IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Performance Evaluation Of Naive Bayes And Random Forest For Sentiment Analysis: A Comparative Case Study

Mrs.K.Sravana Kumari, Assistant Professor
Department of Computer Science and Applications
Kakatiya Government College , Hanamkonda
Telangana State , India

Abstract: E-learning, which includes the integration of modern education and information technology, plays an important role in promoting educational approval. As the field of user groups and e-learning applications expands further, it is guaranteed that the quality of online training is becoming increasingly important. Coursera is a remarkable online learning platform, offering a variety of courses and tutorials in many subjects. Working with over 300 major universities and companies, including Stanford, Duke, Google and IBM, Coursera offers lessons in a variety of fields. His offers include Coursera's Google courses to prepare individuals for careers in digital marketing, data analytics, IT support and user experience design. These courses aim to use Google's tools and technologies to acquire the skills that learners need to thrive in advanced industries. The case study focuses on analyzing participant exams for Google courses from Coursera. Our goal is to use natural language processing (NLP) and machine learning technology processing to predict feelings in these reviews as positive or negative. Using these approaches, we would like to gain insight into participants' opinions and experiences in Coursera and Google courses.

Index Terms - E-Leaning , Coursera, Opinion mining, Sentiment analysis, Machine Learning, NLP

1. INTRODUCTION

Sentiment analysis, also known as expanding opinions, is a powerful technique that extracts subjective information from text data and analyzes the emotional tone expressed within it. In an increasingly digital world where a large amount of textual data is generated every day on social media, customer reviews, news articles, and more, mood analysis plays a critical role in understanding public opinion, customer feedback, and market trends. By using algorithms for natural language (NLP) and machine learning processing, mood analysis categorizes texts and provides valuable insights into people's attitudes, emotions and experiences. This transformative technology enables businesses to discover applications in a variety of industries, including marketing, customer service, product development, reputation management, and more, allowing companies to make data control decisions and improve stakeholder understanding in real time. Mood analysis continues to advance, but there is great potential to shape strategies and drive more relevant global innovations in how businesses and individuals interpret and respond to textual data.

Coursera is a leading online learning platform that offers a variety of courses and tutorials on a variety of subjects to serve learners around the world. With partnerships with over 300 top universities and companies, including well-known institutions such as Stanford, Duke, Google, and IBM, Coursera offers high-quality lessons in areas such as technology, business, arts and science. Learners can access courses on request to allow for flexibility in planning and pace. Furthermore, the interactive learning

approach related to Coursera's practical skills and emphasis on real-world applications is a favourable choice for those who want to improve their knowledge and promote their careers.

Participant reviews on courses offered by Coursera offer valuable insights into the effectiveness and impact of the platform's that the educational is offering. These reviews serve as a testament to the quality of instruction, relevance of content and overall learning experience provided by Coursera. Participant reviews play a crucial role in shaping the Coursera community and guiding prospective learners in selecting courses that align with their goals and preferences. As Coursera continues to evolve and expand its course catalog, participant feedback remains instrumental in maintaining the platform's reputation for excellence in online education.

Sentiment analysis by Coursera Google Reviews provides a differentiated understanding of participants' perceptions and experiences using courses on the platform. By using natural language (NLP) and machine learning processing, researchers can extract valuable knowledge from the emotions expressed in these reviews.

Positive sentiment towards Coursera Google courses often highlights the platform's effectiveness in providing practical skills and knowledge relevant to today's digital landscape. Learners may commend the comprehensive curriculum, engaging instructional materials, and the opportunity to learn from industry experts. Positive reviews often emphasize the convenience of online learning, enabling participants to enhance their skills at their own pace and convenience.

Conversely, negative sentiment in Coursera Google reviews may stem from challenges such as course difficulty, inadequate instructor support, or technical issues with the platform. Participants may express frustration with the learning experience, citing factors that hindered their ability to fully engage with the material or achieve their learning objectives. Constructive criticism in negative reviews provides valuable feedback for Coursera to address areas for improvement and enhance the overall quality of its courses.

Finally sentiment analysis of Coursera Google reviews enables researchers to gain insights into the factors influencing participants' perceptions and satisfaction with the platform's courses. By understanding the sentiments expressed in these reviews, Coursera can continue to refine its offerings and deliver high-quality education that meets the diverse needs of learners worldwide.

2. Related work

In recent years there has been considerable research focusing on mood analysis. This is an area aimed at automatically recognizing opinions and feelings expressed in texts. This analytical technique helps distinguish between objective and subjective content and determine whether the text conveys positive or negative moods. Mood analysis has used applications in a variety of fields, including business, politics, and in particular education, to assess feedback from students. For example, the researchers in [3] have developed a tool to assess teacher performance using loss of opinions in relation to mood analysis. Their research allows students to identify the pros and cons of teachers, whether provided in English or Filipinos, based on student feedback. By integrating the loss of opinions technology, they steal valuable findings from student opinions, thereby using data records for two years of teacher review. The classification model used was used by Bayes, comparing individual words in a sentence with words in the database to determine the probability of positive and negative mood names.

[6], the researchers introduced a hybrid approach combining machine learning and lexicon-based for mood analysis of student feedback. The data record is collected from your Institute's Education Portal at the end of the semester and consists of 1230 comments that provide insight into suggestions to improve the quality of education and teaching methods. The proposed model used for mood analysis used TF IDF and lexicon-based features to assess emotions expressed in text feedback. Learning algorithms such as Random Forest and Support Vector Machine (SVM) were used for analysis.

Similarly, for mood analysis of educational data, a model based on deep learning techniques was proposed in [8]. The main focus was to optimize the accuracy and performance of the training dataset to identify the most effective model. In this study, multilayer Perceptron (MLP) and SVM became top performance models.

Despite the growing interest in sentiment analysis, particularly in the realm of E-learning , there remains a scarcity of studies focusing specifically on Coursera data sourced from kaggle websites. This research endeavors to examine participant opinions concerning online course reviews sourced from one of the largest e-learning platforms, Coursera. By analyzing data extracted directly from Coursera's extensive online repository, this study aims to gain comprehensive insights into participants' perspectives on various courses offered on the platform. As Coursera stands as a prominent player in the e-learning landscape, this research endeavors to shed light on the diverse range of opinions and experiences shared by participants, thereby contributing to a deeper understanding of the effectiveness and impact of online education provided by Coursera.

The main goal of this article is to conduct a polarity categorization of Coursera reviews. This sentiment analysis of participant reviews serves as a valuable tool for evaluating the performance and participant satisfaction of online courses with in Coursera.

3. Proposed approaches

The study included classification of checks in positive and negative categories. This is a process that requires several important steps to achieve optimal classification accuracy. These steps included data collection, preprocessing, characterization extraction, classification, and evaluation. Figure 1 shows a methodological diagram explaining the continuous flow of these steps in the research process. Each step was carefully performed to ensure the robustness and effectiveness of the classification models used in the study.

3.1. Data Set:

Coursera reviews collected by Kaggle include 1048576 reviews in the CSV file format, including attributes such as review, reviewer, review date, and course-ID. The next step was preliminary processing. During pre-processing, tribal, tokenization, suspension discussions and lemmatization measures were taken. 75% of the data was used for training and 25% of the data was used for testing data.

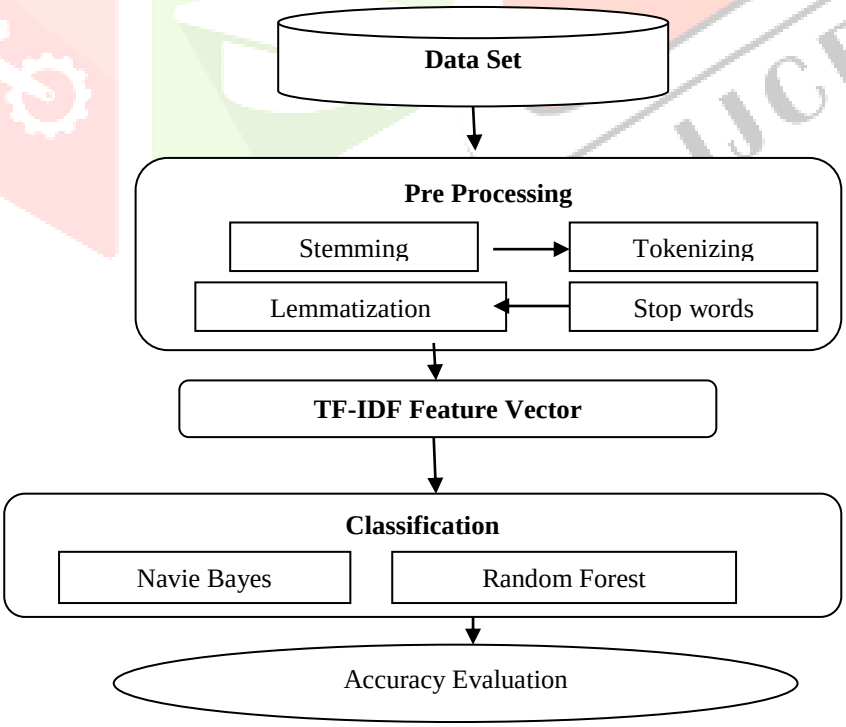

Figure 1. Research method

3.2 Pre processing:

The gathered review data were not well structured yet, therefore, some pre-processing steps were required to make them well structured. Structured data will simplify the classification process. Pre-processing was conducted through some steps, including stop words removal, stemming, tokenization, Lemmatization.

3.3 Feature Extraction:

Feature extraction is a critical process aimed at separating critical features from data records and focusing on people with important relevance. In this study, distinctive extraction was done by wording. Each word in the document has a weight assigned. This is a process executed using the TF IDF method (document frequency during frequency). TF-IDF is a broad technique due to its simplicity and effectiveness in generating more accurate results due to its high callback rate. The TF IDF value for each word in the data record is calculated as achieved in equation (1).

TF-IDF weighting not only emphasizes the frequency of terms (term frequency) within a document, but also takes into account its importance (reverse document frequency) across the
body. This approach helps to emphasize words that not only frequently distinguish, but also distinguish in the context of the entire data record.

$$tf - idf_{t,d} = tf_{t,d} X idf_t \qquad \text{Equation (1)}$$

In the TF-IDF equation, $\text{tf-idf}_{t,d}$ represents the TF-IDF value of term *t* within document *d*. The term frequency within document *d* for term *t* ($\text{tf}_{t,d}$) is utilized in this calculation. Meanwhile, the inverse document frequency (idf_t) for term *t* is also incorporated. The idf_t value is determined using Equation (2).

$$idf_t = \log(\frac{N}{df_t}) \qquad \text{Equation (2)}$$

This equation outlines the calculation process for the inverse document frequency (idf_t) for term *t*. It assists in quantifying the significance of a term across the entire document corpus, thus contributing to the determination of its TF-IDF value within a specific document.

N is the number of all documents. In the meantime, DFT is the number of documents using concept T. Word weighting is used using Python Library, Scikit-Learn [11]. It is weight after each word.

3.4 Text Classification:

Following feature extraction, the research proceeded with text classification to determine the sentiment of each document. This classification task employed a machine learning approach, utilizing Naive Bayes (NB) and Random forest algorithms. The implementation of the classification process was facilitated using the Scikit-learn library in Python.

3.4.1. Naïve Bayesian:

Bayes' theorem, rooted in probability theory and statistical principles, forms the basis for decision-making in classification tasks, considering both probability and associated costs [12]. Naive Bayes (NB) serves as a straightforward application of Bayes' theorem, offering a simple yet effective method for classification. The formula representing Naive Bayes is expressed in equation 3.

$$P\left(\frac{B}{A}\right) = \frac{P\left(\frac{A}{B}\right)P(B)}{P(A)} \qquad \text{Equation (3)}$$

Where P(B/A) is the probability of B appearance when A is known? The value P(A/B) is the probability of an appearance if B is known. P(A) is the probability of an appearance, while P(B) is the probability of B appearance.

3.4.2. Random Forest

Random Forest is an ensemble learning algorithm that combines predictions from several individual models, improving overall performance and generalization. It works as follows:

For data record DS, the classifier first generates k-boat trap samples from the DS. Each Ds_i,It contains the same number of tuples as the DS and includes replacement scans. Replacement sampling means that some tuples from the original DS may not be in Ds_i. It can happen several times by others, but it does exist. Classifiers build decision trees for everyone. As a result, a "Forest" containing the k decision tree is

set. Each tree contributes to the classification of unknown tuples X, with each prediction counting as speech. The final allocation of classes in X is determined by the class with the most votes.

The evaluation process involves testing the created system to assess its accuracy in classifying documents. Accuracy serves as a metric to determine the system's capability in correctly classifying documents. The calculation of accuracy is performed using Equation (4).

$$A = \frac{(tp+tn)}{(tp+fp+tn+fn} \qquad \text{Equation} \quad (4)$$

TP is the number of documents of a positive class that are correctly classified by the system. TN is the number of documents in a negative class, and is correctly classified by the system. The number of positive classic documents that are misclassified by the system is stored in the FN variable. The number of negative categories that are classified by the system is stored in the FP variable.

3.6. Results:

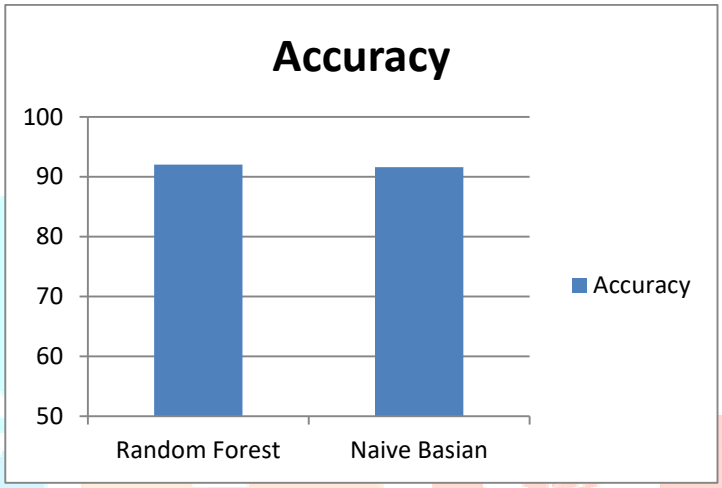

Figure 2: Accuracy comparison between Naïve bayes and Random Forest

The number of reviews in test data is 262144 reviews. The random forest algorithm had produced 6854 reviews as negative and 234513 reviews as positive. Where as Navie bayes algorithm had produced 7963 reviews as negative and 232265 reviews as positive. By the observation of contingency matrix of above algorithms , 20777 reviews were misclassified by random forest and 21916 reviews were misclassified by Navie basian. After accuracy calculations Random forest has given 92.07 % comparing to Navie basian that is 91.63%. Overall performance of both algorithms presented in figure 2:

The Random Forest model exhibits better performance than the Naive Bayes model. Generally Naïve bayes achieve lower accuracy than the Random Forest model across our vector sets and Random Forest surpasses Naive Bayes. This advantage can be attributed to Random Forest's proficiency in accurately classifying both positive and negative categories due to its auto pruning at all levels.

4. Conclusion

In this paper, we plan to classify the text using random forests and Naive Bayes to examine the sense of Coursera assessment. Mood analysis of eLearning Reviews provides clearer photos of actual online courses. This analysis can improve the accuracy of the decision-making process in future online course strategies. The results of the experiment show that random forests achieve good results. Characteristic extraction is performed in TF-IDF. This experiment is carried out at 75% on train data and 25% on test data. This experiment shows that random forests offer higher accuracy than Naive Bayes. Random Forest has an accuracy of 92.07% and a Naive Bayes 91.63%. When compared to both misconception tests, this was more than the number of positive provisions than the number of negative reviews of the training dataset.

References:

[1] Indonesia Internet Service Provider Association (APJII), "Penetrasi dan Perilaku Pengguna Internet Indonesia," 2016. [Online]. Available: https://apjii.or.id/content/read/39/264/Survei-Internet-APJII-2016. [Accessed: 07-Aug-2017].

[2] D. V. N. Devi, C. K. Kumar, and S. Prasad, "A Feature Based Approach for Sentiment Analysis by Using Support Vector Machine," in *6th International Conference on Advanced Computing*, 2016, pp. 3–8.

[3] B. Pang, and L. Lee, "Opinion mining and sentimentanalysis", *Foundations and Trends in Information Retrieval*,2(1-2): pp. 1-13, 2008.

[4] F. F. Balahadia, M. C. G. Fernando and I. C. Juanatas,"Teacher's performance evaluation tool using opinion mining,with sentiment analysis", in *Proceedings of the IEEE Region10 Symposium (TENSYM*P), Bali, 2016, pp. 95-98.

[5] M. H. Saragih and A. S. Girsang, "Sentiment Analysis of Customer Engagement on Social Media in Transport Online," in *International Conference on Sustainable Information Engineering and Technology (SIET)*, 2017, pp. 24–29.

[6] D.D. Dsouza, D.P.N Deepika, E.J Machado, and N.D. Adesh, "Sentimental Analysis of Student Feedback using Machine Learning Techniques", *International Journal of RecentTechnology and Engineering*, 8(14), pp. 986-991, 2019.

[7] D. A. Muthia, "Analisis Sentimen Pada Review Restoran dengan Teks Bahasa Indonesia Menggunakan Algoritma Naive Bayes," *J. IlmuPengetah. dan Teknol. Komput.*, vol. 2, No.2, pp. 39–45, 2017.

[8] J. Sultana, N. Sultana, K. Yadav, and F. Al-Fayez, "Prediction of Sentiment Analysis on Educational Data based on Deep Learning Approach", in *Proceedings of the 21st Saudi Computer Society National Computer Conference (NCC)*, pp. 1-5, 2018.

[9] Zhang Y, Xiang X, Yin C, Shang L. (2013). Parallel sentiment polarity classification method with substring feature reduction. In: Trends and Applications in Knowledge Discovery and Data Mining, volume 7867 of Lecture Notes in Computer Science. Springer Berlin Heidelberg, Heidelberg, Germany. pp 121–132

[10] Pang B, Lee L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics, ACL '04. Association for Computational Linguistics, Stroudsburg, PA, USA

[11] F. Pedregosa *et.al.*, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[12] B. Santosa, *Data Mining: Teknik Pemanfaatan Basis Data untuk Keperluan Bisnis*, 1st ed. Yogyakarta: Graha Ilmu, 2007.

[13] Pang B, Lee L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics, ACL '04. Association for Computational Linguistics, Stroudsburg, PA, USA.

[14] Fisnik Dalipi , Katerina Zdravkova and Fredrik Ahlgren (2021) Sentiment Analysis of students' Feedback in MOOCs: A Systematic Literature Review published in frontiers in artificial Intelligent .

[15] Zenun Kastrati, Blend Arifaj and Arianit Lubishtani (2020) Aspect-Based Opinion Mining of Students' Reviews on Online Courses published on ICCAI 2020 - ttps://doi.org/10.1145/3404555.3404633.