Sentiment Analysis and Summarization of Social Media Content Using Topic Modeling

Prerna Mishra, Ranjana Rajnish, Pankaj Kumar

Abstract: ***Explosion of Web 2.0 had made different social media platforms like Facebook, Twitter, Blogs, etc a data hub for the task of Data Mining. Sentiment Analysis or Opinion mining is an automated process of understanding an opinion expressed by customers. By using Data mining techniques, sentiment analysis helps in determining the polarity (Positive, Negative & Neutral) of views expressed by the end user. Nowadays there are terabytes of data available related to any topic then it can be advertising, politics and Survey Companies, etc. CSAT (Customer Satisfaction) is the key factor for this survey companies. In this paper, we used topic modeling by incorporating a LDA algorithm for finding the topics related to social media. We have used datasets of 900 records for analysis. By analysis, we found three important topics from Survey/Response dataset, which are Customers, Agents & Product/Services. Results depict the CSAT score according to Positive, Negative and Neutral response. We used topic modeling which is a statistical modeling technique. Topic modeling is a technique for categorization of text documents into different topics. This approach helps in better summarization of data according to the topic identification and depiction of polarity classification of sentiments expressed.***

Keywords: Opinion Mining, Sentiment Analysis; Topic Modeling, LDA.

I. INTRODUCTION

Sentiment analysis is contextual mining of text that analyzes the inclination of people's opinions or views. It is one of the basic problems in NLP and is the process that can extract opinions or views expressed from natural language. Sentiment analysis is a classification problem, which mines opinions from reviews and helps in a wide range of applications. NLP is a branch of Artificial Intelligence, which is used in analyzing & understanding the language that humans naturally use to interact with computers. By the analysis of data through different Natural Language processing techniques, classification of data is done whether sentiments expressed are Positive, Negative or neutral.

Revised Manuscript Received on January 05, 2020
***** Correspondence Author
Prerna Mishra*, Amity Institute of Information Technology, Amity University Uttar Pradesh, Lucknow Campus, Lucknow (India), India. Email: prerna21.mishra@gmail.com
Ranjana Rajnish, Amity Institute of Information Technology, Amity University Uttar Pradesh, Lucknow Campus, Lucknow (India), India Email: rrajnish@lko.amity.edu
Pankaj Kumar, Department of Computer Science, Sri Ramswaroop College of Engineering & Technology, Lucknow, India
Email: pk79jan@gmail.com

Polarity classification (Positive, Negative, Neutral) of data helps different product reviewers, in forecasting stock prices, companies, etc to improve their products and services.

With the growing use of the internet, there is a huge data bank of opinions or views expressed by different people. Due to this increasing rate of data availability, people want to get some meaningful or useful information from it. Sentiment analysis / Opinion Mining are used to analyze the opinion and classify them into three categories as Positive, Negative and Neutral, depending upon the views expressed.

However, the data presented by different sources of Social media like Facebook, Twitter, blogs & forums is unstructured. Fig 1 depicts the percentile of data available on the internet which is unstructured and of no use to the consumers due to noisy words, use of slang words and emoticons.

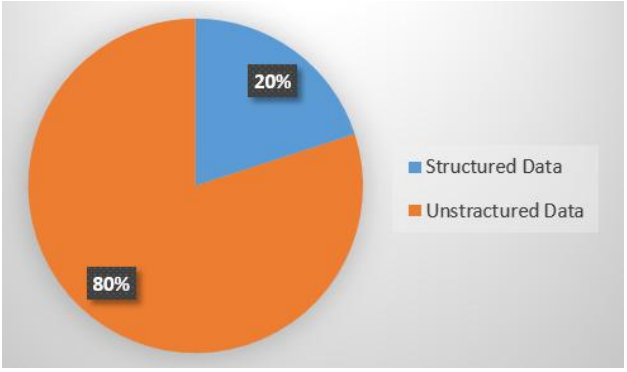


Fig 1-Percentile of Data available on Internet

Study shows that almost 80% of data is unstructured and not organized in a proper manner to get meaningful results. The task of analyzing this type of data is time consuming and cumbersome. Sentiment analysis is a Data mining task, which makes this unstructured data structured by using Natural Language processing techniques. This process is known as Web Scraping. Therefore, by automating the data, a lot of time saved for manual processing and easy for polarity classification (Positive, Negative & Neutral)

There is a saying by Bill Gates-

"**Most unhappy customers are greatest source of learning**"

Therefore, online reviews are a great source of information for consumers. These online reviews even help sellers to gauge the consumer's feedback on the products. Mostly manufacturing companies interested in a different type of questionnaires, this helps to understand the consumer's feedback about any product or services.

- What the customer thinks about the product/services he using.
- What are the expectations from the product/ services and why the seller failed to meet those?
- Quality improvement, as how happy or sad the customer is about any product or service.

CSAT (Customer Satisfaction) is the key factor of this survey response. Results depict the CSAT score according to Positive, Negative and Neutral response.

II. RELATED WORK

Alexander Pak, Patrick Paroubek (2010) has proposed a linguistic analysis of collected corpus from twitter and builds a sentiment classifier. Build Sentiment classifier and performed experimental measures for performance evaluation with existing method, and concluded results with better efficiency measure of proposed method [2].

Jingyi Ye1, Xiaojun Jing1,Jia L (2018) has proposed a modified latent Dirichlet allocation model and used support vector machine for analyzing the sentiments of subjective texts. In this paper results shown that Modified LDA performance is much better in comparison to the traditional LDA model [3].

Ali Daud (2011) presented a modeling approach which is time modeling and named as Temporal-Author-Topic, and is used for modeling the text simultaneously to overcome the problem of exchangeability of topics [4].

ToqirA.Rana, Yu-NCheah, Sukumar Letchmunan (2016), has presented a topic modeling review, using LDA-based techniques for analysis of sentiments. They presented a detailed analysis of approaches and comparison of accuracy measure for various systems. Lot of effort is done to explore various topic-modeling techniques in the capacity of sentiment analysis [5].

David M. Blei, Andrew Y. Ng, Michael I. Jordan (2003), described latent Dirichlet allocation (LDA) for collections of data such as text corpora. LDA is a generative probabilistic model, which is used for discrete data collections and the problem of modeling the corpora text [8].

III. TECHNIQUES AND APPROACH USED IN OPINION MINING

Sentiment Analysis can be classified into two approaches, and these two approaches are widely used nowadays for gathering public opinion.

- Lexicon Analysis-In Lexicon Analysis, the method of semantic orientation of words in the document is used to calculate polarity of the document.
- Machine Learning-It uses the different building models from labeled trained dataset to calculate the polarity of documents.

To analyze the opinions there are different Algorithms which are broadly classified as-

- Rule-Based System- Analysis is done by using manually made rules. In this system a human expert's knowledge is encoded into an automated system. This system consists of a set of IF-THEN rules [1].
- Automatic Based System- Use different Machine learning techniques for analysis like Naive Bayes, Support Vector Machine, Max Entropy, etc. Nowadays Machine Learning techniques are the most widely used approach for the Polarity classification of opinions with better efficiency [6].
- Hybrid System- Combination of both Hybrid and Automatic based system.

In this paper, we used Topic modeling which is an unsupervised class of machine learning approach. LDA is the widely used topic modeling technique for discovering the topics.

IV. TOPIC MODELING USING LDA

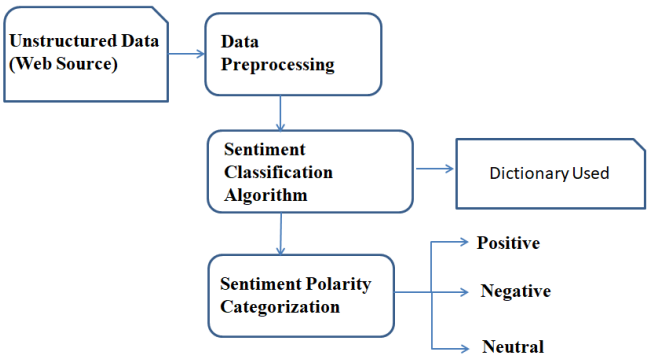


Fig. 2.Sentiment Analysis Process

Process of sentiment analysis is depicted above in fig 2. There are few steps, which need to follow for better results. Detailed explanations of sentiment analysis process are explained below.

A. *Corpus Collection*

First step for Sentiment analysis is Data Collection. Corpus contains about 900 responses, which are expressed by different customers related to survey. 80 % of data, which is required for analysis, is unstructured data that is it requires cleaning process.

Thank you to {Namepii} she was very good.
A very friendly and helpful contact .
The real person was very good but the robot at the beginning is a disaster. V
Was helpful and patient with me during my issue. Appreciated her help.
she was very patient with me since I'm somewhat computer illiterate#N#she was very knowledgeable
The lady who helped me was wonderfully friendly and understanding with no hesitations to ensure i was
{Namepii} and {Namepii} were very patient helpful and efficient. Thank you both for exceptional service.
The remote technician John D was finally able to resolve the issue and was extremely polite and helpful.
A very polite and helpful technician. Many thanks.
Very nice tech agent. The problem will hopefully be solved soon. I still am unable to send and receive em
Polite and somewhat helpful. Beyond technical experience gave forwarding contact info.

Fig.3. Data sample of Survey Response

B. *Data Preprocessing*

Unstructured data which is collected from the data repository is not that precise. For better accuracy and result data collected requires cleaning. Data Cleaning is the step in which we filter the data for further processing like removal of whitespace, removal of punctuation, converting alphabets into lower case etc. We used tidytext package (Silge, Robinson and Hester2016) is an R package used for text mining.

C. *Sentiment Classification*

After cleaning of data, data is ready for analysis. For this, we use the Syuzhet package which comes with four sentiment dictionaries. First, we did analysis using "sentiment" package and matched the specified words in the dictionary to get the specified score. Some packages which are required to load for analysis are dplyr, sentiment, lexicon, textclean etc.

Dictionary used for calculating score is "hash_sentiment_jockers". We added the word in this dictionary also which is not present for analysis by assigning the score value. Sentiment score is calculated on the based on the words expressed in the sentences by customers for the survey.

D. *Sentiment Polarity*

After classification next step is categorization, which displays either the opinion expressed is Positive, negative or neutral. If the score calculated is positive score then data is categorized as positive, if negative score then its categorization is as negative and if it is zero then the opinion expressed is neutral.

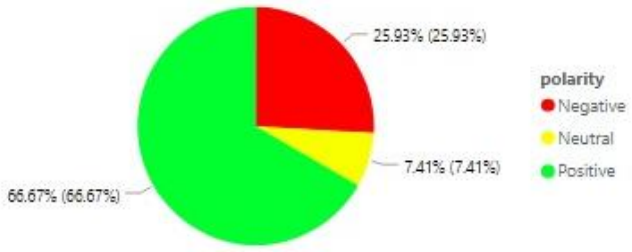


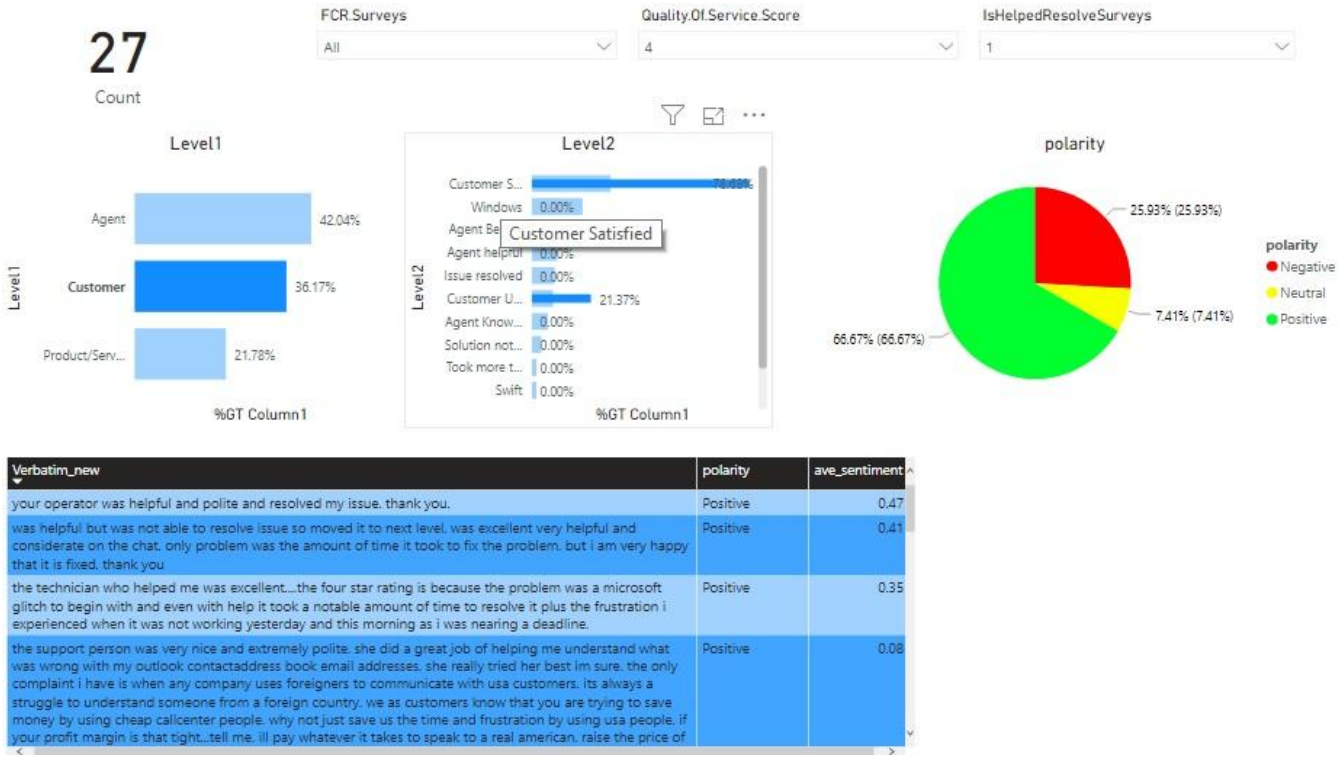
Fig.4 Percentile of CSAT (Customer satisfaction) Score

V. RESULTS

Sentiment analysis of survey data is depicted in fig4. We used a dashboard to display all the required fields. There are 3 dropdowns which are used to select different categories like

A. FCR (First Contact Resolution) survey in which the customer contacted for support and its problem is sorted or not is displayed by value 0(Not resolved) and 1(Resolved) in the first call.

B. Quality of Service Score in the second dropdown which is used to depict the score of quality of the services provided. Customers provide the rating between 1 to 5 based on agent performance where 1 represents "Unsatisfactory", 2 represents "Improvement needed", 3 represents "Meets expectations", 4 represents "Exceed expectations", 5 represents "Exceptional".



Verbatim_new	polarity	ave_sentiment
your operator was helpful and polite and resolved my issue. thank you.	Positive	0.47
was helpful but was not able to resolve issue so moved it to next level. was excellent very helpful and considerate on the chat. only problem was the amount of time it took to fix the problem. but i am very happy that it is fixed. thank you	Positive	0.41
the technician who helped me was excellent....the four star rating is because the problem was a microsoft glitch to begin with and even with help it took a notable amount of time to resolve it plus the frustration i experienced when it was not working yesterday and this morning as i was nearing a deadline.	Positive	0.35
the support person was very nice and extremely polite. she did a great job of helping me understand what was wrong with my outlook contactaddress book email addresses. she really tried her best im sure. the only complaint i have is when any company uses foreigners to communicate with usa customers. its always a struggle to understand someone from a foreign country. we as customers know that you are trying to save money by using cheap callcenter people. why not just save us the time and frustration by using usa people. if your profit margin is that tight...tell me. ill pay whatever it takes to speak to a real american. raise the price of	Positive	0.08

Fig.5 Dashboard of Sentiment Analysis based on Topics

C. Is helped resolved Survey depict yes or no which means the issue is resolved or not.

After selecting the values in the dropdown, we get the counts of records of opinion based on the selection on the left hand side. We can see in Fig 5 the total number of counts is 27 according to our selection criteria.

Now there are two levels, which are Level 1 and Level 2 category. Level 1 classified the important categories, which are used for the analysis of customer survey responses. Similarly, Level 2 classifies the categories which come under the selection of category form Level 1. In this selection, we selected category "Customer"

from Level 1 and thus Level 2 represents the percentage of topics that are based on "Customer" category selection like "Customer Satisfaction" and "Customer Unsatisfied" . Table-I represents the overall sentiment analysis score according to the category selection from Level 1 & Level 2.

Table-I: Percentage of sentiment analysis of customer satisfaction score

Polarity Classification	Customer Satisfaction Score
Positive	66.67%
Negative	25.93%
Neutral	7.41%

VI. CONCLUSION AND FUTURE SCOPE

Nowadays opinion mining/Sentiment Analysis plays an important role in making any decisions related to our daily life. Researchers' important area of interest is Sentiment analysis in NLP.

In this paper, we have seen various steps used to perform sentiment analysis using topic modeling. We did our analysis by topic modeling using R-Language which helps in easy and proper categorization according to different topics categories for survey response data like "Customer", "Agent" and "Product/Service". In results, the Pie chart represents the graphical representation of overall polarity classification done according to the selected categories of "Customer" from Level 1. It overall represents the percentile of "Positive", "Negative "and "Neutral" classification of data according to the category selected from Level1.

In the future, we will work upon various challenges like emoticons, sarcasm etc. those are still unanswered and require efficient work to make Sentiment analysis a successful field.

REFERENCES

1. Grosan C., Abraham A. (2011) Rule-Based Expert Systems. In: Intelligent Systems. Intelligent Systems Reference Library, vol 17. Springer, Berlin, Heidelberg
2. Alexander Pak, Patrick Paroubek, "Twitter as Corpus for Sentiment Analysis", Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010
3. Jingyi Ye1,2, Xiaojun Jing1,Jia Li, "Sentiment Analysis using modified LDA",Signal and Information Processing, Networking and Computers,2018
4. Ali Daud, "Using time topic modeling for semantics-based dynamic research interest finding" ,www.elsevier.com/locate/knosys,2011
5. Toqir A.Rana, Yu-N Cheah, Sukumar Letchmunan, "Topic Modeling in Sentiment Analysis: A Systematic Review," Journal of ICT Research and Applications · June 2016.
6. Prerna Mishra, Ranjana Rajnish, Pankaj Kumar, "Evaluating Performance of Machine Learning Techniques Used in Opinion Mining" 4th International Conference on Computing Communication and Automation (ICCCA),2018
7. Prerna Mishra, Ranjana Rajnish, Pankaj Kumar, "Sentiment Analysis of Twitter Data: Case Study on Digital India," International Conference on Information Technology-IEEE, pp 21,2016
8. David M. Blei, Andrew Y. Ng, Michael I. Jordan, "Latent Dirichlet allocation", Journal of Machine Learning Research 3 (2003) 993-1022

AUTHORS PROFILE

Ms. Prerna Mishra is a Research Scholar, pursuing Ph.D. from Amity Institute of Information Technology, Amity University, Lucknow. She has completed her Masters degree from U.P Technical University. Her area of interest is Opinion Mining/Sentiment Analysis. She has published five research papers in National & International conferences. She has total 7 years of Industry experience as System Analyst in different domains. She also exposes the Onsite experience of handling different projects.

Dr. Ranjana Rajnish is an Assistant Professor at Amity Institute of Information Technology at Amity University, Lucknow. Dr. Ranjana possesses approximately 25 years of experience in academics/research. She has been engaged with institutions like U.P. Technical University and Amity University in roles ranging from a faculty in computer science to Academic Head. Her area of interest includes Software Engineering,Opinion Mining/Sentiment Analysis and Healthcare.

She has several publications in national and international journals and conference proceedings of National and International Conferences of repute. She is also member of various professional bodies like Computer Society of India (CSI), Association of Computing Machinery (ACM), International Association of Engineers (IAENG), Internet Society and Computer Science Teaching Association (CSTA).

Along with being a committed teacher and a passionate researcher, Dr. Ranjana is reviewer for various International Journal and member of editorial board for different International Journals. She is also reviewer, member of technical programme committee in various conferences of repute in and outside India. She has many Ph.D. scholars pursuing Ph.D. under her.

Dr. Pankaj Kumar is currently working as Assistant Professor (Reader) in Department of Computer Science & Engineering in Sri Ramswaroop Group of Professional College, Lucknow. He has more than 18 years of teaching experiences. He received his MCA degree in 2001, M.Tech in 2010 and Ph.D degree in Computer Application in 2011. His Area of Expertise is Parallel Computing/ Mining/Security.

More than 60 research papers of Dr. Pankaj Kumar have been published in various National/International Journals and IEEE/Springer/ACM sponsored Conference proceeding. He is Senior Member of IEEE, Professional Member of ACM and Life member of CSI, IETE, ISTE, IAENG, ISOC and IACSIT. He is member of Management Committee of CSI and IETE Lucknow Chapter.

He is reviewer for various International Journal and member of editorial board for different International Journals. He also participated in various conferences as reviewer, member technical committee, and co-chair. One PhD thesis is awarded and eight students are enrolled as PhD scholar under his guidance. More than 10 students are guided by him in M.Tech Thesis.