Prediction of N-Gram Language Models Using Sentiment Analysis on E-Learning Reviews

P.Rajesh
Ph.D., Research Scholar,
Department of Computer Science,
Vels Institute of Science Technology
and Advanced Studies (VISTAS),
Chennai.
email: itsrajesh91@gmail.com

G.Suseendran
Assistant Professor,
Department of Computer Science,
Vels Institute of Science Technology
and Advanced Studies (VISTAS),
Chennai.
email: suseendar_1234@yahoo.co.in

***Abstract*- Sentiment Analysis describes the branch of the study of Natural Language Processing that seeks to identify and learn insights from the text or sentences considered to be reviews or opinions about a product or service. These opinions are collected from any platforms like social media, online surveys, online product selling applications, and blogs, etc. The process of sentiment analysis roughly starts by collecting the reviews or opinions, pre-processing of text or sentences, classifying the text to find the polarity whether it is found to be as positive, negative, or neutral. The main objective of this research work is to apply sentiment analysis to the e-learning review dataset. To attain the above-said objective, we predict which n-gram model best suits in feature extraction with machine learning algorithms.**

Keyword- e-learning, sentiment analysis, feature extraction.

I. INTRODUCTION

The application of computers in education has impacted both learners and instructors. The latest cutting edge technologies used across the learning community emerging into lightning speed. The interesting fact is that, one can learn online with various controls over the learning application by getting flexible features throughout the course they learn. To enrich the understanding of any concepts clearly, wide range of animation and graphics been used throughout the course. Traditional classroom methods are getting little bored these days, since students are well attentive towards online classroom environment. To get the clear vision on e-learning that, it always a best technology partner to every student to develop their own skills theoretically as well as practically through learning from the application to acquire depth insights.

Hence to be honest, the laboratory sessions on sciences like physics, chemistry, botany zoology etc., at school level, or graduate level or even for the research level hands on sessions are done only in real-time, but still e-learning medium helps in the situation very well. This can be achieved through watching the lecture online and with the assistive guidance from the teacher; one can perform it in a great manner. Some of the major advantages of e-learning are to learn wherever and whenever you want at your own space and time. The evolution of implementable e-learning systems has been driven by the inclusion of adaptation methods and strategies. In this context, proper guidance in the moment of learning is delivered to every learner. The student models act as storage for the details about a student, thereby providing personalization. The goals of the adopted e-learning system align with the gathered information. The elements of emotion and affection appear to impact the motivation of the student and the results of the learning process. Hence, in the context of learning, it is easy for the instructor to detect and administer information regarding the emotion of the students at a particular time make an impact on understanding the details related to those needs. The information plays a significant role in the mind of the learners because it recommends to them on possible ways to overwhelm the emotional suffering. Conversely, information regarding the emotion of the student towards a course serves as a response to the instructor. This is important in online courses whereby there is less physical contact between the learners and instructors. In this context likelihood of the instructors to obtain any form of response from the student.

Conversely, an adaptive e-learning setting makes use of this. Having an awareness of the emotions of the users is essential in the context of education, politics, eCommerce, and marketing. For the system to be capable of making decisions regarding the information about the users, they need to obtain and store such details. One of the procedures to get information regarding the users constitutes inquiring about whether it is possible to fill the questionnaire.

Fig. 1. E-Learning community

However, users can discover their task to be consuming much time. Lately, non-intrusive approaches are given priority. In the world of e-learning, the parties should understand the significant of obtaining details from the student models without engaging in any form of compromise. From a sentiment analysis point of view, the mood of the authors and their perception regarding a specific entity is

978-1-7281-4097-1/20/$31.00 ©2020 IEEE

demonstrated by written sentiment. It is possible for the prevalent users to remark and put up their thoughts or opinions regarding whatever they wish. The internet is abundant is in expressions as the Web 2.0 technologies. Currently, the sentiment analysis using certain tools has made it easy to come up with inferences about the opinions of masses on many topics. It tends to bring serious technical challenges [1].

The content that is expressed in many languages on the internet by social media users will be instrumental in explaining this statement. The application of the process of text analysis automatically defines sentiment analysis in online education. The text analysis is vital in obtaining views and recognizing the diversity of sentiments that are discussed in the electronic learning forums and blogs. It is important when the students are expressing their perceptions grounded on the present services. The acknowledgement of the patterns during free-text that is shown in emotions is instrumental in the present approaches. These patterns entail character n-grams that are grounded on the n-grams model.

The paper seeks to predict the best fit n-gram model for individuals to find out how people choose their best e-learning providers through their positive opinions and give negative reviews that the e-learning providers should improve by applying the concept of Natural Language Processing.

Fig. 2. Sentiment Analysis Tasks

II. RELATED WORKS

Most of the people confuse e-learning and distance learning because they perceive these two phrases as synonyms. Online learning is revolutionizing the education sector because it has made it a reality for individuals to study in spite of the fact that they have tight schedules. The mode of training in e-learning has continued to be dynamic for a long time. It is likely to define e-learning [2] as a form of education, whereby students work at home and converse with teachers and other learners through e-mail, videoconferencing, electronic forum, bulletin boards alongside other computer-oriented communication means. As the world continues to face economic crisis, the working class have chosen to study online and reach their academic goals at the convenience of their homes. Software and web developers have collaborated to build systems that allow the students to learn easily and hearing well from the instructors. Each institution is attempting to include the online learning model because of the technological trend in the education sector. There is a lot of research in recent years on online education than classroom education. The grouping of certain conditions and characteristics, including flexibility, has facilitated the embracement and acceptance of the learning model in institutions. E-earning has improved the flexibility between tutor and learner [3], interactivity [4] and solid infrastructure to support the system and offer the learners with easy and fast access to the system.

Many perceptions have been raised in the best way to carry out sentiment analysis. Notably, a significant portion of the approaches aimed at grouping the sentiment at a particular degree. The authors present their views that the algebraic sum of the orientation terminologies for grouping the documents is important in establishing the approach to sentiment analysis. The previous researches have shown that the other methods have been established. It is mandatory for manually selected adjectives and computing the details on appears to be applied on the adjectives. The latter is enabled through the frequency on the web counts. It has become easy for most of the academicians to develop the 'sentiment' lexicon starting from this technique. Relying on the strength of its relationship with many insensitive positive words minus the strength of the relationship with negative words, it has become effortless to gauge the semantic orientation of a specified word. The data analysts have continued to use the sentiment lexicon to formulate the polarity score at every text.

The research by Pang on classic topic grouping presented a wealth of information on sentiment analysis. Pang aimed at instilling the audience with the skills on the best way to apply the machine learning algorithms to produce better results. He positioned his study in a moment where sentiment analysis is considered as either positive or negative. The Maximum Entropy, Naïve Bayes, and Support Vector Machine algorithms were instrumental in getting the desired outcomes. Pang collected outcomes of the study that the machine learning algorithms improved the solutions from 71% to 85% taking into account the approach and datasets.

In the reference [6] it understands the significance of applying equivalent classifiers to group the reviews of movies and blogs that feature the films and vehicles. He considers the exceptional characteristics that entail the bigrams, unigrams, adjectives, and unigrams + subjectivity, bigrams. The author is confident that Maximum Entropy together with characteristics of unigrams + subjectivity is instrumental in improving the level of precision when reviewing the e-learning blogs. In terms of ranking, the quickest technique is the Naïve Bayes NB. The [7] explores the polarity of a movie review relies on the characteristics, linkages, and adjectives. It is important to note that the subjective characteristics are instrumental in improving the accuracy level. On the other hand, the results will not be improved when the data analysts filter the features of the objectives. The online reviews of products can be realized through the various machine learning algorithms taking into account the experiments [8]. The results of the experiments have shown that improved performance is possible when the highly ranked n-grams are combined with the discerning classifier [9].

Notably, a new approach has allowed the integration of the fundamental-rule classification. The latter entails the

algorithms of generations, supervised learning vector machines, and rules. The review of the films and products from the MySpace justifies one of the applications of this approach. The findings grounded on the reviews are that a hybrid classification will go a long way in improving the categorization effectiveness. Also, [10] familiarizes the audience that Hidden Markov Models (HMM) and the Support Vector Machines (SVM) are the drivers of the grouping of sentiments. The individual classifiers are overwhelmed by the integrated approaches with different rules based on the findings from MySpace. The [11] makes use of the supervised machine learning algorithms of Support Vector Machines and the feature-grounded N-gram model to the reviews on the internet about the popular travel destinations across the universe. The feature-based N-gram model [13] garners the highest performance, followed by Support Vector Machines model and lastly the Naïve Bayes Method [12]. The level of user analysis and the analysis on user survival [21] taken here as a major part based on the online MOOC discussion forums. Through observing the students' reviews from the forum by analyzing the sentiments on specific courses will help the course trainer to know the pulse of the students and their requirements while learning online. Another work [22] on sentiment analysis uses the feedback of students about their teachers based on the principle of voting, which helped in predicting the teachers' individual skills, their style of teaching. Here the researchers follow an ensemble approach combining five major machine learning algorithms providing accuracy at the maximum level.

III. PROPOSED WORK

This section is essential for familiarizing the audience with the suggested algorithms for the learning-centered sentiment categorization to assist the education stakeholders in understanding the opinions of the students regarding the system.

A. Feature Selection Methods

Feature Selection is a crucial process applied in data mining and statistical methods. Especially dealing with data mining, the advanced root that shoots out from mining called machine learning which takes more privilege on feature selection. To achieve high performance impact by using the data features on machine learning models and selecting proper variables.

1) Information Gain (IG): One of the feature selection methods that scale the lack of predictability leads by fissuring out the dataset with respect to the gain of random variable. Information gain [24] is formulated by the requirement of the word or term can used for classifying information with respect to scale the significance of lexical elements for classification. Calculation of information gain is given below:

$$G\,(D,t) = -\sum_{i=1}^{m} P(\mathrm{C_i}) \log \mathrm{P}\,(C_i) + P\,(t)\,\sum_{i=1}^{m} P(\mathrm{C_i})|\ \mathrm{t}) \log P(\mathrm{C_i})|\ \mathrm{t}) + P\,(t) \sum_{i=1}^{m} P(\mathrm{C_i})|\ \mathrm{t}) \log P(\mathrm{C_i})|\ \mathrm{t}) \quad (1)$$

2) CHI Square (CHI): While dealing with the problems on classification, where categorical input variables occurs, tests on statistical methods can be used to know whether the variable output is either dependent or independent with respect to the input variable.

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \quad (2)$$

3) Mutual Information (MI): It is a scale of calculating any two mutually dependent random variables. MI weighs the measure of information secured by a random variable via another one.

B. N-Gram Language Models

Where there is a major role play of text and words, especially in the case of Natural Language Processing like sentiment analysis, the statistical language models comes with a huge participation to help with the probabilities of word sequences and phrases. It gives the circumstances to differentiate between the phrases and words that are unique phonetics. The three main types of n-gram language models are explained below:

1) Unigram: This model defines the statement of words or sequence of words, presented in the way like separating each term individually to find its own chances of presence in the statements, a paragraph or the whole document.

2) Bigram: Bigram model groups the statement of words or sequence of words, into two, that is starting from the first and the second as a pair and the subsequent second term and its neighbor term as another group and so on.

3) Trigram: Trigram model is nothing but grouping up the words as a three term that is starting from the beginning of the statement first three words as a group and vice versa.

C. Hidden Markov Model (HMM)

The hidden markov model is the model that comes out from the fields of statistical methods. This model is widely applied in various branches of science especially in the case of text and its classifications. It is actually used in the recuperation of sequence of data which has the dependency with other data sequences which are to be predicted.

D. Support Vector Machines (SVM)

The support vector machines are the machine learning models with the supervised terminology and learning algorithms. SVM are used for classification and analysis on regression.

Paving a way [23], machine learning applications utilizes the help of feature selection methods for getting better clarity among the data. This makes sense that performing data classification with the use of finding the better variables to deal with the algorithm becomes very easier. The feature selection methods will be assessed together with the HMM and SVM-oriented hybrid learning. The researcher used the e-learning corpus as the platform for conducting a study. The sentiment analysis acknowledges the perceptions of the learners considered the presentation of the data in a structured format.

The sentiment analysis involves testing the polarity score of the opinions and perceptions of the students. The score garnered regarding the subjectivity represents a particular meaning. Here, the score of subjectivity is diverse. The table shows the methodology of the e-learning platform.

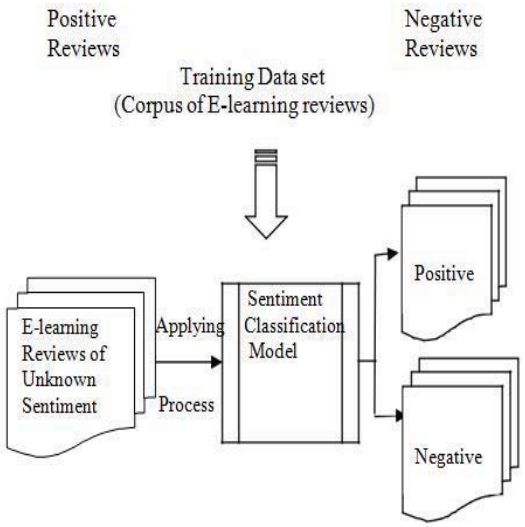

Fig. 3. Process of Sentiment Classification

The tabular columns for results will be presented as follows:

TABLE 1: CONFUSION MATRIX

Original Class	Positive	Negative
Positive	${}^{N}P, P$	${}^{N}P, N$
Negative	${}^{N}N, P$	${}^{N}N, N$

TABLE 2: RESULTS USING THE FEATURE SELECTION METHODS ON OUR HYBRID LEARNING TECHNIQUE

	Positive			Negative		
	P	***R***	***F***	***P***	***R***	***F***
IG	0.791	0.815	0.803	0.809	0.785	0.796
MI	0.772	0.760	0.765	0.763	0.775	0.769
CHI	0.727	0.735	0.731	0.732	0.725	0.728

IV. METHODOLOGY

The public forums including Udemy, Swayam, and nptel (India) have made it a reality for the users to express their perceptions and opinions on many topics. It is interesting how the social media users have fitted themselves into the dynamic social media world. The social media users express their opinions by using slangs, short forms, and vocabulary. It is impossible for a data analyst to conduct a manual analysis of sentiment data.

Preparation of text is important because it involves the sorting of the extracted data on Udemy, nptel, and Swayam. The researcher is cautious at recognizing data and ensuring that the content that is relevant is obtained for the attainment of the aims and objectives of the study. When the data analyst is preparing the data, the former is cautious about subjectivity and explores each sentence and opinion. Here, there is sorting of the phrases depending on their subjectivity.

A sentiment classification is important because it is the phase that allows the data analysts to assign a group to a document from a predefined set of categories. The negative and positive classes formulate the predefined category. The research by [17] demonstrates that a trained statistical classifier engages in the grouping of sentiment. The sentiment orientation of the input documents is predicted by the trained professional. The machine learning algorithms were implemented through the standard bag of characteristics. The goal of this was to group a feature selection method during the categorization of sentiments of e-learning reviews. It aimed at evaluating if the examination of the e-learning blogs can be a challenge for grouping opinions.

The learning algorithms cannot handle the texts directly. This justifies why a preliminary phase regarded as preprocessing is necessitated. Prior to the categorization of the sentiments, it is essential for the text details to be processed. Preprocessing makes it easy to convert the documents into an epitome that is ready to proceed to the next phase. The documents that are relevant to the stemming and feature selection [18] are relatable. Further to this, the test corpus undergoes the equivalent process. From a sentiment analysis standpoint, the stop words are considered as noise words because they are not vital when classifying the opinions. The research by [19] entailed the listing of phrases essential for sentiment classification.

V. CONCLUSION

Evidently, it is fascinating to extract the data from nptel and Udemy and effectively use it well through the application of the sentiment analysis. Currently, the trends demonstrate that the exploration of the forum conversation is going to bring data accuracy. The examination of the sentiments will allow it for the audience to grasp the perceptions of the user regarding the e-learning system.

The combination of the sentiment classification on e-learning is possible based on the study. Even though there are benefits associated with sentiment analysis, it does not fall short of challenges including complexity to understand and this presents a source of losing accuracy. Considerable work has been carried out in the field of sentiment analysis stems from sentiment lexicons.

However, this study is concentrated on offering a comparison between sentiment lexicons for the best to be implemented for sentiment analysis. We validated the sentiment analysis lexicons with algorithms that we have not discovered in an earlier form of research. Due to this, we computed sentiments from analyzers.

We tested their outcomes a realized that they were comparatively better as we extracted better outcomes when assessing the text, as it is demonstrated in the results. Ultimately, the present paper is opening doors for the data scientists to address future research. In the future, the data scientists should combine some of these feature selections, pre-process refinement, and consider misspelled phrases and apply distinct linguistic approaches [20] in the processes of classification.

REFERENCES

[1] "Sentiment Analysis - A Review", International Journal of Science and Research (IJSR), vol. 4, no. 12, pp. 1842-1845, 2015. Available: 10.21275/v4i12.nov152437.
[2] Z. Pozgaj, B. Knezevic, " E-Learning: Survey on Students' Opinions" Information Technology Interfaces, 2007. ITI 2007. pp: 381 – 386
[3] L. C. Seng, T. T. Hok; "Humanizing E-learning" 2003 International Conference on Cyberworlds, 2003; Singapore.
[4] H. Giroire, F. Le Calvez, G. Tisseau; "Benefits of knowledge-based interactive learning environments: A case in combinatorics"; Proceedings of the Sixth International Conference on Advanced Learning Technologies, 2006. pp:285-289.

[5] M. Taboada, J. Brooke, M. Tofiloski, K. Voll and M. Stede, "Lexicon-Based Methods for Sentiment Analysis", Computational Linguistics, vol. 37, no. 2, pp. 267-307, 2011. Available: 10.1162/coli_a_00049.
[6] A. Sheshasaayee and R. Jayanthi, "A Text Mining Approach to Extract Opinions from Unstructured Text", Indian Journal of Science and Technology, vol. 8, no. 36, 2015. Available: 10.17485/ijst/2015/v8i36/88609.
[7] X. Wan, "Bilingual Co-Training for Sentiment Classification of Chinese Product Reviews", Computational Linguistics, vol. 37, no. 3, pp. 587-616, 2011. Available: 10.1162/coli_a_00061.
[8] H. Cui, V. Mittal and M. Datar. 2006. Comparative experiments on sentiment classification for online product reviews. In Proceedings of AAAI-2006.
[9] R. Prabowo and M. Thelwall. 2009. Sentiment analysis: A combined approach. Journal of Informetrics Volume 3, Issue 2, (April 2009), pp: 143-157.
[10] Z. Kechaou, A. Wali, M. Ben Ammar and A. M. Alimi, "Novel Hybrid Method for Sentiment Classification of movie reviews," in The 6th International Conference on Data Mining, 12-15 july 2010 Las Vegas, Nevada USA, pp: 415-421.
[11] Q. Ye, Z. Zhang, R. Law: Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. Expert Syst. Appl. 36(3): 6527-6535 (2009).
[12] H. Binali, V. Potdar, and C. Wu, "A State Of The Art Opinion Mining And Its Application Domains," in Proceedings of ICIT09, 2009.
[13] D. Song; H. Lin; Z. Yang; « Opinion Mining in e-Learning System « 2007 IFIP International Conference on Network and Parallel Computing – Workshops. pp: 788 – 792
[14] M. F . Porter. 1980. An Algorithm for Suffix Stripping Program, vol 14, (1980), pp. 130-137.
[15] Y. Yang, , Pedersen, O. Jan (1997). A comparative study on feature selection in text categorization. ICML, 412–420.
[16] L. Galavotti, F. Sebastiani, M. Simi (2000). Feature selection and negative evidence in automated text categorization. In Proceedings of KDD
[17] "Feature Extraction for Sentiment Classification on Twitter Data", International Journal of Science and Research (IJSR), vol. 5, no. 2, pp. 2183-2189, 2016. Available: 10.21275/v5i2.nov161677.
[18] "Sentiment Classification using Machine Learning Techniques", International Journal of Science and Research (IJSR), vol. 5, no. 4, pp. 819-821, 2016. Available: 10.21275/v5i4.nov162724.
[19] Zheng and G. Li, "Identifying Negative Sentiment with Sentiment Based LDA and Support Vector Machine Classification", International Journal of Control and Automation, vol. 9, no. 9, pp. 331-342, 2016. Available: 10.14257/ijca.2016.9.9.32.
[20] Z. Kechaou, M. Benammar and M. A. Alimi. 2010. A new linguistic approach to sentiment automatic processing. The 9th IEEE International Conference on Cognitive Informatics, 7-9 July, Beijing pp:265-272.
[21] M. Wen, D. Yang, and C. Rosé, "Sentiment Analysis in MOOC Discussion Forums: What does it tell us?," Proc. Educ. Data Min., no. Edm, pp. 1–8, 2014.
[22] J. A. P. Lalata, B. Gerardo, and R. Medina, "A sentiment analysis model for faculty comment evaluation using ensemble machine learning algorithms," ACM Int. Conf. Proceeding Ser., pp. 68–73, 2019.
[23] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," Comput. Electr. Eng., vol. 40, no. 1, pp. 16–28, 2014.
[24] S. Lei, "A feature selection method based on information gain and genetic algorithm," Proc. - 2012 Int. Conf. Comput. Sci. Electron. Eng. ICCSEE 2012, vol. 2, pp. 355–358, 2012.