Performance Analysis of Logistic Regression Using Term Frequency-Inverse Document Frequency For Sentiment Analysis

1st Khin Than Nyunt
Department of Computer Engineering and Information Technology
Naypyitaw Technological University
Nay Pyi Taw, Myanmar
khinthannyunt.ktn@gmail.com

2nd Naw Thiri Wai Khin
Department of Information Science
University of Technology (Yatanarpon Cyber City)
Pyin Oo Lwin, Myanmar
ntrwk87@gmail.com

***Abstract*— This study is a sentiment analysis of applying feature extraction method such as Term Frequency-Inverse Document Frequency (TF-IDF) and logistic Regression techniques on YouTube Video comments data of US. Although, movie makers may be able to predict how the public will respond to their once it has been published with the help of the comments that appear on YouTube, not exactly. Due to privacy concerns, YouTube has since December 14, 2021, stopped making some content statistics count such as the number of dislikes, publicly available. Indenting to assist them, we split movie comments into four popular movie genres: Entertainment, Business, Technology, Medical. Figure out the machine learning model that performs sentiment analysis the most effectively is an ongoing challenge. Each machine learning algorithm's performance scores have been calculated. The accuracy of the six machine learning algorithms – Logistic Regression, Naïve Bayes, Light GBM, K-Nearest Neighbors, Random Forest, and XGBRF was 91.01%, 90.16%, 70.23%, 63.18%, 52.58%, and 60.07%, respectively. The most accurate and well-fitted algorithm for predicting YouTube video comments in the US is the Logistic Regression model. Therefore, this system implements the TF-IDF feature extraction method with the combination of logistic regression method to explore the model performance.**

Keywords—Logistic Regression, Sentiment Analysis, TF-IDF, YouTube Video Comments

I. RESEARCH METHODOLOGY

The research methodology section mainly includes the following steps.

A. Data Collection and Preparation

The dataset YouTube Videos Comments US is obtained from Kaggle. It includes 65535 rows and 4 columns. These are divided into four categories with category Encoded 0 to 3 respectively: Entertainment, Business, Technology, and Medical. The data must be translated using google translate afterwards because some comments are not in English language.

B. Data Preprocessing

Data preprocessing was done after the previous stage. The data preprocessing stage aims to reduce noisy data from comments by performing data cleaning, word tokenizing, lowercasing, stemming, lemmatization, punctuation mark removal, stop word removal, and normalization. Using the automatic labelling tool Valence Aware Dictionary and Sentiment Reasoner (VADER) for the three classes: positive, negative, and neutral. This stage's last step involves labelling all of the US comment data.

C. TF-IDF

The term weighting process, also known as word weighting, is a technique used to express the significance of words in numerical numbers. It determines the contribution of a term to a document's class (Positive, Negative, and Neutral). The weighting value becomes more relevant for the classifier when assessing a term on a dataset with positive, negative, and neutral labels. In this study, the pre-processed data is converted into a vector of terms and count of terms using Scikit-learn's CountVectorizer. The TF-IDF score frequency assigns a word's importance based on its frequency. The Term Frequency (TF) of a term is the ratio of the number of times the term appears to the total number of words in the document. IDF is calculated the Logarithm of the number of the document in the dataset is divided by the number of documents in the dataset contain the term. TF-IDF is created by multiplying TF and IDF[3]. The values will be low for words that appear frequently.

D. Logistic Regression

After being vectorized, the data is divided into train and test sets before being classified using the Logistic Regression model. A supervised machine learning method for classification problems is logistic regression. Assisted machine learning algorithms are trained on labelled datasets and accuracy is measured using an answer key. The model's objective is to discover and approximation of a mapping function from input variables x1, x2, and x_n to output variable (Y). Because the model predictions are continually evaluated and adjusted in accordance with the output values, the process is known as supervised learning. When employing the maximum likelihood technique, logistic regression produces the best predictions. A mathematical function called a sigmoid has the property of mapping every real value between - and + to a real value between 0 and 1. Therefore, we can classify something as belonging to the positive class if the sigmoid function's result is more than 0.5, and as belonging to the negative class if it is less

than 0.5, and if it is equal to 0.5 then we can classify it as neutral class. Although informally we sometimes use the shorthand logistic regression even when we are talking about multiple classes, technically logistic regression refers to a classifier that classes an observation into one of two classes, and multinomial logistic regression is used when classifying into more than two classes [4].

II. SYSTEM DESIGN AND IMPLEMENTATION

This section is clearly explained about the system design and implementation in Fig. 1. There are three main portions in this section which are sentiment analysis, preprocessing result, and visualization result.

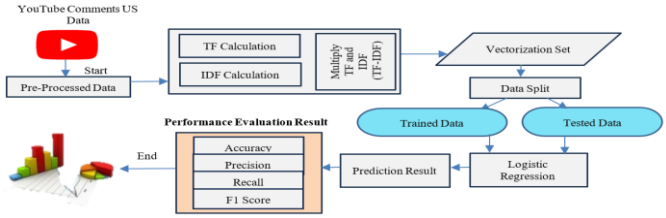

Fig. 1. System Design and Implementation

A. Sentiment Analysis

Sentiment analysis, also known as opinion mining, is used to gauge consumer opinion and assess patients' mental health based on social media posts [1]. It identifies patterns in data and categorizes it into positive, negative, and neutral categories. The analysis is conducted at document, sentence, and aspect levels, and text can be categorized based on polarity and subjectivity [1] [2]. The primary libraries used in this study are NLTK version 3.7 and Scikit-learn library 1.2.1, which can access modern classifiers with minimal configuration. The Logistic Regression model is used for sentiment analysis for three classes.

B. Preprocessing Result

The dataset for analyzing US YouTube comments includes six preparation stages, including case folding, punctuation removal, lowercase transformation, tokenization, stopword removal, stemming, and lemmatization. A google translation method translated into English is used to obtain a complete preprocessing dataset. The automated labelling program VADER uses three sentiment labels: Positive, Negative, and Neutral. YouTube comments are divided into four categories, with categories Encoded value 0-3 shown in Fig. 2 using Mathplotlib library. TABLE I displays the result of text preprocessing including the transformation of categorial values to numerical value of category encoded data, cleaned data, and the value of subjectivity and polarity.

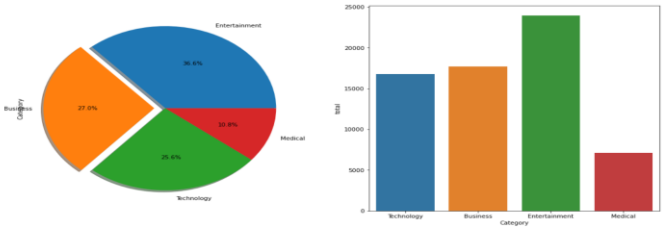

Fig. 2. Categorizing the YouTube Comments

C. Visualization Result

Fig.4 demonstrates that the cleaned text data is visualized using word cloud by category. Word cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance. Significant textual data points can be highlighted using a word cloud and to analyse the customer reviews/comments on YouTube Social Media Platform. TABLE II shows the weight of each word of the cleaned final preprocessing results calculated by using the TF-IDF method. Because it offers both the important in terms of word analysis. Then, by removing the terms that are less crucial for analysis, we may make the model-building process simpler by reducing the dimension of the input data. [2] Additionally, the TF-IDF gives us an opportunity to quantify the relevance of each word in a document by associating it with a number. Consequently, documents containing similar, pertinent terms will have comparable vectors, which is what we need in a machine learning method.

TABLE I. PREPROCESSING RESULTS

	No	News Title	Category	categoryEncoded	news_title	news_title_clean	Subjectivity	Polarity
0	1	Google+ rolls out 'Stories' for tricked out ph...	Technology	3	Google+ rolls out 'Stories' for tricked out ph...	google roll out story for tricked out photo pl...	0.000000	0.000000
1	2	Dov Charney's Redeeming Quality	Business	0	Dov Charney's Redeeming Quality	dov charneys redeeming quality	0.500000	0.500000
2	3	White God adds Un Certain Regard to the Palm Dog	Entertainment	1	White God adds Un Certain Regard to the Palm Dog	white god add un certain regard to the palm dog	0.285714	0.107143
3	4	Google shows off Androids for wearables, cars,...	Technology	3	Google shows off Androids for wearables, cars,...	google show off android for wearable car tv	0.000000	0.000000
4	5	China May new bank loans at 870.8 bln yuan	Business	0	China May new bank loans at 870.8 bln yuan	china may new bank loan at 8708 bln yuan	0.454545	0.136364

TABLE II. TF-IDF VECTORIZING THE CLEANED TEXT RESULT

	8708	add	android	bank	bln	car	charneys	china	dog	dov	...	redeeming	regard	roll	story	tricked
0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.0	...	0.0	0.000000	0.420669	0.420669	0.420669
1	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.5	0.000000	0.000000	0.5	...	0.5	0.000000	0.000000	0.000000	0.000000
2	0.000000	0.377964	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.377964	0.0	...	0.0	0.377964	0.000000	0.000000	0.000000
3	0.000000	0.000000	0.463693	0.000000	0.000000	0.463693	0.0	0.000000	0.000000	0.0	...	0.0	0.000000	0.000000	0.000000	0.000000
4	0.377964	0.000000	0.000000	0.377964	0.377964	0.000000	0.0	0.377964	0.000000	0.0	...	0.0	0.000000	0.000000	0.000000	0.000000

III. PERFORMANCE EVALUATION

This section is carried out to evaluate the performance of the Multinomial Naïve Bayes classification model using the parameters of accuracy, precision, recall, F1 score, and support. Additionally, the accuracy completeness of the macro average and weighted average are evaluated performance metrics as well. These metrics are: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

TABLE III. PERFORMANCE TEST RESULT OF BUSINESS CATEGORY

Business Category					
	Accuracy	Precision	Recall	F1Score	Support
Scale 9:1	91.01%	89%	89%	89%	1811
Scale 8:2	91.01%	89%	89%	89%	3571
Scale 7:3	91.01%	88%	89%	88%	5333
Scale 6:4	90%	88%	88%	88%	7139

TABLE IV. PERFORMANCE TEST RESULT OF ENTERTAINMENT CATEGORY

Entertainment Category					
	Accuracy	Precision	Recall	F1Score	Support
Scale 9:1	91.01%	93%	97%	95%	2410
Scale 8:2	91.01%	93%	97%	95%	4758
Scale 7:3	91.01%	92%	97%	95%	7219
Scale 6:4	90%	91%	97%	94%	9569

TABLE V. PERFORMANCE TEST RESULT OF MEDICAL CATEGORY

Medical Category					
	Accuracy	Precision	Recall	F1Score	Support
Scale 9:1	91.01%	94%	82%	87%	680
Scale 8:2	91.01%	93%	82%	87%	1409
Scale 7:3	91.01%	94%	80%	87%	2044
Scale 6:4	90%	93%	78%	85%	2853

TABLE VI. PERFORMANCE TEST RESULT OF TECHNOLOGY CATEGORY

Technology Category					
	Accuracy	Precision	Recall	F1Score	Support
Scale 9:1	91.01%	91%	89%	90%	1653
Scale 8:2	91.01%	90%	89%	89%	3369
Scale 7:3	91.01%	90%	88%	89%	5065
Scale 6:4	90%	90%	87%	88%	6653

According to the research finding in TABLE III, IV, V and VI, the Logistic Regression model's performance testing results show minimal prediction defects compared to those that performed well in line with the classification model. The model's accuracy levels were tested on scales of 9:1, 8:2, 7:3, and 6:4 data, with the highest accuracy, precision, recall, and F1 scores at these scales. The 9:1, 8:2, and 7:3 scales achieved the highest accuracy of 91.01%, with no change in accuracy. The 6:4 scale showed a slight decrease in accuracy to 90%. Precision, recall, and F1 scores have little changes. However, there was a significant change in support count depending on the occurrence of the class according to the ratio, suggesting that performance increases as the amount of data increases. This suggests that the model's performance improves with more data.

Fig. 3. Sentiment Analysis Result for US YouTube Comments

In addition, subjectivity and polarity values are also calculated for each sentiment result. In my research study, polarity is the output that occurs between [-1,1], where -1 denotes negative sentiment, +1 denotes positive sentiment, and 0 denotes neutral. Subjectivity is the output that is within the range [0,1] and pertains to subjective assessments. Subjectivity measures how much factual information and subjective opinion are present in the text. The content contains personal opinion rather than factual information due to the text's heightened subjectivity. The findings result from sentiment analysis of US-based YouTube comments are displayed in Fig. 3: There are 58.3% of neutral comments, 14.0% of negative comments, and 27.7% of positive comments respectively. In TABLE VII, VIII, and IX, display the results of three classes: positive, negative, and neutral of YouTube sentiment comments with subjectivity, and polarity values. Fig. 4 shows the performance analysis results that were achieved after doing sentiment analysis on YouTube video comments using six machine learning models, in accordance with my research observation. The Logistic Regression Model is employed because it has the greatest accuracy value among the six machine learning models, and findings show that it generates accurate results.

TABLE VII. DISPLAYING THE POSITIVE COMMENTS

No		News Title	Category	categoryEncoded	news_title	news_title_clean	Subjectivity	Polarity	Sentiment
1	2	Dov Charney's Redeeming Quality	Business	0	Dov Charney's Redeeming Quality	dov charneys redeeming quality	0.500000	0.500000	Positive
2	3	White God adds Un Certain Regard to the Palm Dog	Entertainment	1	White God adds Un Certain Regard to the palm Dog	white god add un certain regard to the palm dog	0.285714	0.107143	Positive
4	5	China May new bank loans at 870.6 bln yuan	Business	0	China May new bank loans at 870.6 bln yuan	china may new bank loan at 8706 bln yuan	0.454545	0.136364	Positive
7	8	Apple & Google's Motorola end legal battle	Technology	3	Apple & Google's Motorola end legal battle	apple google motorola end legal battle	0.200000	0.200000	Positive
11	12	Angela Bassett to direct Whitney Houston biopi..	Entertainment	1	Angela Bassett to direct Whitney Houston biopi..	angela bassett to direct whitney houston biopi..	0.400000	0.100000	Positive

TABLE VIII. DISPLAYING THE NEGATIVE COMMENTS

No		News Title	Category	categoryEncoded	news_title	news_title_clean	Subjectivity	Polarity	Sentiment
20	21	Most people ready to just assume Colin Kaepern..	Business	0	Most people ready to just assume Colin Kaepern..	most people ready to just assume colin kaepern..	0.666667	-0.100000	Negative
34	35	Bloomberg Soda Ban Really Dead This Time	Medical	2	Bloomberg Soda Ban Really Dead This Time	bloomberg soda ban really dead this time	0.400000	-0.200000	Negative
41	42	Experts call secret Facebook experiment unethical	Technology	3	Experts call secret Facebook experiment unethical	expert call secret facebook experiment unethical	0.700000	-0.400000	Negative
44	45	Eurozone down as China growth accelerates	Business	0	Eurozone down as China growth accelerates	eurozone down a china growth accelerates	0.288889	-0.155556	Negative
61	62	'Game of Thrones' Mon-Sat: How to Survive the Gap	Entertainment	1	'Game of Thrones' Mon-Sat: How to Survive the Gap	game of throne monsat how to survive the gap	0.400000	-0.400000	Negative

TABLE IX. DISPLAYING THE NEUTRAL COMMENTS

No		News Title	Category	categoryEncoded	news_title	news_title_clean	Subjectivity	Polarity	Sentiment
0	1	Google+ rolls out 'Stories' for tricked out ph..	Technology	3	Google+ rolls out 'Stories' for tricked out ph..	google roll out story for tricked out photo pl..	0.000	0.0	Neutral
3	4	Google shows off Androids for wearables, cars,..	Technology	3	Google shows off Androids for wearables, cars,..	google show off android for wearable car tv	0.000	0.0	Neutral
5	6	Firefox Windows 8 Metro Browser Development Ca..	Technology	3	Firefox Windows 8 Metro Browser Development Ca..	firefox window 8 metro browser development can..	0.000	0.0	Neutral
6	7	Destiny Beta Kicks Off In July	Technology	3	Destiny Beta Kicks Off In July	destiny beta kick off in july	0.000	0.0	Neutral
8	9	UPDATE 2-Facebook Q1 revenue grows 72 percent ..	Business	0	UPDATE 2-Facebook Q1 revenue grows 72 percent ..	update 2facebook q1 revenue grows 72 percent o..	0.000	0.0	Neutral

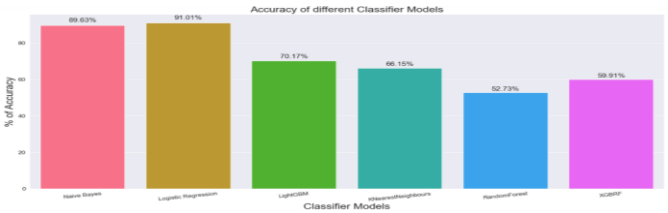

Fig. 4. Comparing the Accuracy Results of Six Machine Learning Model

IV. CONCLUSIONS

The authors discovered that certain comments on YouTube contain slang and misspellings, making them difficult to classify. The automated labeling process may produce inaccurate results due to the noise in the data. Despite these issues, the Logistic Regression model remains stable due to its best prediction. The model achieved the highest accuracy of 91.01%. For those who want to analyze about YouTube, this research highlighted that the opinions of audience from the comments of those videos can be obtained by doing sentiment analysis for attributes such as dislike that don't show a statistic count to public.

REFERENCES

[1] Mayur Wankhade[1,2] Annavarapu Chandra SekharaRao[1,2], . Chaitanya Kulkarni[1,2] ,"A survey on sentiment analysis methods, applications, and challenges", published on 7, February 2022 by Springer.

[2] [1st] Muhammad Alkaff, [2nd] Andreyan Rizky Baskara, [3rd] Yohanes Hendro Wicaksono, "Sentiment Analysis of Indonesian Movie Trailer on YouTube Using Delta TF-IDF and SVM", published on May 14, 2021 by IEEE Xplore.

[3] Akash Addiga, Sikha Bagui, "Sentiment Analysis on Twitter Data Using Term Frequency-Inverse Document Frequency", Journal of Computer and Communications, 2022, published by Scientific Research Publishing, USA.

[4] George B. Aliman, Tanya Faye S. Nivera, Jensine Charmille A. Olazo, Daisy Jane P. Ramos, Chris Danielle B. Sanchez, Timothy M. Amado, Nilo M. Arago, Romeo L. Jorda Jr., Glenn C. Virrey, Ira C. Valenzuela, "Sentiment Analysis using Logistic Regression", Journal of Computational Innovations and Engineering Applications JULY 2022.