International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-6, March 2020

Exploring Classification Techniques for Sentiment Analysis

Mahesh G., Satish Kumar T., Shreya S., Sushmitha N., Sripad T.

Abstract*: *In the recent years, there has been tremendous improvements in the Information Technology industry. Every day 2.5 quintillion bytes of data is generated, this provides sentiment analysis to become a key tool to make something out of it. This has allowed companies to grab this key acumen and automate the needed processes. This information can serve as good source for extracting useful insights regarding the products/services. It would be helpful for the product owners if these thousands of reviews could be summarized using the latest technologies. Also it would be more useful if the sentiment scores are provided for each aspects of products/services. This paper presents Sentiment Analysis for various predefined aspects for Hotel Reviews. We have used Naive Bayes Classifier for classifying hotel reviews as positive or negative and we have evaluated the performance and suggested few future enhancements.

Keywords*: *Sentiment Analysis, Opinion Mining, Naive Bayes.

I. INTRODUCTION

In today's world, the rising usage of the internet and the services provided through it has scaled very high over the years. This brings in the question of quality and genuineness of these services which go through a tough competition out in the wild web. One way for the customers to build trust in these services, they have always sought to criticism and reviews.

Criticism always helped us to give a contemporary perspective and enlighten us to things we may have omitted or never considered. Whether it's a customer review, professional critique or a rival's statement, productive and practical criticism and feedback can guide you to flourish by elucidating and providing the means to improve and grow.

It is observed that about 92% of professionals in the field of marketing envisage that the social media has deep-seated repercussion on the business, which implies that there is more competition than you think in the game of social media to captivate your potential customers. To do this is not an impassable task but building an accurate sentiment analysis will assure that you win.

Basically, Sentiment analysis is the tool to find and fathom about how the customers feel about your product or service. It is also known as opinion mining and is much reported about but often a misconceived term. Sentiment analysis is the process of understanding and determining attitudes of customer, their opinions and emotions conveyed within an online mention with the emotional tone behind a series of words.

Among all the services, hotels play a prominent role in a country's economy and also offering hospitality to travellers since the earliest civilizations. The need for proper facilities to be provided has been a challenge to both the hotel owners and the customers. Now since the culture of marketing and reviews has spread over internet and media, there is a constant need for sentiment analysis for the hotel owners on their customer reviews just that they can understand and provide better for their customers.

All the reviews are directly made available to the owner of the product so that he/she can read it manually and make amendments to his/her products for the better in the future. But the main problem that we face here is that it is very difficult for the owner to go through all these reviews manually especially when the product or the service is very popular. In such scenarios what happens is that most of these owners skip through all these reviews and try to finish up looking at all the reviews. Due to which the very essence of taking the review for the products and services loses its effectiveness and as a result, the product manufacturer doesn't get any information about what features of his or her product did the customers like and dislike.

According to the feedback review, we have to classify whether the review is written is the positive, negative, or neutral sentiment. There exist few reviews where it conveys both positive and negative sentiment, for such reviews we should choose the sentiment which is stronger than the other. Businesses are trying their best to unlock the hidden value of text reviews in order to understand their customers' opinions (Customer Satisfaction Assessment) and to make a better business decision and more informed, (Brand Perception Analysis). Traditionally the common businesses have entrusted on workshops, surveys, and groups to gain better insights into their customers' opinions and emotions.

Manuscript received on February 10, 2020.
Revised Manuscript received on February 20, 2020.
Manuscript published on March 30, 2020.
* Correspondence Author
Dr. Mahesh G.*, Department of Computer Science and Engineering, B M S Institute of Technology and Management, Bangalore, India.
Dr. Satish Kumar T., Department of Computer Science and Engineering, B M S Institute of Technology and Management, Bangalore, India.
Ms. Shreya S., Department of Department of Computer Science and Engineering, B M S Institute of Technology and Management, Bangalore, India.
Ms. Sushmitha N., Department of Computer Science and Engineering, B M S Institute of Technology and Management, Bangalore, India.
Mr. Sripad T., Department of Computer Science and Engineering, B M S Institute of Technology and Management, Bangalore, India.

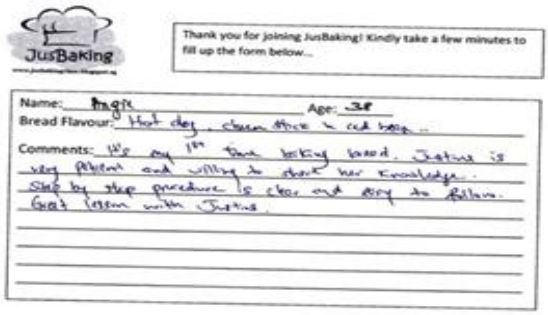
Using the new age technology we will be able to use the power of Natural Language Processing, Data mining and Machine Learning to extract meaning from textual reviews and get deeper into the thoughts of the customers and learn to see them outside of the controlled environment such as a survey.

This paper is organized as follows: section II presents the previously developed research works related to this project. Section III describes the design of our sentiment analysis model. Section IV presents a general algorithm of the project, a brief summary of the tools used and the snapshots of our work. Section V presents the results that are obtained along with discussions regarding it. Section VI presents the conclusion and future work.

II. LITERATURE REVIEW

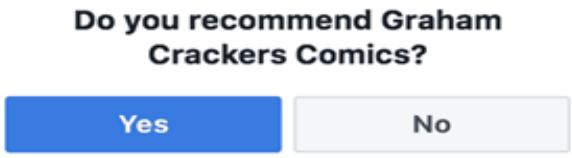
This section describes some of the older methods of review system and we have referred to the previous related works and tried to draw conclusion on the method preferred for hotel reviews. The Review System was introduced in order to inform the manufacturer of a specific product about the faults in it so that he can use this information and know what faults he needs to rectify in his next upcoming product. Few of the older methods of review system are:

A. Handwritten Reviews



Disadvantage - Sloppy Handwriting And Time Consuming.

B. Yes / No Reviews



Disadvantage - Not much information.

C. Star Rating Review

Disadvantage - Less Expressive And Limited Review

D. Online Text Review

Disadvantage - Difficult to Analyse And Time Consuming

E. Present Review system (combination of star rating and textual review)

Jitendra, Kim-Kwang, Amiya, SambitBakshi, Sanjay Kumar, Karen [1], used both in supervised and unsupervised approaches on various datasets. They concluded that by combining supervised approach with unigram, bigram and Part-of-Speech resulted in more efficient results. Minqing Hu, Bing Liu [2], have used several techniques for providing aspect-based summary of reviews of an online product provided by customers.

P. BavithraMatharasi, Dr A.Senthilrajan [3], used the naive Bayes classification algorithm on twitter dataset and showed many errors that could not be handled properly by the algorithm. Huma Parveen, Prof. Shikha Pandey [5], have again used naive Bayes on twitter dataset. They have concluded that when performing pre-processing by considering emoticons resulted in better performance and accuracy.

Upma Kumari, Dr.Arvind K Sharma, Dinesh Soni [7] have used different classification models on smart phone reviews and concluded that support vector machine algorithm has achieved highest accuracy compared to others. Preety, Sunny Dahiya [9] have proposed a method by combining naive Bayes and modified k-means on mobile reviews and stated that it gives better accuracy compared to techniques considered individually

Shweta Rana, Archana Singh [10] have worked on movie reviews using naive Bayes, linear SVM and synthetic words. They have concluded that linear SVM performs better among the three. Ms Aarti A. Patil, Prof. Seema Kolkur [4], have applied the advanced naive Bayesian algorithm on product reviews and achieved a precision of 82.85%.Joana, Alcione de Paiva, Guidson Coelho, Alexandra [13],

presented a convolutional neural network approach on hotel reviews and stated that CNN has performed better than the previously considered models on the same corpus.

Vijay B. Raut, Prof. D.D. Londhe [15], have basically presented a summary of sentiment analysis approaches and stated that machine learning approaches perform better for reviews regarding the movie, product, hotel etc and lexicon-based approach is more suitable for blogs, tweets and comments on the web. Chhaya Chauhan, Smriti Sehgal [17], have presented a review of algorithms and techniques for feature wise sentiment analysis of product reviews. They have concluded when good features are extracted, better results were given by naive Bayes classifier.

Naive Bayes Algorithm

Bayesian network classifiers are one of the most popular classification paradigms which use a supervised approach. A well-known of its kind of classifier is the Naïve Bayes' classifier which is based mainly on probability on the Bayes' theorem, and mainly it considers Naïve (Strong) independence assumption.

The starting point is to calculate conditional probability, considering, x for a point in data and class C:

$$P(C / x) = P(x/C)/P(x)$$

Further, make the assumption x = {x1,x2,...xj} is independent of each other,, we can estimate the probability of x as follows:

$$P(C/x) = P(C) . \prod P(xi/C)$$

Vader (Valence Aware Dictionary and sEntiment Reasoner) is one of the open source lexicon and rule based sentiment analysis tool. It uses sentiment lexicons which are generally labelled either as positive or negative. It not only specifies whether a sentence is positive or negative but also how much positive and negative it is. It considers emojis, slangs and acronyms in the sentences.

III. SYSTEM MODEL

This section describes the overall process of our model and detailed design of the subcomponent systems.

The proposed methodology discusses how to apply sentiment analysis algorithms and machine learning techniques to gain insights and knowledge about the online reviews for hotels and the revenue of the performance.

We extract these online reviews from various platforms and store in a local database for further pre-processing and analysing purposes.

Pre-processing of data is important to remove unwanted, noisy and inconsistent data. Pre-processing is a mandatory step inorder to proceed with any data mining functionality. We used the following pre-processing activities before proceeding with any of the sentiment analysis approaches - converting into lowercase, removing punctuation, digits, newline characters, special characters etc, performing a spell check.

Certain features of the hotel industry are considered. Reviews pertaining to these features are extracted into various files for which classification algorithms will be applied for sentiment analysis.

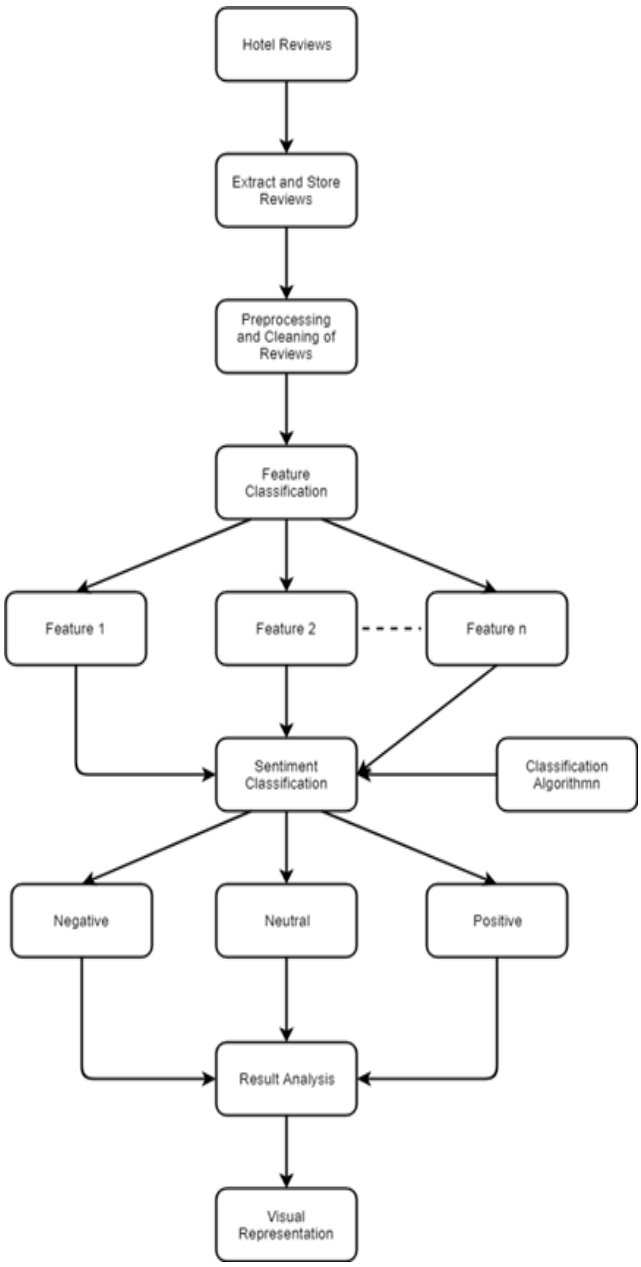


Fig. 1.System Model

A supervised learning model needs to be trained with training data which are labelled and generates trained model that can be used for classifying new instances. Ideally, this model can classify accurately the unknown data instances to labels as they are. This is approximated by the generalization of the training data instances so that the trained model can label new instances in an accurate way.

For every feature discovered, related opinions are segregated into positive and negative classes. We can show what part of reviews are positive/negative about each feature. In this manner, all the aspects present are ranked based on their appearances in the review dataset.

To represent the derived output, graphs and plots will be used to show the frequency of words in the customer reviews and the sentiment scores.

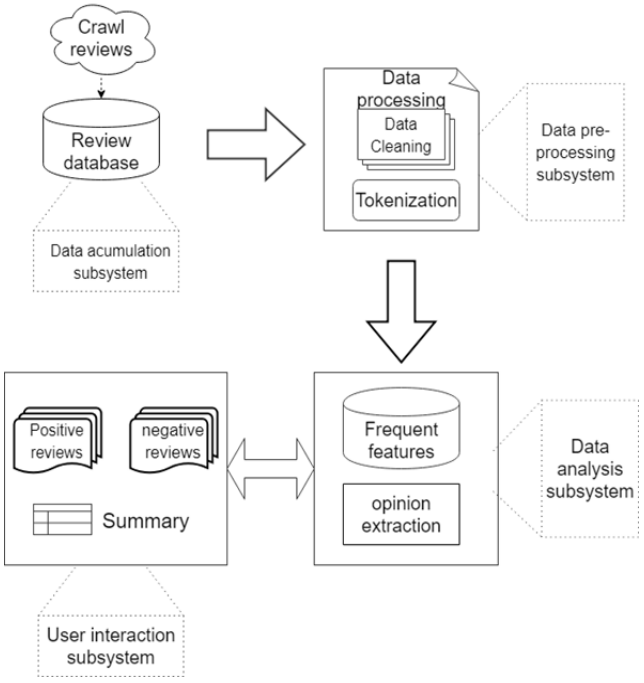


Fig. 2.High Level Design

There are four subsystems in our model.

- *Data Accumulation:* The optimal review accumulation is a process of collecting opinion reviews from major reviews providing channels. Many review channels will contain customer opinions for different products, services, movies, hotels, news etc. We should be considering only the reviews which contain subjective data pertaining to our interests but not the objective data.
- *Data Preprocessing:* Pre-processing of data is one step in data mining technique that involves converting raw input data obtained into a format that is understandable to the machine. Real-time data often is inconsistent, not complete, containing missing values, and usually contains a lot of errors. One of the methods proven to resolve such issues is Data pre-processing. This step of pre-processing of data prepares the obtained raw input data suitable for further processing.
- *Data Analysis:* Feature extraction is the tedious process of collecting discriminative information from a collection of samples. Feature classification is process of grouping features based on some known criteria. Sometimes feature classification can also be related to selection of feature which is to obtain a subset of the extracted features that would optimize the machine learning algorithm by possibly reducing noise removing unrelated features.
- *User Interaction:* With the help of our sentiment analysis tools, the obtained unstructured data should be transformed into structured data and represented using GUI, graphs, plots etc. This system can be very useful for market target applications like public relations, marketing analysis, product reviews, feedback, and customer service.

IV. IMPLEMENTATION

This section presents a general algorithm of our model, and description of the tools used.

Algorithm

Input: *Training-Dataset T,*
Testing Dataset t,
Hotel-Reviews h .

Output: *Feature-wise sentiments of reviews for each h.*

Steps:

1. Pre-processing the datasets T, t and h for spell check, punctuation removal, digits removal, lower case conversion, stripping etc.
2. Comparing each review in h with predefined set of features and segregating it into respective files f.
3. Train the model using Multinomial NB using T.
4. Test the model for accuracy, precision and recall score using t.
5. For each h
   For each review in f
   a. Run the classifier algorithm
   b. Predict the sentiment
   Calculate the cumulative sentiment score for f
6. Plot the sentiment scores obtained of each hotel h and display using graphs.

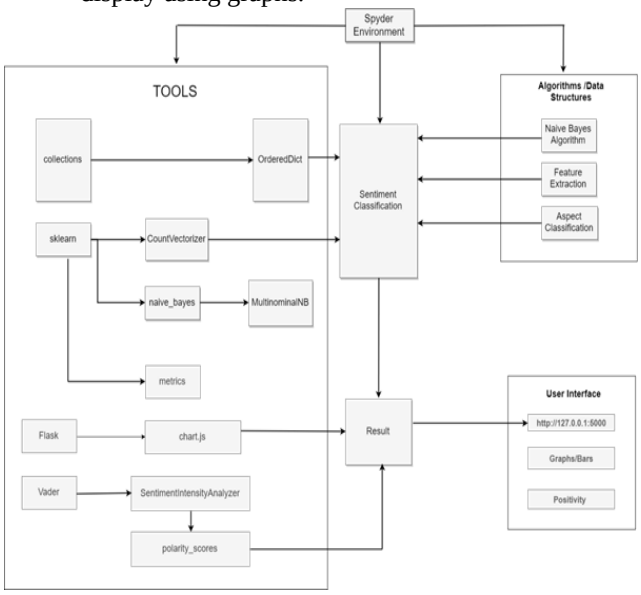


Fig. 3.Tools Used

- Spyder is the IDE which we have used for our development work.
- Scikit-learn is a machine learning library in Python which contains many classifications, regression, clustering algorithms.
- We are using the Naive Bayes algorithm for our review classification purpose. Multinomial NB classifier is a part of the Naive Bayes algorithm mainly used for classification with discrete features.
- Count Vectorizer makes it easy for us to tokenize a text document.
- VADER(Valence Aware Dictionary and sEntiment Reasoner) is a rule-based sentiment analysis tool which uses its own lexicons that are labelled.
- Flask is a web application framework which allows us to integrate python scripts with the frontend HTML code.

Retrieval Number: E6897018520/2020©BEIESP
DOI:10.35940/ijrte.E6897.038620
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

V. RESULTS

Some of the standard metrics to evaluate how good our model is performing are Precision, recall, and accuracy.

Confusion matrix = [[TP FP]
[FN TN]]

Precision = TP / [TP+FP]

Recall score = TP / [TP+FN]

Accuracy = (TP+TN) / (TP+FP+FN+TN)

The evaluation metrics of our trained classification model is shown below:

Confusion matrix = [[863 135]
[103 897]]

Precision = 0.8691860465116279

Recall score = 0.897

Accuracy = 0.8808808808808809

Fig 4. shows the feature-based summary of the four hotel review datasets considered using Naïve Bayes Classifier. Fig 5. shows the feature-based summary of the four hotel review datasets considered using VADER Classifier.

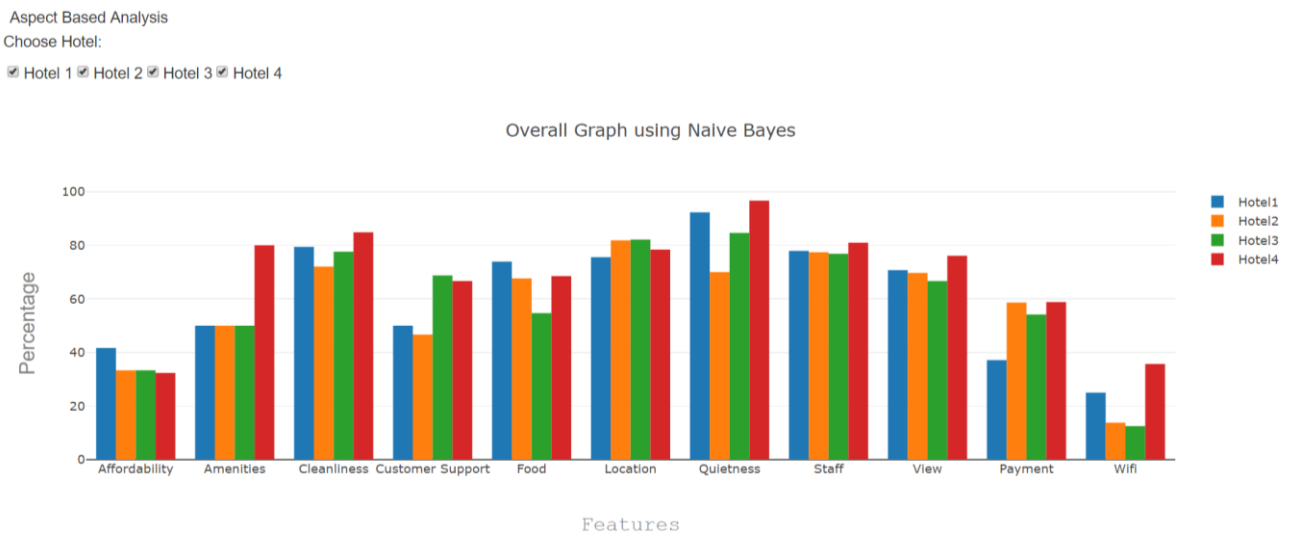


Fig. 4.Feature Based Summary Using Naive Bayes Classifier

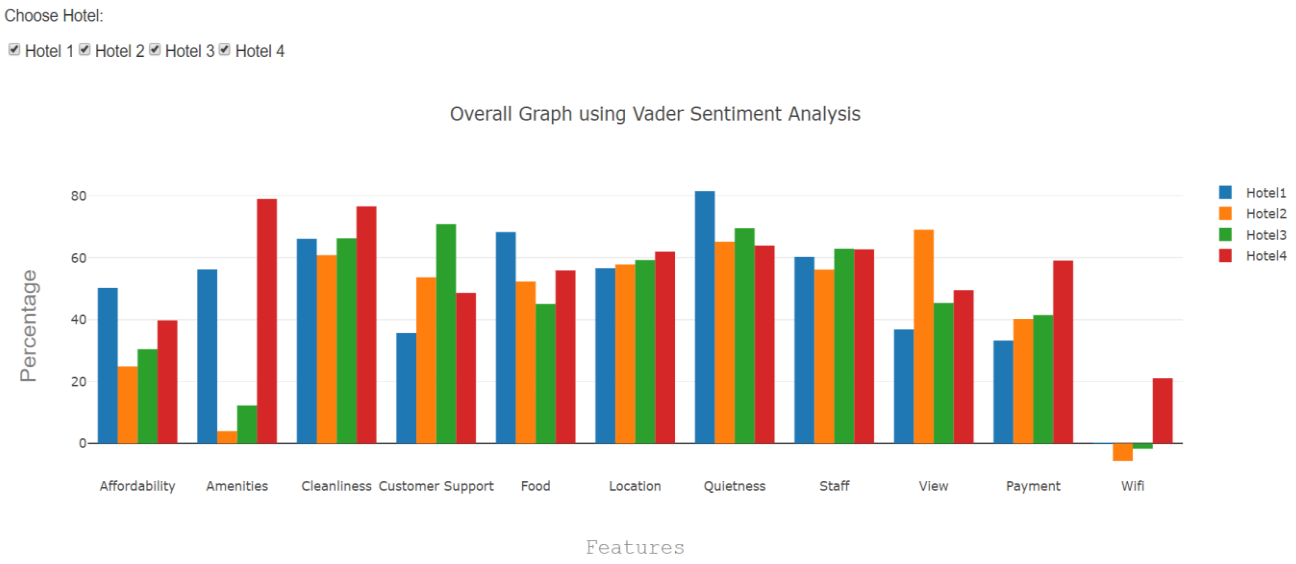


Fig. 5.Feature Based Summary Using VADER Classifier

According to these graphs, hotel management can understand in which aspect their hotel is lacking and needs improvement. For example, hotel 3 should improve in providing a better internet facility and making it more affordable to customers. From a customer's perspective, he/she can come to a conclusion that hotel 1 is more affordable among all the 4 using this comparison graph. In case, a customer wants to look into cleanliness of the hotel, hotel 4 is more better compared to others in cleanliness.

VI. CONCLUSION

The customers' opinions on products and services have become one of the important ways of communication due to the expansion of social media. We have covered sentiment analysis using machine learning tools like Naive Bayes and VADER sentiment - a built-in library tool.

This paper presents one of the many available methods of aspect-based sentiment analysis of hotel reviews by considering four hotel review datasets. We have used the Naive Bayes algorithm and VADER for classifying the reviews as positive or negative. We have used a pre-classified set of reviews as the training dataset. Testing our model with test dataset resulted in 88.08% accuracy using Naive Bayes.

For future work, we would like to try and come up with an efficient and more optimized sentiment analysis using various algorithms such as random forest, decision trees, Support Vector Machine and so on. Also, there can be a possibility to improve accuracy by finding the combinations of these algorithms. We also would try to analyse slang and sarcasm in order to correctly predict the opinion of the reviewer and to improve our models to do better data forecasting.

REFERENCES

1. Jitendra Kumar Rout, Kim-Kwang Raymond Choo, Amiya Kumar Dash, SambitBakshi, Sanjay Kumar Jena and Karen L. Williams, "*A model for sentiment and emotion analysis of unstructured social media text*", Electronic Commerce Research, Volume 18, Issue 1, March 2018, pp. 181–199.
2. Minqing Hu and Bing Liu, "*Mining and Summarizing Customer Reviews*", In Proceedings of the tenth ACM SIGKDD International Conference on Knowledge discovery and data mining, Seattle, Washington, USA, August 22–25, 2004, pp. 168-177.
3. P. BavithraMatharasi and Dr.A.Senthilrajan, "*Sentiment Analysis of Twitter Data using Naïve Bayes with Unigram Approach*", International Journal of Scientific and Research Publications, Volume 7, Issue 5, May 2017, pp. 337-341.
4. Ms. Aarti A. Patil and Prof. Seema Kolkur, "*Sentiment Analysis for Product Reviews*", International Journal of Advanced Research in Computer Science, Volume 5, Issue 5, 2014, pp. 202-204.
5. Huma Parveen and Prof. Shikha Pandey, "*Sentiment Analysis on Twitter Data-set using Naive Bayes Algorithm*", In proceedings of Second International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), Bangalore, 21-23 July, 2016, pp. 416-419.
6. Pooja jain, Neetu verma, "*Sentiment Analysis using Naïve Bayes Classifier*", International Journal of Innovations in Engineering and Technology, Volume 8 Issue 2, April 2017.
7. Upma Kumari, Dr.Arvind K Sharma and Dinesh Soni, "*Sentiment Analysis of Smart Phone Product Review using SVM Classification Technique*", In Proceedings of International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS-2017), Chennai, 1-2 August,2017, pp. 1469-1474.
8. Bhumika M. Jadav ,Vimalkumar and B. Vaghela, "*Sentiment Analysis using Support Vector Machine based on Feature Selection and Semantic Analysis*" ,International Journal of Computer Applications, Volume 146, No.13, 2016, pp. 26-30.
9. Preety and Sunny Dahiya, "*Sentiment Analysis Using SVM And Naïve Bayes Algorithm*", International Journal of Computer Science and Mobile Computing Volume 4, Issue 9, 2015, pp. 212 – 219.
10. Shweta Rana and Archana Singh, "*Comparative Analysis of Sentiment Orientation Using SVM and Naïve Bayes Techniques*", In Proceedings of Second International Conference on Next Generation Computing Technologies (NGCT-2016) Dehradun, India, 14-16 October 2016, pp. 106-111.
11. Mathieu Cliche, "*BB_twtr at SemEval-2017 Task 4: Twitter Sentiment Analysis with CNNs and LSTMs*", In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), August 2017, pp. 573-580.
12. Marzieh Saeidi, Guillaume Bouchard, Maria Liakata and Sebastian Riedel, "*SentiHood: Targeted Aspect Based Sentiment Analysis Dataset for Urban Neighbourhoods*", In Proceedings of the 26th International Conference on Computational Linguistics, Osaka, Japan, December 2016, pp. 1546–1556.
13. Joana Gabriela Ribeiro de Souza, alcione de Paiva Oliveira, Guidson Coelho de Andrade and Alexandra Moreira, "*A Deep Learning Approach for Sentiment Analysis Applied to Hotel's Reviews*", In Proceedings of International Conference on Applications of Natural Language to Information Systems, Natural Language Processing and Information Systems, Salford, United Kingdom, 26-28 June 2018, pp. 48-56.
14. Yi Tay, Luu Anh Tuan and Siu Cheung Hui, "*Dyadic Memory Networks for Aspect-based Sentiment Analysis*", CIKM'17 , Singapore, DOI:10.1145/3132847.3132936, 2017.
15. Vijay B. Raut and Prof. D.D. Londhe , "*Survey on Opinion Mining and Summarization of User Reviews on Web*" ,International Journal of Computer Science and Information Technologies, Volume 5, Issue 2, 2014, pp. 1026-1030.
16. AmlaanBhoi and Sandeep Joshi, "*Various Approaches to Aspect-based Sentiment Analysis*", arXiv:1805.01984v1 [cs.CL] 5 May 2018.
17. Chhaya Chauhan and Smriti Sehgal, "*Sentiment Analysis on Product Reviews",* In Proceedings of International Conference on Computing, Communication and Automation (ICCCA2017), Greater Noida, 5-6 may 2017, pp. 26-31.
18. Walter Kasper and Mihaela Vela, "*Sentiment Analysis for Hotel Reviews*", In Proceedings of the Computational Linguistics-Applications Conference, Jachranka, 2011, pp. 45–52.
19. Nadeem Akhtara, NashezZubaira, Abhishek Kumara and Tameem Ahmadi, "*Aspect based Sentiment Oriented Summarization of Hotel Reviews*", In Proceedings of Seventh International Conference on Advances in Computing & Communications, ICACC-2017, Cochin, India, 22-24 August 2017, pp. 563-571.

AUTHORS PROFILE

Dr. Mahesh G., holds B.E., M.Tech and Ph.D in Computer Science & Engineering from VTU. He was a 2nd Rank holder in M.Tech. He is currently working with Department of CSE, BMSIT. Prior to this he was associated with Acharya Institute of Technology, Bangalore. He has 15 years of professional experience, which spans from academics, research and consultancy. He has published around 20 papers in reputed International Journals / Conferences. His current research interests include stochastic and Petrinets modeling of wireless networks. He is a member of Society of Digital Information & Wireless Communication, International Association of Engineers and Indian Society for Technical Education. He was also the BOS member of Dr.AIT, Bangalore. Cognizant Technology Solutions has honored him with the Best Faculty Award in 2017.

Dr. Satish Kumar T., holds B.E from Bangalore University, M.Tech and Ph.D in Computer Science & Engineering from ANNA University. He is currently working with Department of CSE, BMSIT. Prior to this he was associated with RNS Institute of Technology, Bangalore. He has 19 years of professional experience, which spans from Industry, academics, research and consultancy. He has published around 28 papers in reputed International Journals / Conferences. His research primarily focuses on Code Optimization on High performance Computing (HPC) systems using heuristics methods. Specifically, using Multi-core clusters with high degree of computations. Over the years Code optimization, while highly relevant to HPC is slowly picking up grounds in mobile computing and IOT. His current research extends to development of Compiler Optimization Algorithms, design of RTOS and Embedded system design, and its synchronization. He is a member of Indian Society for Technical Education and IEEE. He was also the BOE member of VTU, Belagavi and journal reviewer for several journals.

Ms. Shreya S., holds B.E in Computer Science & Engineering from BMS Institute of Technology affiliated under VTU. She has consistently remained one among top 10 in her college. She is currently working as a software engineer at Temenos. Prior to this she has worked as an intern at Guestasy Pvt Ltd as a Software Developer and at PLV Technology as a Web developer.

Ms. Sushmitha N., holds B.E in Computer Science & Engineering from BMS Institute of Technology affiliated under VTU, Belgaum. She is currently working as a Software Developer at DXC Technology. Priorly, she has interned at Guestasy Pvt Ltd as a Software Developer and at Quadwave Consulting Pvt Ltd as a Web developer.

Mr. T Sripad holds B.E. in Computer Science & Engineering from VTU. He has secured Silver Medal in Cyber Olympiad competition. He is currently working in Evertz Microsystems as Jr. Project Engineer. Prior to this he was associated with Google's Developer Student Club(DSC's) in BMSIT, as core TA of the club he has Conducted workshop on Android Application Development in the college. He has also been a Treasurer and Board Member in Rotaract Club BMSIT, Yelahanka (RotaryClub) under which he has Project Chaired fundraising events to help educating poor children. He holds certification in Microsoft Technology Associate, Game Development and Android Development.