SVM is projected to accomplish this task and several other experiments square measure conducted to judge the projected technique. The results show that the projected technique is promising in text classification issues. Ouyang, Yuanxin et. al. [4] has projected the news title classification with support from auxiliary long texts. during this paper, the authors have targetted on the matter of stories title classification that is an important associated typical member in brief text family and propose an approach that employs external data from long text to deal with the matter the sparsity. later on Restricted Boltzman Machine square measure utilized to pick options and so finally perform classification mistreatment Support Vector Machine.

Ouyang, Yuanxin, Yao Huangfu, Hao Sheng, and Zhang Xiong [9] projected the News Title Classification with Support from Auxiliary Long Texts. The performance of text classification is proscribed as a result of its intrinsic shortness of sentences that causes the sparsity of vector house model. ancient classifiers like SVM square measure very sensitive to the options house, thereby creating classification performance dissatisfactory in brief text connected applications. it's believed that mistreatment external data to assist higher represent computer file might yield satisfying results. during this paper, the authors target on the matter of stories title classification that is an important associated typical member in brief text family and propose an approach that employs external data from long text to deal with the matter the sparsity. later on Restricted Boltzman Machine square measure utilized to pick options and so finally perform classification mistreatment Support Vector Machine. D. Morariu, R. Cre,Tulescu and L.,Vin,tan [12] says that build up on the meta- classifier bestowed supported eight SVM elements, we have a tendency to boost these a replacement mathematician sort classifier that results in a major improvement of the higher that the meta classifier will reach. Lie Lu, Stan Z. Li Associate in Nursingd Hong –Jiang Zhang [5] bestowed well our approach that uses SVM for classification and segmentation of an audio clip. The projected approach classifies audio clips into one in all 5 classes: Pure speech, Music, atmosphere sound and silence. We have conjointly projected a collection of recent options to represent a 1 second sub clip, as well as band regularity , LSP divergence form and spectrum flux. Krishnlal G, S adult male Rengarajan, K G Srinivasagan [6] The intelligent news classifier is developed and experimented with on-line news from internet for the class sports, finance and politics. The noval approach combining 2 powerful algorithms, Hidden mathematician Model and Support vector machine, within the on-line news classification domain provides extraordinarily smart result compared to existing methodologies. By the introduction of many preprocessing techniques and also the application of filters we tend to scale back the noise to an excellent extent, that successively improved the classification accuracy.

3. Experimental Design

Sentiment Analysis is the technique to analyze the emotion from a text document, message, or similar content. Sentiment Analysis algorithms are used to predict the public opinion on various issues getting discussed in the social network threads. Hence, it is also called opinion mining or emotion mining. In this paper, we have proposed the algorithm to calculate the sentiment in the social threads with positive or negative messages. At first, the messages are scanned for their positivity or negativity. Then, the messages are automatically classified according to their product feature and re-analyzed for further emotion. The algorithm uses the weighted dictionary matching to analyze the sentiment in the messages. The social thread is also recording the results on the basis of every user taking part in the social discussion thread. The results have shown the high accuracy of 97.3%. The accuracy is measured on the basis of type 1 and type 2 statistical errors.

We propose the use of a rich set of sentiment analysis features like positive, negative, automatic product feature classification and automatic summarization. The proposed feature selection method can improve opinion classification performance. The proposed Feature Relation Network is rule-based sentiment classification method that finds the text features and emotions from the given message data. The proposed algorithm consisted of four basic components: Post/Thread Acquisition, Tokenization, Polarization & Negative Emotion Analysis.

Post/Thread Acquisition: The first step is used to read the post saved in the excel file. The comments are read and classified as the user comments by grouping the comments of one users in one group.

Polarization (Positive or Negative rating module): The user comments are polarized in three major categories under this step. The three major categories are positive, negative and neutral. The tokenized comments are compared with a list of words. The file contains the ranking for each of the word listed on the list. The rank or weight or strength of the words has been listed in the document, which ranges between -5 to +5. The words are classified on the basis of their use and its impact in the natural English language spoken in our daily lives.

Negative Emotion Analysis: All of the user comments marked as negative the undergoes the negativity analysis, which checks the comments for the different negative emotions. The user comment is compared with two different files, out of which one is containing the words representing product feature classification and other for automatic summarization. The comment is marked on the basis of higher weight. For example if public review on the specific product is found high positive then the text summarization module shows the significance emphasis on the emotion analysis.

3.1 Algorithm 1: Brief Design of Sentiment analysis algorithm

1. Obtain the data from the social network thread Tr
2. Extract the list of users U from the social networking thread
3. Extract N number of words from Message M using dictionary based tokenization
4. Filter message content with STOPWORD list of common English words while Tokenization
5. Load product and feature classification (PFC) knowledge data
6. Classify the message after comparing it with the PFC data
7. Acquisition of the sentiment and expression classification (SEC) knowledge data
8. Calculate the message score after comparing it with SEC data
9. Classify the message according to the score and increment the product or product feature index accordingly
   a. If score is more than zero
      Increment the positive index
   b. If score equals zero
      Increment the neutral index
   c. If score is less than zero
      Increment the negative index
10. Load product review summarization (PRS) knowledge data
11. Prepare the summarization content according to the sentiment report

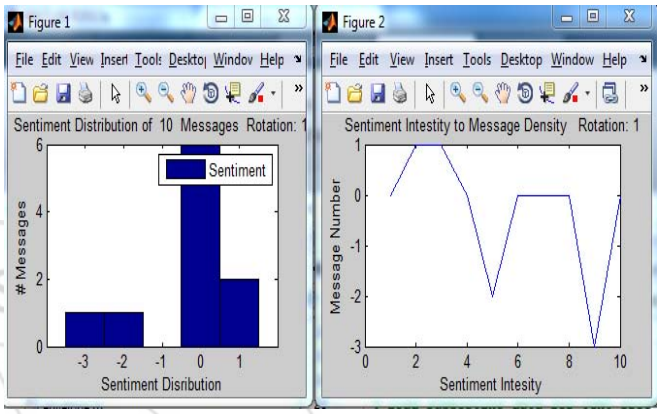
3.2 Algorithm 2

Detailed Explanation of Sentiment Analysis Model

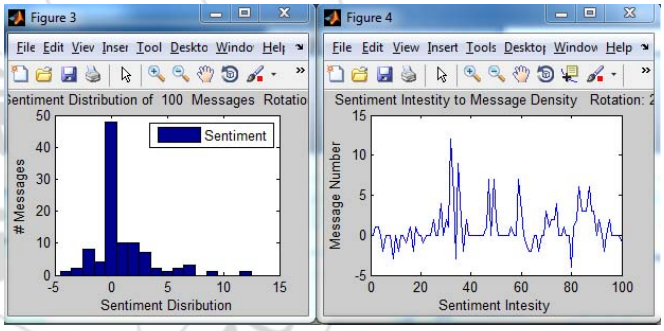
12. Obtain the data from the social network thread **Tr**
13. Extract the list of users **U** from the social networking thread
14. Extract **N** number of Message **M** using dictionary based tokenization
15. Filter message content with STOPWORD list of common English words while Tokenization
16. Load negative and positive sentiment expression word classification information file
17. Calculate word weight score to measure the sentiment **Sn**
18. Count the final sentiment score **S** of each message (Positive/Negative)
19. Calculate sentiment score for tokenized message number **N**
20. Find the sentiment type **St** (Positive/Negative) by validating the sentiment specific word dictionary
21. If **St >0**
    a. Mark the message as positive
    b. Add 1 to **posMsg**
22. If **St<=0**
    a. Mark the message as negative
    b. Add 1 to **negMsg**

4. Result Analysis

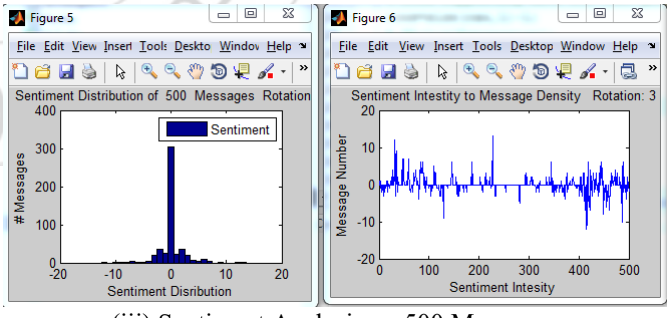
The results have been obtained from the proposed model. The proposed model have been given a discussion thread from the social network website collected in the excel sheet as the input data. The sentiment analysis has been performed on the excel file containing messages to find certain emotions automatically. The proposed algorithm returns the positive, negative, automatic product feature classification after analyzing the messages. The emotions are calculated by analyzing the words weight in the certain combinations & counting the whole emotion weight to calculate the resultant weight of the message. The messages are firstly broken in the words, phrases or combination of words, collectively called tokens, matching with the pre-programmed dictionary file stored up in the proposed model.



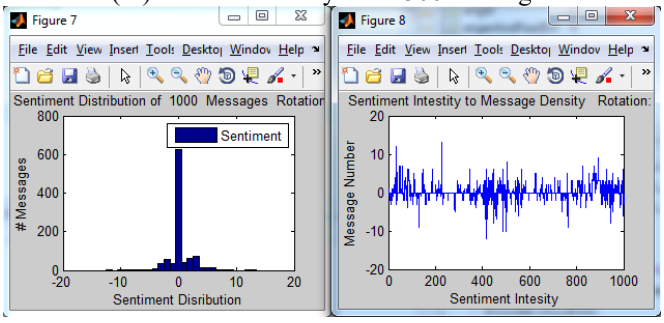
(i) Sentiment Analysis on 10 Messages



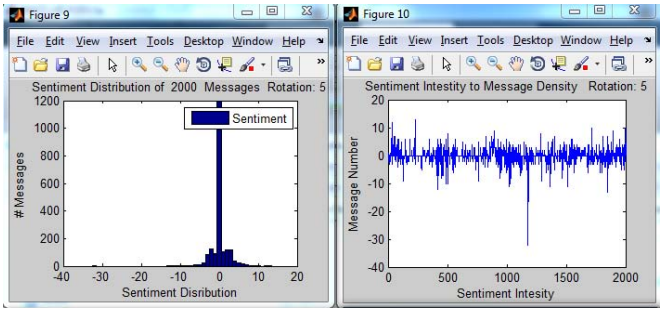
(ii) Sentiment Analysis on 100 Messages



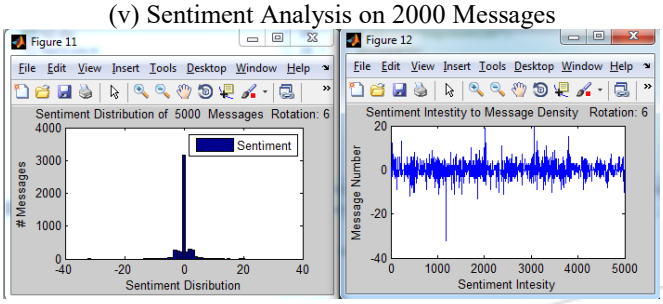
(iii) Sentiment Analysis on 500 Messages



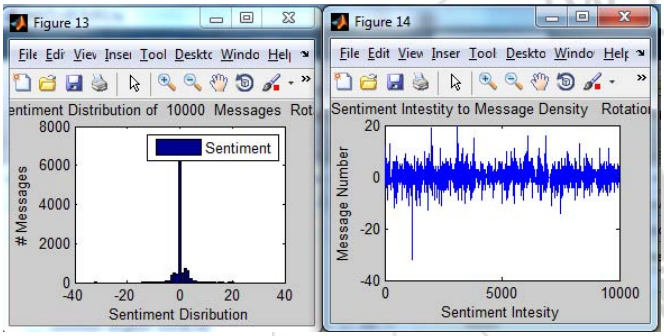
(iv) Sentiment Analysis on 1000 Messages



(v) Sentiment Analysis on 2000 Messages



(vi) Sentiment Analysis on 5000 Messages



(vii) Sentiment Analysis on 10000 Messages

Then the score or weight is calculated for each token, & a total of the token score is calculated and published. The published score is analyzed against the slab to mark the message as positive or negative. The dataset messages are analyzed against the pre-programmed dictionaries for positive and negative words and their scores. After the calculation of the emotion score, the message is marked as negative or positive and further classified according the product feature, where the message is again analyzed using the sentiment analysis module which gives the user review on that specific feature. The number of negative messages, positive messages has been given in the table 1. Table 1 is also equipped with the number unreadable messages from the other languages. Table 1 shows the normal sentiment analysis using the proposed scheme. The following table data has been designed to measure the accuracy of the proposed model in terms of analyzing the emotion from the given dataset. The accuracy has been tested against the manual classification. The manual classification is entirely based upon the natural emotion selection by the human.

Table 1: Analysis of various numbers of message

Total Number of Messages	Positive Messages	Negative Messages	Message From Other Languages
10	2	8	0
100	37	63	5
500	100	400	24
1000	219	781	24
2000	430	1570	27
5000	984	4016	47
10000	2190	8810	212

Table 2: A table of statistical error calculated manually on the given database

Total Number of Messages	True Positive	True Negative	False Positive	False Negative	Recall (TP/TP+FN)	Precision TP/ TP+FP
10	10	0	0	0	100%	100%
100	94	1	5	0	100%	94.90%
500	483	5	12	0	100%	97.57%
1000	978	8	14	0	100%	98.58%
2000	1943	24	29	2	99.80%	98.52%
5000	4897	43	56	3	99.91%	98.86%
10000	9790	89	113	8	99.18%	98.85%

The results have been obtained by applying the proposed algorithm for 10, 100, 500, 1000, 2000, 5000 and 10000 messages. The overall percentage of correct results by the proposed algorithm is 97.3. The algorithm has shown a higher accuracy in term of sentiment analysis.

In this thesis project, the proposed model has shown significant accuracy in calculating the sentiment in the social thread. The social thread collected from social network Facebook, Twitter or GSMArena has been analyzed with this proposed algorithm for analyzing its performance. The results in the table 1 are showing the different types of emotions calculated on the messages using the proposed algorithm. The accounted emotions the message been marked with are positive, negative, product feature classification and unreadable messages from other languages. The proposed algorithm has been tested with various sizes of messages data in each rotation as given in the table 1. The results of type 1 and type 2 statistical errors given in table 2 are manually validated. The results are proving the good performance of the proposed algorithm with accuracy of almost 97%.

In the future, the proposed algorithm can be enhanced to calculate product features of more products or with a wider range. The above emotions can be probably calculated using the dictionary based phrase specification methods. Also, the proposed algorithm can be improved on the basis of execution time and accuracy.

5. Conclusion

The product evaluation & summarization process consists of two major components: sentiment analysis & summarizer. The sentiment analysis component gives the sentiment spread in order to evaluate the user opinion on the product being evaluated using the input data. The proposed model has been designed in the different components for sentiment analysis and product review summarization. The proposed model

performance has been measured in terms of Precision and Recall. Both precision and recall have produced the satisfactory results in terms of product review auto classification and automatic text summarization. The precision has been recorded near 98%, whereas the recall values have been measured at almost 99%. The existing system has been measured at almost 46% recall value, which is way lower than the recall produced by the proposed model. The proposed model has been also evaluated for its performance on the sentiment analysis. The sentiment analysis is the core system in the proposed model. Product review evaluation and automatic product review summarization depends upon the sentiment analysis report. The sentiment analysis report generates the emotion weights of the messages or reviews given by the users of the product specific. The specific product review classification is automatically done on the basis of product review ontology. The sentiment analysis system has been well tested for its performance on the various numbers of the product reviews. The system has been tested with 100, 1000, 10000 and other number options of the product reviews. The system has been proved its accuracy at nearly 96% for the sentiment analysis, which makes it the robust system.

6. Future Work

In the future, the proposed model will be designed for multiple product review and product feature classification. Also the proposed model can be attached with some of the web source offering the product review API to classify the large amounts of data automatically. The proposed model can be enhanced for different N-gram and product feature classification.

References

[1] Li, Jinyan, Simon Fong, Yan Zhuang, and Richard Khoury. "Hierarchical classification in text mining for sentiment analysis of online news." Soft Computing (2015): 1-10.

[2] Prollochs, Nicolas, Stefan Feuerriegel, and Dirk Neumann. "Enhancing Sentiment Analysis of Financial News by Detecting Negation Scopes." InSystem Sciences (HICSS), 2015 48th Hawaii International Conference on, pp. 959-968. IEEE, 2015.

[3] Cui, Limeng, Fan Meng, Yong Shi, Minqiang Li, and An Liu. "A Hierarchy Method Based on LDA and SVM for News Classification." In Data Mining Workshop (ICDMW), 2014 IEEE International Conference on, pp. 60-64. IEEE, 2014.

[4] Ouyang, Yuanxin, Yao Huangfu, Hao Sheng, and Zhang Xiong. "News Title Classification with Support from Auxiliary Long Texts." In Neural Information Processing, pp. 581-588. Springer International Publishing, 2014.

[5] Bieliková, Mária, Michal Kompan, and Dušan Zeleník. "Effective hierarchical vector-based news representation for personalized recommendation."*Computer Science and Information Systems* 9, no. 1 (2012): 303-322.

[6] Krishnlal G, S Babu Rengarajan, K G Srinivasagan, " A new text mining approach based on HMM-SVM for web news classification" International Journal of Computer Applications (0975-8887) Volumn 1- No.19,2010.

[7] Vandana Korde, C namrata Mahender, "Text classification and classifier a survey," International Journal of Artificial Intelligence and Application (IJAIA), vol.3, No.2, March2012.

[8] Mita K. Dalal, Mukesh A.Zaveri," Automatic text Classification," International Journal of Computer Applications (0975-8887) Volumn 28- No.2, August 2011.

[9] Ouyang, Yuanxin, Yao Huangfu, Hao Sheng, and Zhang Xiong. "News Title Classification with Support from Auxiliary Long Texts." In Neural Information Processing, pp. 581-588. Springer International Publishing, 2014.

[10] Balahur, Alexandra, Ralf Steinberger, Mijail Kabadjov, Vanni Zavarella, Erik Van Der Goot, Matina Halkia, Bruno Pouliquen, and Jenya Belyaeva. "Sentiment analysis in the news." *arXiv preprint arXiv:1309.6202* (2013).

[11] Yu, Liang-Chih, Jheng-Long Wu, Pei-Chann Chang, and Hsuan-Shou Chu. "Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news." *Knowledge-Based Systems* 41 (2013): 89-97.

[12] Dilrukshi, Inoshika, and Kasun De Zoysa. "Twitter news classification: Theoretical and practical comparison of SVM against Naive Bayes algorithms." In *Advances in ICT for Emerging Regions (ICTer), 2013 International Conference on*, pp. 278-278. IEEE, 2013.

[13] Bandari, Roja, Sitaram Asur, and Bernardo A. Huberman. "The pulse of news in social media: Forecasting popularity." *arXiv preprint arXiv:1202.0332*(2012).

[14] De Choudhury, Munmun, Nicholas Diakopoulos, and Mor Naaman. "Unfolding the event landscape on twitter: classification and exploration of user categories." In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pp. 241-244. ACM, 2012.

[15] Hagenau, Michael, Michael Liebmann, Markus Hedwig, and Dirk Neumann. "Automated news reading: Stock price prediction based on financial news using context-specific features." In *System Science (HICSS), 2012 45th Hawaii International Conference on*, pp. 1040-1049. IEEE, 2012.