IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 12, Issue 1, January 2023

DOI: 10.17148/IJARCCE.2023.12113

Classifying Social Media Comments Using MachineLearning

Rajath S[1], Swarna N T[2], Arpitha M[3], Prathima K P[4], Dr. Chethan Chandra S Basavaraddi[5]

Prof. Shashidhara M S[6], Prof. Sapna S Basavaraddi[7]

B.E, Dept. of CSE, Kalpataru Institute of Technology Tiptur, India[1-4]

Associate Professor, Dept. of CSE, Kalpataru Institute of Technology Tiptur, India[5]

Associate Professor and HOD, Dept. of CSE,Kalpataru Institute of Technology Tiptur, India[6]

Assistant Professor, Dept. of CSE,Kalpataru Institute of TechnologyTiptur, India[7]

Abstract: The demand to reach and satisfy audiences world- wide increases the number of influencers and content creators onsocial media, which is the primary platform to disseminate their work. Each video could potentially get thousands of commentsas a content creator grows, and these comments acts as direct feedback from the viewers, also as major means of understandingviewer expectations and improving channel engagement. We haveproposed approach to classify social media comments into five categories namely good, discussion, motivational, demotivating and abusive. In this paper we have elaborated comparative analysis between the available machine learning classificationalgorithm like Logistic Regression, SGD Classifier, and Random Forest.

Index Terms: Machine Learning, Natural language Process- ing, Logistic Regression, SGD Classifier, Random Forest

I. INTRODUCTION

In recent years, content creators have grown in prominence on social media platforms. A large number of content creators upload their content in the form of videos or photos on this platform. These contents get thousands of views and comments. The content creators need to continuously workon maintaining the quality and quantity of their contents. Todo so, they must collect feedback from their viewers through the comments section. This feedback lets them understand the influence of their creations. In addition to improving audience engagement, feedback also provides information on the aspectsof the content that need improvement.

However, not all content creators have the time to read through all of the comments on each video. On the contrary, they must read all of the comments in order to completely understand the public's interest in their material. Our study addresses the answer to this annoyance. We use the technique of extracting all comments from a film and categorising them into numerous categories based on both sentiment and sen- tence type: Good, Discussion, Motivational, Demotivational, and Irrelevant or Abuse. These categories might assist content authors in focusing solely on comments that are relevant to their interests.

There have been multiple studies in the field of sentiment analysis such as Twitter sentiment analysis [1], YouTube polarity trend analysis [2], user comment sentiment analysison YouTube [3], and so on. However, not enough research hasbeen carried out on sentiment analysis through classificationof a sentence based on its type. We have approached thisissue from the perspective of YouTube comments. Conse-quently, it is a challenging task to categorize the comments into different sentence types because of various factors suchas non-standard language, spelling errors, unformatted texts, and trivial comments. Apart from these, sometimes there are multiple sentences of different classes on a single comment. The combination of these issues poses a unique challenge in sentiment analysis based on sentence types.

Our approach is to extract features from preprocessed data and then train those features using well-known supervised learning algorithms. Since the model's performance is de-pendent on the text corpus, we choose and compare several common fine-tuned methods for this purpose. To acquire the best results for each model, we tested our comments data-set with three distinct fine-tuned classification models utilisingfeature extraction approaches. The accuracy of the models is calculated using the cross-validation score and the F1 score. Despite the fact that our technique is basic, the results are effective, allowing content producers to readily examine their comments of interest.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 12, Issue 1, January 2023

DOI: 10.17148/IJARCCE.2023.12113

II. RELATED WORKS

There is a research scope for augmenting the methods to classify social media comments. Three-stage Naive Bayes classifier is employed to detect "flame" stated by Razavi [4]. By training the dataset of 1153 Uesnet comments the test accuracy turned out to be 97%.Warner [5] detected hate speech using N-gram SVM classi-fier. The F1 score turned out to be 0.63 and unigram features are the most suited to their Yahoo dataset. Since hate speech and insults are different, the models exhibited similar F1.

Most of the literatures use similar linear models and bagof N-grams [6] [7] [8] [9] [10] [11] [12] [13]. In fact, onlyin the past two years have new papers come out which use other models. These papers have both come from Yahoo, and they use Paragraph Vector [14]. The papers are Nobata [10] and Djuric [15]. These papers provide useful comparisons andstarting points for this project.

In most of the existing work, the sentences of the imper- ative class have not been researched adequately. Khoo [16] performed experiments on different models for 14 different classes of sentences (including imperative sentence types like request, instruction and suggestions). The models used for the experiment were Naive Bayes, Decision Tree, and Support Vector Machine. Support Vector Machine overshadowed all other models and had insignificant effect from feature selec-tion. In their work, they only chose the standard responseemails because these emails have well-structured sentences and few grammatical errors. It eases the classification task. However, our work consists of a large number of unstructured sentences with huge grammatical errors.

III. METHODS

The models have been developed using python language. The various python packages used for analysis are numpy, pandas, sklearn, NLTK (Natural Language Tool Kit) and matplotlib. Our experimental set up consisted of the below mentioned steps:

- Data Collection
- Data Pre-processing
- Converting text into features
- Defining Model

A. Data Collection

The data set has been procured from Kaggle. Most notable of these being a Toxic Comment Classification Challenge [17], and a Kaggle insult detection dataset [18]. The dataset consistsof 4589 comments, we then manually label the comments into5 different classes: Good, Discussion, Motivational, Demoti- vational, and Irrelevant or Abuse. These classes are defined based on general needs of the content creators. Note that further categories can be established if or as needed. These classes belong to two broader classes: Sentiment Analysis(good, Motivational, and Demotivational) and Sentence Types(Discussion, Imperative, Corrective and Miscellaneous). TableI shows different classes and the content of that class. The classes are explained in more detail next.

Good tells that the viewers perceived the content as wor-thy and that the content created has a positive impact on them. Demotivational provides information on what is wrong with the content and why the viewers are not attracted to it. Discussion conveys viewer's doubts and questions. It isa useful feature because the content creators can increase

TABLE I
CLASSES OF COMMENTS WITH CONTENT TYPE

Class	Content
Good	appraisals, positive
Discussion	all type of questions, queries
Motivational	appreciations
Demotivational	scoldings and negative
Irrelevant or Abuse	Abusive, promotions, chitchat

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 12, Issue 1, January 2023

DOI: 10.17148/IJARCCE.2023.12113

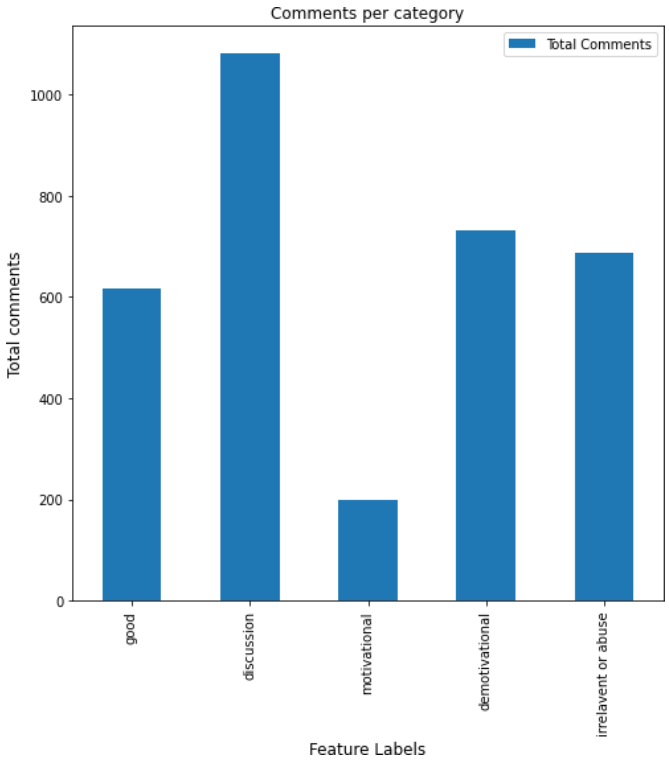

Fig. 1. Number of comments in each class

their influence by addressing viewer's questions and issues. Motivational provides viewer's expectations and requests for actions. Irrelevant or abuse includes declarative sentences and all other trivial comments.

As mentioned previously, some of the comments can belongto more than one class. For instance, the first sentence in"Your solution is not practical. Can you suggest another one?",suggests the Demotivational class while the second sentence suggests the Discussion class. In such situations, we classifiedthe comment based on the importance to the content creator. Inthis example, we assumed it as an Discussion sentence becauseit is more important to answer the question to increase oddsof the viewer to return and stay engaged in the content of the channel. Fig. 1 shows the visual presentation of the quantityof comments in each class.

B. Data Pre-processing

It is important to clean the data and have them in appropriateformat to improve classification. The data pre-processing step handles the following factors that make the classification process difficult:

- Non-standard language: The texts used in the comments section do not always employ standard English. Comments often contain slangs and improper form of words, making it difficult to extract features from them.
- Unformatted texts: These refer to comments containing computer codes. These do not contribute to the feature extraction accuracy; rather, they add unnecessary load to the feature matrix.
- Trivial comments: Not all the comments posted wereabout the video or related to the channel. A large number of viewers comment in order to market their productsor just to show their presence. These comments are not useful to the content creators and only add unnecessary overhead.

Above issues are common in platforms like YouTube because of the informal nature of communication. We addressed these issues using the following pre-processing steps:

- lowercasing
- removing URLs
- removing integers
- removing punctuation
- lemmatizing
- removing stopwords

DOI: 10.17148/IJARCCE.2023.12113

Given the nature of this study, lowercasing was relevant because the same word would have been identified as a different feature had some letters been capitalized (for example, ”Love”and ”love” are two different words for a computer). The removal of URLs, punctuation, and integers was performed because they did not provide useful information for feature extraction; rather adding unnecessary complexity to the model.

We used Lemmatization to analyze the words morphologically and group the similar words together. Furthermore, there are certain frequent words that do not add significant meaning to the sentence such as ”is”, ”are”, and ”it”. They were removed from the corpus. However, stopwords were not removed from all the classes of comments because stopwords for one category might be important for another category. For example, ”not” and ”no” are important for the negative class whereas they are not important for other classes. Stopwords were used from NLTK English corpus, which consists of 179 stopwords.

C. Converting text into features

The well-known techniques for vectorizing a corpus of text include document frequency, tf-idf vectorizer, hashing vectorizer, and Word2Vec. We selected document frequency vectorizer and tf-idf vectorizer for this paper. Using these two methods we can study the behaviour of different classification models under two different conditions. Document frequency (df) vectorizer gives importance to the term that has higher frequency in the document; whereas, tf-idf can incorporatethe terms that are rarely present in the document. Unlike hashing vectorizer, we can examine the text features which are important to the model using the vectors generated by df and tf-idf. For rare or out of vocabulary terms (which mightbe important to a model), Word2Vec can not create an ideal vector for them and it is difficult to interpret those vectors because of hidden layers.

When calculating document frequency (Eqn. 1), if the sameterm is present multiple times in a comment, then its additional counts are not considered. Also, the terms that appear in less than or equal to 5 comments are ignored because they donot add value to the features. In the same way, if any term appears in the majority of the comments, it does not add valueto the feature because it is not the distinguishable feature for a class. These terms are likely already filtered by the stopwords removal process. However, we ensure that only terms that significantly add value to their comment’s class are considered.

$$df = \frac{n}{N} \qquad (1)$$

where, df denotes document frequency, n denotes number of documents in which the term appears, and N denotes total number of documents where document means comment.

After above steps, 2210 terms (features) were derived and scaled from 0 to 1 using a min-max scalar (normalization). We performed this because some of the machine learning models cannot handle large ranges of data. Doing so also helps in speeding up some of the calculations.

$$MinMaxScaler(x) = \frac{x - x_{min}}{x_{max} - x_{min}} \qquad (2)$$

where, x is observed value, x_{min} is the minimum value of that class and x_{max} is the maximum value of that class. The second feature extraction technique used in this paperis $tf - idf$ (term frequency - inverse document frequency). It not only considers the frequent terms, but also the rare terms.

$$tf - idf = tf * log\frac{1}{df} \qquad (3)$$

where, tf is the term frequency and df is the document frequency For $td - idf$, we got 4304 features when both unigram and bigram were taken into account.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 12, Issue 1, January 2023

DOI: 10.17148/IJARCCE.2023.12113

D. Defining And Training Model

The existing available models like Logistic regression, stochastic gradient descents, and a Random forest are tai-lored for our dataset in python. Accuracy and 10-fold cross- validation Acts as measurable criteria to measure primary evaluation metrics. The feasibility of training accuracy is checked with testing data.

IV. RESULTS AND DISCUSSIONS

Precision, Recall, Accuracy, and AUC are considered as theevaluation criteria after training the model, and it is com-pared with logistic regression, SGD Classifier, and Random Forest.

Table II shows the Precision, Recall, F1-score of the selected machine learning classification Algorithms. Based onthe experiment Table III shows the ROC AUC Score of the algorithms on test data.

TABLE II
COMPARISON OF ALGORITHMS

Algorithm	***Precision***	***Recall***	***F1-score***
Logistic regression metrics	65	60	62
SGD metrics	68	46	55
Random Forest metrics	69	47	56

TABLE III
ROC AUC SCORE TEST

Algorithm	***ROC AUC Score Test***
Logistic regression	69.78
SGD	63.53
Random Forest	63.75

The experimental result show in the Table III suggest logistic regression with result:

- High AUC Score - 69.78%
- Balance Precision / Recall values : 65%-60%.

V. CONCLUSION

The evaluation of the results suggested that logistic re-gression was more robust for the data set compared to SGD and Random Forest. With the successful classification of the comments into their respective categories, a content creator can easily access each category of comment.

This can helpthe content creators to avoid scrolling through hundreds of comments and filtering them manually for each video. Pre- vious researchers focused either on sentiment analysis or classification of sentence of a niche, we have incorporated both the aspects. In this paper, we classified the comments using 3 different models on feature selection method. Theexperiments showed that best scores for cross validation and F1 were obtained by Logistic Regression.

VI. FUTURE WORK

In future work, the number of classes and sub-classes canbe increased to represent a more comprehensive comment classification. Likewise, the classification models and overall feature selection approach can be further improved for the comments that belong to more than one class. Performing a comparative study with Explainable Neural Networks (xNNs) for comments classification can be carried out.

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 12, Issue 1, January 2023

DOI: 10.17148/IJARCCE.2023.12113

REFERENCES

[1] Agarwal, A.; Xie, B.; Vovsha, I.; Rambow, O.; and Passon- neau, R.J. 2011. Sentiment analysis of twitter data. In Pro- ceedings of theworkshop on language in social media (LSM 2011), 30–38.

[2] Krishna, A.; Zambreno, J.; and Krishnan, S. 2013. Polarity trend analysis of public sentiment on YouTube. In Proceed- ings of the 19th international conference on management of data, 125–128.

[3] Aung, K. Z.; and Myo, N. N. 2017. Sentiment analysis of students' comment using lexicon based approach. In 2017 IEEE/ACIS 16th international conference on computer and information science (ICIS), 149–154. IEEE.

[4] Farzindar, A.; and Keselj, V. 2010. Advances in Artificial Intelligence: 23rd Canadian Conference on Artificial Intelli- gence, Canadian AI2010, Ottawa, Canada, May 31-June 2, 2010, Proceedings, volume 6085. Springer.

[5] Warner, W.; and Hirschberg, J. 2012. Detecting hate speech on the worldwide web. In Proceedings of the second work- shop on language in socialmedia, 19–26.

[6] Goyal, P.; and Kalra, G. S. 2013. Peer-to-peer insult detection in online communities. IITK Unpubl.

[7] Reynolds, K.; Kontostathis, A.; and Edwards, L. 2011. Using machine learning to detect cyberbullying. In 2011 10th Inter- national Confer- ence on Machine learning and applications and workshops, volume 2, 241–244. IEEE.

[8] Kontostathis, A.; Reynolds, K.; Garron, A.; and Edwards, L. 2013. Detecting cyberbullying: query terms and techniques. In Proceedingsof the 5th annual acm web science conference, 195–204.

[9] Nobata, C.; Tetreault, J.; Thomas, A.; Mehdad, Y.; and Chang, Y. 2016. Abusive language detection in online user content. In Proceedings ofthe 25th international conference on world wide web, 145–153.

[10] D. Bhimani, R. Bheda, F. Dharamshi, D. Nikumbh and P. Ab- hyankar, "Identification of Hate Speech using Natural Language Processing and Machine Learning," 2021 2nd Global Conferencefor Advancement in Technology (GCAT), 2021, pp. 1-4, doi: 10.1109/GCAT52182.2021.9587652.

[11] Dhaoui, C., Webster, C.M. and Tan, L.P., 2017. Social media sentiment analysis: lexicon versus machine learning. Journal of Consumer Mar- keting.

[12] Chakrabarty, N., 2020. A machine learning approach to comment toxi- city classification. In Computational intelligence in pattern recognition (pp. 183-193). Springer, Singapore.

[13] Salminen, J., Almerekhi, H., Milenković, M., Jung, S.G., An, J., Kwak,H. and Jansen, B.J., 2018, June. Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In Twelfth International AAAI Conference on Web and Social Media.

[14] Le, Q.; and Mikolov, T. 2014. Distributed representations of sen- tences and documents. In International conference on machine learning, 1188–1196. PMLR.

[15] Djuric, N.; Zhou, J.; Morris, R.; Grbovic, M.; Radosavljevic, V.; and Bhamidipati, N. 2015. Hate speech detection with comment embeddings. In Proceedings of the 24th interna- tional conference on world wide web, 29–30.

[16] Khoo, A.; Marom, Y.; and Albrecht, D. 2006. Experiments with sentenceclassification. In Proceedings of the Aus- tralasian Language TechnologyWorkshop 2006, 18–25.

[17] 2018. Toxic Comment Classification Challenge.

[18] 2012. Detecting Insults in Social Commentary.