Hybrid Sentiment Analyzer for Opinion Mining: Indian Admission Scenario

Priti Jagwani

Abstract: Social media has become one of the widely acclaimed tool for sharing information as well as expressing ideas and emotions. The work depicts the dual aspect task of analyzing and comprehending data available on Twitter platform. This is done using NLP techniques. Using Latent Dirichlet Allocation (LDA) topic technique; the major topics discussed in tweets (of data set taken), have been identified. The input for this Latent Dirichlet Allocation is given by NLP technique – Bag of Words. For further processing, identification of the underlying emotions contained in tweets using the techniques of Sentiment Analysis is done. The result of sentiment analysis is in the polar form. As a case study, a scenario of admissions in India for UG and PG has been considered. The whole process has captured the opinions of stake holders taking part in the admission process. Tweeter data of Indian Institute of Technology (IIT) admission has been used to collect the data in order to conduct the experiment. Major topics discussed in tweets and the fundamental emotions contained are obtained as results along with the polarity of the tweets

Keywords–: Sentiment Analysis, Latent Dirichlet Allocation, Opinion Mining.

I. INTRODUCTION

In the present era of vast usage of internet, masses are living a different life in the online world. This is often called "second life". According to a statistics number of internet users has grown up to 3.01 billion. Around 52.4% of global online population access internet from their mobile device. In this ""second life" of masses which they are leading online, social media particularly is having a dominant role. Out of 4 hours 25 mins which is an average of online time for a user, 2 hours 25 minutes are spent on social media. Social media has witnessed an exponential growth during the last decade. It has become a medium to share views and opinions, as a tool to educate the population about various social and political issues while on the other hand it can be used to share feelings, experiences or adventures of masses with the global community. With this, another life on social media a lot of content has been generated by people. This content is in the form of textual and audio visual data. Among various social networking and microblogging sites, Twitter has become a go-to place for sharing news, opinions and events. With its huge repository of data, twitter has also become a valuable tool for researchers working in various fields whether it is management, arts, science of any interdisciplinary field for that matter. Because of the huge availability of data, the problem of information overload has become a prominent one. In order to extract meaningful content/information from the available data various mining techniques can be used like Topic Modelling, sentiment analysis etc. are used.

Ultimately it's the information derived from content repositories are of use may be because of its commercial use or a personal one. Sentiment analysis as the name indicates deals with finding the sentiment of a given corpus of text. These sentiments can be in binary form like either positive or negative sentiments or may cover a wide range of feelings and emotions. Along with Sentiment Analysis, Topic Modelling is another technique used to infer abstract topics occurring in a collection or a corpus. The work presented here is an extension of the work [10]. The work presented in that paper focused on analysing the opinions of users using the above said techniques namely sentiment analysis and topic modelling. The domain explored in the work was of a typical Undergraduate scenario. In the current effort, the above work is extended and along with UG scenario, tweets about PG admissions (including MTech, PHD etc) are also analysed to get the user opinion. This detailed research will help stake holders to know about the public opinion (separately about UG and PG admissions, as both the scenarios are entirely different), their sentiments, difficulties faced during admission process, also suggestions and suggested points for improvement. Along with covering Post graduate scenario the size of data set has also been extended as compared to the previous work.

This paper is divided into five sections. Along with the line of Introduction in section 1, Section 2 elaborates the related work while section 3 contains background and motivation. Methodology is contained in section 4 followed by results and inferences in section 5. Section 6 contains limitations and future work. Finally, conclusion is elucidated in section 7.

II. RELATED WORK

With the popularity of social media, the field of opinion mining through sentiment analysis techniques have been very popular in one decade. Various researches have been performed under the umbrella of this domain.

In fact, Twitter data had been used for predicting election results, for performing various data mining and natural language processing tasks, gauging user feedback about products and services. Authors in [2] used Twitter data to gauge sentiment around the topic of demonetization in India. [7] Used Twitter data to analyse public opinion about political issues In the paper [5] the importance of sentimental analysis has been highlighted. Sentiment analysis has been used for emotion detection, building resources, transfer learning etc. various approaches to sentiment analysis have also been compared. In the paper [6] authors emphasize the need for automated analysis techniques to extract sentiments and opinions conveyed in the user comments.

Revised Manuscript Received on October 20, 2020.

Dr. Priti Jagwani Assistant Professor Department. of computer science, Aryabhatta College, University of Delhi.

The authors describe Sentiment analysis, also known as opinion mining as the computational study of sentiments and opinions conveyed in natural language for the purpose of decision making. Authors in [3] studied online movie reviews using sentiment analysing approaches. The authors compared three supervised machine learning approaches SVM, Naive Bayes and kNN for Sentiment Classification of Reviews. Empirical results state that SVM approach outperformed the Naive Bayes and kNN approaches, and the training dataset had a large number of reviews, SVM approach reached accuracies of atleast 80%. The work [4] describes the Automatic keyword extraction system for Punjabi language to find words from a document which convey the complete meaning of the text.

III. BACKGROUND

80% of the data available on the internet is unstructured. It's highly unorganized. Huge volumes of text data (emails, support tickets, chats, social media conversations, surveys, articles, documents, etc), as well as other forms of data (audio visual and images) are created every day but it's extremely difficult to analyze, understand, and draw meaning out of it, also it's time-consuming and expensive. Sentiment analysis is the process which helps by making sense of all this unstructured data by automatically tagging it.

In every country, there are some special institutes which are of special importance. India is also not an exemption. A typical Indian admission scenario necessarily includes discussion about getting admissions in the autonomous bodies of National Importance- IIT's (Indian Institute of Technology). Tweets gathered from the Twitter handle during the admission process of one of the IIT's are used for analysis.

In India, every year hundreds of thousands of students want to secure admission into dream schools and their chosen fields, but the whole admission process is complicated and cumbersome. Students face many difficulties; many times there are technical issues due to which the whole process is stalled. Now a days students extensively use social media as a medium to communicate the issues with the institution, public and stake holders which results in generating a huge corpus of text data.

This data can be utilized to draw a clear picture of the issues faced by students during the admissions process. However, because of the huge amount of textual data it is a daunting to assess and enumerate this data which requires expensive human resource. The paper describes the technique to analyse the corpus collected during the admission season and to generate some useful results out of it.

A. Dataset

Various tweets were searched and filtered containing screen name of IIT'S Twitter handle. IIT Twitter handle screen name was used as a search criterion. The data was collected from Twitter using the publicly available Twitter API over a period of one and half year. All the tweets which satisfy the search criteria were extracted and stored in a CSV file. This CSV file also has tweet ID, Name of the author along with timestamp. Approximately 1500+ tweets were collected. Number of tweets is 1500 as admission process is not a whole year long process; it lasts only for 3 months. Further on this data sentiment Analysis and Topic Modelling are applied and the results were used to identify potential issues during the admission process and to identify the sentiments of students and parents alike.

B. Bag of Words (BOW)

The NLP technique used in the work is Bag of Words which is the most elementary technique of NLP. In Bag of Words, unique words are selected from the corpus to make a dictionary. Every document is presented in the form of a binary array. In this array the words that are present in the document are represented as 1 and the words which are not present in the document are represented as 0. Bag of Words is popular because of its simplicity and ease of implementation but it neglects the sequence of words because of which context may get lost. BOW output is serving as the input for LDA as well as for sentiment analysis.

C. Latent Dirichlet Allocation(LDA)

Topic Modelling allows us to find the abstract topics that occur in a collection of documents. It is a frequently used text-mining tool for the discovery of hidden semantic structures in a body of text. Latent Dirichlet Allocation (LDA) is an example of topic model [1]. LDA is a generative statistical model. It views each document as a random mixture of corpus-wide topics and each word is drawn from one of

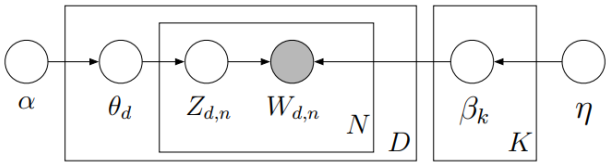

FIGURE 1 THE GRAPHICAL MODEL OF LDA. [SHARMA ET AL, 2019]

these topics. The graphical model of LDA is represented as:

Above is what is known as a plate diagram of an LDA model where:

α is the per-document topic distributions,
β is the per-topic word distribution,
θ is the topic distribution for document m,
φ is the word distribution for topic k,
z is the topic for the n-th word in document m, and
w is the specific word

Plate model shows how the variables are related to each other. α and η are the parameters for prior distributions of θ and β. The per-word topic assignment $Z_{d,n}$ depends upon the per-document topic proportions θ_d. The observed word $W_{d,n}$ depends upon both $Z_{d,n}$ and β_k. $W_{d,n}$ is the only observed variable in the model; rest are latent variables and will be estimated using one of the approximate posterior inference algorithm. Gibbs Sampling has been used to derive distributions of θ_d and β_k. Gibbs Sampling is a Markov chain Monte Carlo algorithm for obtaining a sequence of observations approximated from a specified multivariate probability distribution.

D. Sentiment analysis

Sentiment Analysis is a process of comprehending and categorization of sentiments available within a corpus of the text. It is done using various text analysis tools. It uses various techniques of NLP, text analysis and computational linguistics in order to perform the above tasks. Sentiment analysis can be used either to find polarity i.e. identifying polarity (such as positive, negative and neutral) or it may go beyond polarity till aspect based sentiment analysis. Sentiment analysis uses various Natural Language Processing (NLP) methods and algorithms like rule based, automatic and hybrid. The work uses Sentiment Analysis to analyse the opinion of the public by taking emotions into account which plays a critical role in defining whether the people are in support of or against towards a particular subject.

IV. METHODOLOGY

The whole work Flow is shown in Figure 2. In order to perform dual task of sentiment analysis and topic modelling, data was collected using Twitter API. The work is focused on the textual data contained in various tweets. The collected data needs to be pre-processed in order to make it suitable for further analysis. This textual data is too raw to be worked upon as it contains various text characters and elements like URLs, special symbols like * (asterisk), % (percentage), @ (at-the-rate), # (hash) etc. along with images and emoticons. These all elements are considered as noise in the data. Also, the tweets contain plenty of stop words which should be safely avoided or removed before analysing the text corpus. These stop words don't have any ~~significance~~ significant value for text analysis. List of stop words is easily available on the internet. With the help of this list, stop words occurring in the corpus can be identified and removed. After this, the data will be used for further analysis. In the task of pre-processing data is being cleaned and stop words are removed.

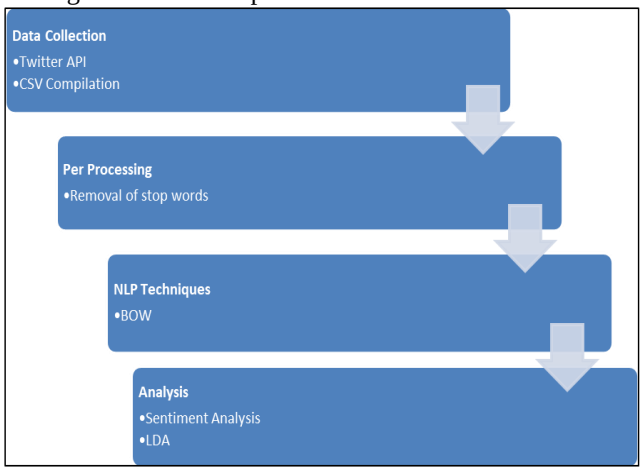

Figure 2 : Complete Workflow of the System

A. Application of BOW and LDA Model

After pre-processing and cleaning the tweets, a Bag of Words (BOW) was generated as explained in the background section. This Bag of Words serves as the input for both LDA technique and Sentiment Analysis. As the next step this BOW data was passed to the Latent Dirichlet Allocation (LDA) model. LDA model on operating upon the bag of words produces the output in the form of word with their respective probabilities. These probabilities indicate the chances of a word to form a topic. Higher the probability higher is the chances for the word to get recognized as a topic in the output list. The LDA implementation was provided by genism [9] which is a popular topic modelling toolkit implemented in Python.

B. Sentiment Analyzer

A hybrid Sentiment Analyzer [8] which combines machine learning techniques with the lexicon-based methods was used to identify the sentiments of the given corpus. The hybrid Sentiment Analyzer uses the NRC Emoticon Lexicon along with the Liu's Opinion Lexicon. As the output of BOW is given to LDA model , similarly that is supplied to sentiment Analysis process also. The work mainly deals with polarity based sentiment analysis which classifies the data in mainly two categories positive or negative. Along with this one more category has been taken into consideration and that is 'neutral

V. RESULTS AND INFERENCES

Through this technique, the topic words and their associated words will be identified as shown in the figure. After applying these techniques on the tweets collected; various topics and associated words have been found [10]. These are shown in the table below.

Table 1 : Topics of UG Data

Topic words	Gate	Joint	Problem	Website
Associated Topics	Server	Madras	Site	Gate
	Application	Yesterday	Gate	Form
	Trying	Website	Server	Students
	Today	Payment	Application	Server
	Site	Problem	Yesterday	Working
	Extend	Facing	Help	Madras

For the underlying case study, four topic words have been identified by the system. Topic 1 reflects that examination being talked about is "GATE" with students trying to access site. Topic 2 lists 'Joint', 'Madras', 'Yesterday', 'Website', 'Payment', 'Problem' which means that perhaps there is a problem being faced by the students which is related to payment. Topic 3 reflects that the problem is with the site and server. It also reveals that the applicants must have used use social-media for help. Topic 4 lists 'website', 'gate', 'form', 'students', 'server', 'working' which indicates (tentatively) that the site was functional. It can be inferred form the topics that the major difficulties were website, payment and server related. Figure 3 shows the word cloud of the topics obtained. While analysing the tweets related to PG admissions the following topics have been identified which are given in the table below along with their associated words.

Table II: Topics of PG Data

Topic words	Admissions	Exam	Research	Eligibility
Associated Topics	Exam	Fee	PhD	Portal
	Entrance	University	Gate	students
	MTech	Diploma	Announcement	Hall ticket
	Apply	NTA	July	Joint
	Online	Competitive	Session	Score
	Website	Hard	Integrated	interview

Word cloud (for a better pictorial representation) have been generated for the tabular results of both UG and PG tweets analysis in order to have a clear and quick idea about major words in the tweets. These are shown in figures 3 and 4 respectively.

Figure 3 : WordCloud of UG data

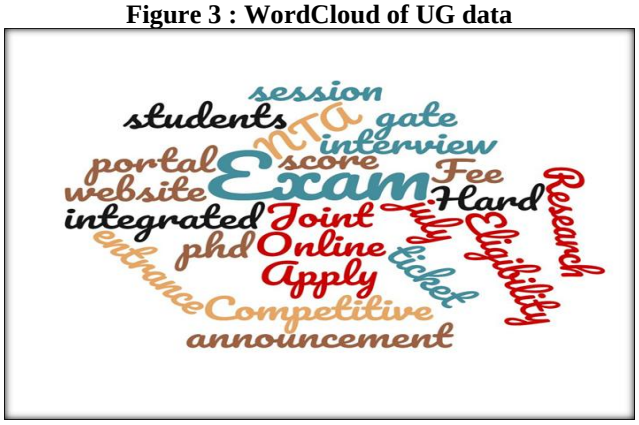

Figure 4: WordCloud of PG data

. Figure 5 shows the results of the sentiment analysis done on the dataset (for UG). It is clear from the figure that there is a mixed sentiment. Overall, 28% of the people expressed positive sentiment, with 25% expressing negative sentiment. A whopping 47% of people showed neutral sentiment. It can be inferred from the sentiments analysis that admission process definitely encompasses a space of improvement further.

In order to, extend the current research and the work tweets containing PG information are also analysed by using the technique sentiment analysis also. It so evident from the results shown in figure 6 that for PG admissions 46% people have shown positive sentiments; 24 % have shown the negative sentiments while rest were neutral. It can be inferred from the sentiments analysis that admission process strongly encompasses a space of further improvement. By carefully observing the results of sentiment analysis, it can be concluded that there is more satisfaction prevailing towards PG admissions as compared to UG admissions.

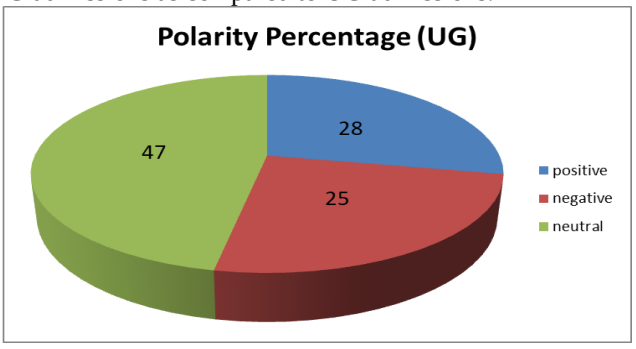

Figure 5: Sentiment Analysis for UG Data

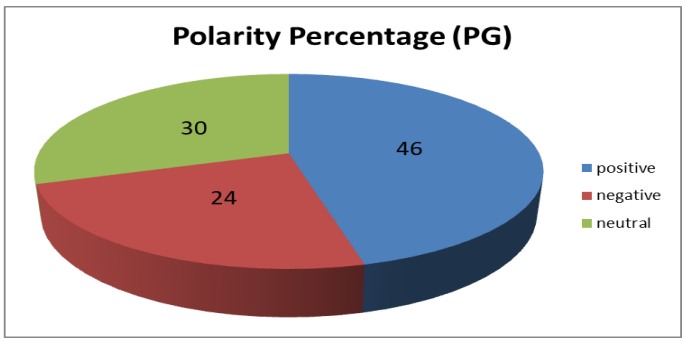

Figure 6: Sentiment Analysis for PG Data

VI. LIMITATIONS AND FUTURE WORK

The method described in the current work is focusing on textual data in tweets. It is not taking into considerations ofemoji and other graphical emotions. Further, these methods work very well with English language tweets with the roman script while many tweets are written in Hindi transcribed in Roman alphabets. Because of this limitation, many tweets cannot be analysed. To overcome this limitation, Deep learning techniques can be explored as solutions. Also rather than just limiting to sentiment analysis in binary form other forms of sentiment analysis owing to different degrees of various sentiments can be analysed.

VII. CONCLUSION

The work is focusing on dual aspect analysis of tweet data obtained from social media (Twitter) in order to analyze the opinion of stake holders. Sentiment analysis and topic modelling are used as major techniques for the above task. Case study of Indian UG and PG admissions has been presented. Topics discussed through tweets and their underlying emotions have been identified which indicates that admission process needs to render the problems present. Also on the basis of results obtained from the case study; it has been clearly identified that there is a larger sense of satisfaction towards PG admission process as compared to UG admissions.

REFERENCES

1. David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
2. Hyui Geon Yoon, Hyungjun Kim, Chang Ouk Kim, and Min Song. Opinion polarity detection in twitter data combining shrinkage regression and topic modeling. *J. Informetrics*, 10(2):634–644, 2016.
3. Kalaivani, P., and K. L. Shunmuganathan. "Sentiment classification of movie reviews by supervised machine learning approaches." *Indian Journal of Computer Science and Engineering* 4, no. 4 (2013): 285-292.
4. Kaur, Kamaldeep, and Vishal Gupta. "Keyword extraction for punjabi language." *Indian Journal of Computer Science and Engineering (IJSCE)* 2, no. 3 (2011): 364-370.
5. Kumar, Praveen, and Umesh Chandra Jaiswal. "A comparative study on sentiment analysis and opinion mining." *Int J Eng Technol* 8, no. 2 (2016): 938-943
6. Mathapati, Savitha, and S. H. Manjula. "Sentiment analysis and opinion mining from social media: A review." *Global Journal of Computer Science and Technology* (2017).

7. Mitodru Niyogi and Asim K. Pal. Discovering conversational topics and emotions associated with demonetization tweets in india. *CoRR*, abs/1711.04115, 2017.
8. Pedro Paulo Balage Filho, Lucas Vinicius Avanço, Maria das Graças Volpe Nunes, and Thiago Alexandre Salgueiro Pardo. NILC_USP: An improved hybrid system for sentiment analysis in twitter messages. In Preslav Nakov and Torsten Zesch, editors, Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 428–432, Dublin, Ireland, 23–24 August 2014. Association for Computational Linguistics and Dublin City University.
9. Radim Rehurek and Petr Sojka. Šoftware Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/-884893/en.
10. SHARMA, YASH, ISHANT MALIK, AND PRITI JAGWANI. "DUAL ASPECT SENTIMENT ANALYSIS FOR OPINION MINING." IN 2019 IEEE INTERNATIONAL CONFERENCE ON ELECTRICAL, COMPUTER AND COMMUNICATION TECHNOLOGIES (ICECCT), PP. 1-4. IEEE, 2019.

AUTHORS PROFILE

Dr. Priti Jagwani is an alumnus of IIT Delhi. She has received her MTech and PhD from IIT Delhi. Currently, she is serving as an assistant professor in the dept. of computer science, Aryabhatta College, University of Delhi. She has served academia for more than 15 years. She has published many papers with Springer, IEEE and Inderscience publishers. She has also successfully completed the Innovation project of the University of Delhi. She is also working as a reviewer for many conferences of repute and presented her research papers in international conferences at Japan and China. She is the recipient of the prestigious Top 50 Women in Education Leaders award (2019) by World Education Congress.