Argumentation Analysis On Twitter and YouTube Comments Based On Social Events Using Ensemble Learning Of Language Models

RAJALAKSHMI R, Vellore Institute of Technology Chennai, India
SHIBANI ANTONETTE, University of Technology Sydney, Australia
SUJITH M, Vellore Institute of Technology Chennai, India
RITHVIK G, Vellore Institute of Technology Chennai, India

Twitter and YouTube are two vital virtual spaces where numerous opinions and arguments are posted. Analyzing those arguments paves the way to extract nine different attributes such as stance, relevancy and overall quality. For this research, tweets and comments pertaining to five different social events were scraped from the two social platforms. Unlike English and Tamil, fine-tuning the model to code-mixed and romanized data is a challenging task. This work proposes a method of majority voting of three models namely BERT-base-multilingual-cased, XLM-RoBERTa-Large and XLM-MLM-100-1280 for the task of text classification in English, Tamil and code-mixed data. Individual models are fine-tuned for each of the nine attributes. The proposed method surpasses the traditional machine learning and deep learning-based methods. The precision, recall, F1-score and accuracy metrics are used for evaluation. Different ensembling techniques such as majority voting and any class voting are experimented with. Ensembling the model improves the recall and F1 scores when compared to the individual models.

CCS Concepts: • **Computing methodologies → Artificial intelligence**; **Natural language processing**; **Information extraction**;

Additional Key Words and Phrases: Argument analysis, Twitter, YouTube, Tamil, Code-Mixed, mBERT, XLM-RoBERTa, XLM-MLM, Ensemble Learning

ACM Reference Format:
Rajalakshmi R, Shibani Antonette, Sujith M, and Rithvik G. 2018. Argumentation Analysis On Twitter and YouTube Comments Based On Social Events Using Ensemble Learning Of Language Models. *J. ACM* 37, 4, Article 111 (August 2018), 21 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Social media provides a virtual space for people to share their opinions and feelings on issues and events. Sharing an opinion tends to attract other people to provide their points of view as arguments. The argument may be valid or invalid, it may be in accordance with or a counterpoint to the main context. The overall quality of the comments varies from low (poor structure of word formation with derogatory language) to high (presents examples and support statements). People may

Authors' addresses: Rajalakshmi R, Vellore Institute of Technology Chennai, Chennai, India, rajalakshmi.r@vit.ac.in; Shibani Antonette, University of Technology Sydney, Sydney, Australia, antonette.shibani@uts.edu.au; Sujith M, Vellore Institute of Technology Chennai, Chennai, India, sujith.m2020@vitstudent.ac.in; Rithvik G, Vellore Institute of Technology Chennai, Chennai, India, rithvik.g2021@vitstudent.ac.in.

ACM 0004-5411/2018/8-ART111
https://doi.org/XXXXXXX.XXXXXXX

pass on remarks to the content. It could be a positive remark which includes praise and appreciation or a negative remark with hate and offensive language. At times, the tone and characteristics of the author are criticised. Certain comments tend to be spam or news irrelevant to the topic of debate. Although there are numerous social media platforms available, Twitter and YouTube are the top two platforms where most of the online discussions and arguments occur. On YouTube, users tend to argue in the comments section after watching a video or shorts. Twitter allows users to tweet their opinions openly on the platform and others can comment on those tweets. The capability to argue in languages other than English has increased the popularity of these platforms and allowed for a wide audience from several parts of the country to engage in arguments in their language. Tamil is one of the culturally richest languages spoken by about 90 million people in the world. Posting tweets and videos in Tamil is prevalent and they provoke a large number of arguments and comments from the language speakers. However, very few comments are written in pure Tamil language, whereas a majority of them are in code-mixed or romanized form.

A social event is a large-scale event that involves the participation of a large group of people. An event can belong to one separate community or it can be a global event involving the participation of people from all over the world. Several social events take place in a country or a state throughout the year. Such events tend to lure interactions among people in physical and virtual spaces. The COVID-19 pandemic was a global event that provoked arguments on social media. Due to the pandemic situation, most of the arguments were on virtual spaces. Apart from that, there were a few other events specific to the state of Tamilnadu such as Jallikattu, a traditional sport representing the pride of Tamil culture that stimulated lots of arguments with regards to whether the sport has to be banned or allowed. There have been arguments in the state about the sale of Alcohol, free bus passes for women and NEET (NATIONAL ELIGIBILITY ENTRANCE TEST) examination for medical aspirants. All these events prompted many comments and tweets from the people of the state.

Analysing the arguments can lead to the identification of several attributes such as their overall quality, stance regarding the topic and content, relevancy and remark. Certain attributes were classified based on Graham's hierarchy of disagreement [6]. The analysis process can be automated by building appropriate deep-learning models specific to the task of text classification. This results in the analysis of arguments for several attributes in more than one language.

2 RELATED WORK

Several works are present for analysing arguments in English from social media platforms such as Twitter and YouTube. A cooperative clustering framework based on majority voting is proposed for sentiment analysis on Twitter [2]. Unsupervised learning is implemented using three different clustering techniques namely single linkage, complete linkage and average linkage along with their combinations. This approach outperforms the state-of-the-art algorithms such as SVM and Naive Bayes classifiers. Majority voting-based cooperative clustering thus provides good overall quality clusters with a tradeoff of poor time efficiency. In another study [1], Twitter sentiment analysis on tweets regarding the monkeypox outbreak is performed using VADER and TextBlob. 56 different classification models including support vector machine, K Nearest Neighbours and Multilayer Perceptron are built to evaluate and compare the accuracy and F1-score. Lemmetization techniques along with a count vectorizer are also used. SVM yields the highest accuracy.

The introduction of transformers was an important milestone in the field of Natural Language Processing leading to the development and fine-tuning of several transformer-based models for various tasks such as text and token classification, text generation, translation and masked language modelling. The BERT model [4] was simple yet powerful and could be fine-tuned with ease for a variety of tasks such as question-answering and text classification. The XLM-RoBERTa model

[3] outperforms mBERT and supports one hundred languages. XLM-MLM [8] is another model making use of a masked language modelling technique that provides better results than XLM-RoBERTa. This model also supports one hundred languages and has more parameters than the Roberta model. The combination of these three models has the potential to provide state-of-the-art results in the task of text classification involving English, Tamil and Code-Mixed data.

[11] proposes a transformer-based capsule network built on top of the Bidirectional Encoder Representations from Transformers (BERT) model for Tweet Act classification. The model learns the attributes by joint optimization of features from the capsule layer and the BERT model. It attains a benchmark accuracy of 77.52% and an F1-score of 0.77. The model outperforms the base BERT and several other state-of-the-art models. Another study [7] experiments with the classification of health mentions on Twitter with nine different transformer-based models. The classification scores are obtained and compared with the model's parameters. The RoBERTa model achieves the best F1-score of 0.93 followed by other modified versions of the BERT model. Despite the huge number of parameters, the classification score of GPT-2 is just about 88%. The models with about 300 to 400 million parameters tend to turn out with good performance.

The works done in English were then applied to low-resource languages such as Tamil. Detection of offensive languages was a vital research field in low-resource languages due to the availability of data from social platforms. Adapters and knowledge transfer techniques are used to detect hate speech on YouTube in the Tamil language [13]. The study proves the efficiency of transformer-based models over traditional machine-learning techniques. It also proves that the adapter-based techniques outperform the fine-tuned models for languages like Tamil. The XLM-RoBERTa-Large model gives the best performance with an accuracy of 88%. The adapter-based models ensure parameter efficiency and provide good results in cross-domain datasets. Research is undertaken for abusive comments detection using machine learning, deep learning and multilingual transformer-based models on English and Tamil datasets [9]. A total of eight machine learning, two deep learning algorithms and a transformer model are used for this purpose. Random forest gives the best F1 score among all machine learning techniques and bidirectional LSTM gives the best score among all deep learning techniques.

Working with code-mixed data is yet more challenging than working with multilingual datasets. Another study [12] is done for the sentiment analysis of Tamil code-mixed data by deploying hybrid deep learning techniques such as the combination of CNN with LSTM and Bi-LSTM. This hybrid approach is compared with the traditional machine learning algorithms and transformer-based models such as IndicBERT. The proposed model outperforms all of them by achieving an F1-score of 0.66 on code-mixed data. HOTTEST, hate and offensive content detection in the Tamil language using transformers and advanced stemming techniques [10] is done with the data from YouTube. All the state-of-the-art transformer-based models such as mBERT, XLM-RoBERTa and MuRIL are used. The results tend to be better with the stemmed data rather than unstemmed data. A majority voting-based ensembling is performed as a downstream classifier. An F1-score of 0.84 is achieved with this advanced stemming technique.

3 DATASET INFORMATION

This section provides detailed information on the collection, preprocessing and annotations of the dataset. The statistics and inferences obtained as a result of exploratory data analysis are also reported.

3.1 Data Collection

The data was gathered from two different social media platforms, YouTube and Twitter, to analyse the comments and tweets of the people on these platforms. The tweets obtained from Twitter are replies to a parent tweet whereas the comments on YouTube are replies to the videos and shorts. The data collected is based on five important social events namely Jallikattu sport (Ban or Allow), NEET Examination (Ban or Allow), Alcohol consumption (Ban or Allow), Covid Vaccination (Boon or Bane), and Free bus pass for women (Boon or Bane) that took place in the state of Tamilnadu over seven years.

The keywords used for searching the content are the same as the event names. Different cases were used, which included all capitals, all small letters and the first letter capitalised. Searching with hashtags provided better results on Twitter. On YouTube, both normal videos and shorts are considered. Any YouTube video or a parent tweet is selected only if it has a minimum of forty-five replies. A minimum of five different parent tweets are taken from Twitter for each topic to ensure variety in the tweets. No two parent tweets are from the same user. Similarly, no two videos on a topic are from the same channel. Both national and state YouTube channels are selected to gather data in English, Tamil and a code-mix of both languages. It is ensured that the proportion of comments from each topic is almost identical between YouTube and Twitter.

3.2 Data Preprocessing

The collected data has several unnecessary tokens such as the HTML tags representing line breaks and links. The hashtags on YouTube are also represented as links. All these tokens are removed from the data. The emoticons and special symbols are also eliminated. Regular expressions are used for this purpose. The language of the data is auto-predicted using the pycld2 package.

As the data is also collected from sources outside the state of Tamilnadu, there are comments in languages other than English and Tamil. All such comments are removed from the dataset. To protect the user identity, the tags and mentions of individuals in the data are replaced using false stand-in tags such as "@USER_1" and "@USER_2". However, all the offensive content in the data is left unchanged as the model must also learn to predict offensive arguments.

3.3 Data annotations and coding scheme

The data is manually annotated using three human annotators for nine different attributes which are overall quality, Stance with respect to the topic, Stance with respect to the content, Argumentation, Comment, Responding to tone, Person's Characteristics, Remark and Relevancy. All three annotators are well-versed in English and Tamil. The reliability between the annotators is ensured by performing an inter-rater reliability test [5]. The rhoR and irr libraries in Rstudio were used to calculate Shaffer's Rho, Krippendorff's Alpha and Cohen's Kappa values. A total of 3480 comments were annotated over two months.

The coding scheme used for the key attributes of overall quality, Stance and Relevancy are presented in Table 1. The overall quality is computed based on the usage of grammar and structuring of the sentences. The clarity of the statement is also considered. Relevancy is the attribute that checks for the relevancy between the comments and the content or topic presented in the video or tweet. The stance with respect to topic and content are the attributes to check whether the stance of the user is in alignment with or opposing to the topic of debate and the content respectively.

The hierarchy of disagreement [6] proposed by Paul Graham is a model that categorizes arguments into different levels of disagreement based on their effectiveness and intellectual rigour. As depicted in Figure 1, it is represented in the form of a pyramid with each level representing a more constructive form of disagreement going down the pyramid. A similar structure was proposed for the hierarchy of agreement based on the disagreement hierarchy. Table 2 shows the coding scheme used for the rest of the five attributes that were constructed based on the hierarchy of agreement and disagreement.

The attribute argument is the most constructive form of agreement or disagreement where the users use high overall quality language with proper reasoning to put forth their opinions. Comments are points of view and opinions without any reasoning. The last three attributes responding to tone, discussing the person’s characteristics and remarks focus on attacking or praising a person rather than the content and topic of debate. The users may point out the tone of the speaker in the video or the tone of the tweet as being angry, happy or sarcastic. The responding to tone attribute is used for finding out all such references to the tone. Similarly, the writer characteristics attribute is used to find the discussion on the characteristics of the user rather than the content. Remarks include praise and derogation of the users or the content.

Table 1. Coding scheme for three key attributes of the dataset

Attributes	Values	Coding Scheme
Overall quality	High	• Usage of understandable grammar. • Comments with clear reasoning or evidence. • Good clarity on the opinion presented. • Example: ”So many retired employees of SETC are yet to get part of the final settlement. A good scheme should be sustainable .. from what I see transport dept will go bankrupt along with EB.”
	Med	• Usage of understandable grammar. • Commenting without reasoning or evidence. • Lacking clarity in the opinion presented. • Example: ”This is the difference between social revolution/social justice and communal politics/closed mindedness.”
	Low	• Usage of derogatory words. • Very poor grammatical structure resulting in difficulties in understanding. • Example: ”you shut your mouth.”
Stance with respect to Content	For	• Agree with the content in the tweet or video rather than the main topic. • Implicit and sarcastic stances are considered directly. • Example: ”Sir, rightly said.After so much propaganda for vaccination, if it is not available, people are disappointed.”
	Against	• Disagree with the content in the tweet or video rather than the main topic. • Implicit and sarcastic stances are considered directly. • Example: ”You are completely wrong because Covishield is an mRNA vaccine....”
	Undetermined	• A clear stance not taken with respect to the content in the tweet or video. • Example: ”I am having allergy by ciprofloxacin tablet, can i take corona vaccine?”
Stance with respect to Topic	For	• Agree with the main topic rather than the content in the tweet or video. • Implicit and sarcastic stances are considered directly. • Example: ”#Jallikattu is a part of their tradition n culture.Tamil ppl has an emotional attachment wit it.Respect their culture atleast!”
	Against	• Agree with the main topic rather than the content in the tweet or video. • Implicit and sarcastic stances are considered directly. • Example: "தடுப்பூசி உயிரை காப்பதில்லை நோய் எதிர்ப்புச் சக்தியே உயிரை காக்கிறது அதனை அதிகரித்துக் கோள்ளுங்கள்."
	Undetermined	• A clear stance not taken with respect to the main topic. • Example: ”1st injunction pottum, safe aa irundhum, silarukku corona varudhu, en?”
Relevancy	Relevant	• Proper connection between the comment and the content is established. • Sarcastic content which may not be explicitly relevant can still be relevant. • Example: ”Thanks for sharing the video as it is clear to understand the content with easy examples.”
	Irrelevant	• Spam content that is not related to the context. • Example: ”Hi guys, please do check out my new channel. Do no forget to subscribe.”

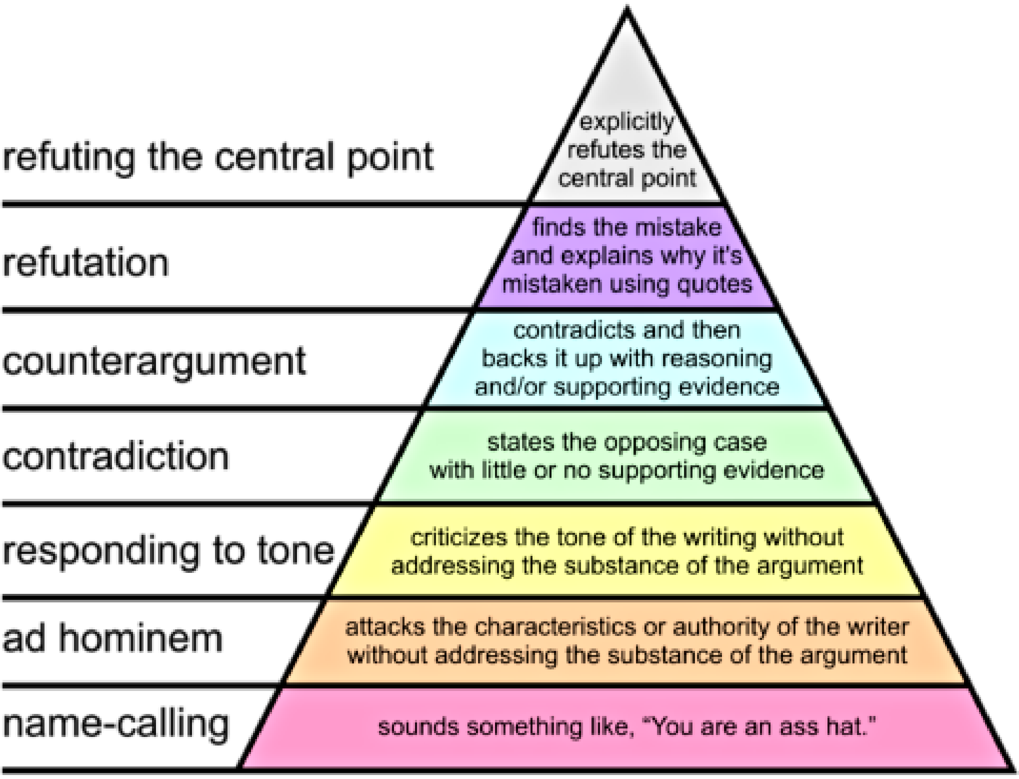
refuting the central point
explicitly refutes the central point
refutation
finds the mistake and explains why it's mistaken using quotes
counterargument
contradicts and then backs it up with reasoning and/or supporting evidence
contradiction
states the opposing case with little or no supporting evidence
responding to tone
criticizes the tone of the writing without addressing the substance of the argument
ad hominem
attacks the characteristics or authority of the writer without addressing the substance of the argument
name-calling
sounds something like, "You are an ass hat."

Fig. 1. Grahams model for hierarchy of disagreement

Table 2. Coding scheme for sub categories based on the hierarchy of disagreement

Attributes	**Coding Scheme**
Argument (1/0)	• Provide reasoning and evidence. • Detailed explanation of the context. • Proper stance is taken with evidence to back it up. • Grammatically and logically consistent. • Words such as ‘because’, ‘since’, and ‘as’, imply the reasoning. • Words in Tamil that could be present are ‘ஏனெனில்’, ‘இருப்பதால்’ and ‘காரணத்தினால்’ • Example: ”NEET is an unnecessary test. It only fills the pockets of coaching centers. NEET permits multiple attempts and so it does not always bring out real talents. The propaganda that one can clear NEET by studying 11th and 12th text books alone clearly and throughly is a big lie. It requires extra coaching which everyone cannot afford. On the other hand, plus 2 board exam tests students both on a descriptive as well as objective basis. Because of tests like NEET, students have lost the ability to write descriptive answers. Tests like NEET have totally killed the cognitive skills, creative skills, and writing skills of students. Since students concentrate only on tests like NEET they totally ignore plus 2 curriculum and students are not even able to write official letters and leave letters. It’s good time we scrap NEET totally.”
Comment (1/0)	• Providing a point of view and own opinion related to the content. • The content may or may not be accurate. • The content may or may not have reasoning. • Example: ”Neet is important”
Responding to tone (1/0)	• Responding to the tone of the user. • The content of the argument may or may not be addressed. • Example: ”அதோட விடு என்று சொல்வதே ஆணவம் கலந்த வார்த்தை தான்"
Discussing a person’s characteristics (1/0)	• Talking about the characteristics of the writer or host. • It could be praise or attack. • The content of the argument may or may not be addressed. • Example: ”Excellent.. In 10 minutes video you have explained a himalayan size biological matters. Now i am very clear. Among the worst youtubers, I found a best and social responsible youtuber. Thank you sir..”
Remark (1/0)	• Passing remarks on the user or the context. • It could be praise or attack. • Includes name-calling, offensive words and negative intent of argument. • Example: ”Super brother , I am watching your videos very very super”

3.4 Dataset statistics and inferences

This section presents all the results and inferences obtained by performing exploratory data analysis on the dataset. Results generic to the entire dataset and specific to the attributes are present.

The Twitter and YouTube datasets consist of 1350 tweets and 2130 comments respectively. The proportion of comments from Twitter and YouTube belonging to each topic is given in Table 3.

Table 3. The proportion of comments from Twitter and YouTube on each topic

Topic	Twitter (in %)	YouTube (in %)
Covid vaccine - Boon or bane	28.3	32.5
Jallikattu	25.6	24.4
Alcohol and Drugs	21.2	20.1
Free bus commute for women. Boon or Bane	19.8	16.4
NEET: Boon or Bane	5.1	6.6

3.5 Word and Character counts

Computing the word and character counts reveals the users' intention and commitment to argue on social platforms. Comments with short word counts portray that the users only want to provide short opinions or remarks on the matter of debate rather than a strong argument. Table 4 gives the mean and mode of the word counts and the average character counts for the Twitter and YouTube datasets. The mean word count is much less on Twitter as compared to YouTube. The restriction on the word count of a tweet on Twitter may be the reason for this occurrence. Thus, YouTube is the platform for large-scale arguments.

Table 4. Word and character counts for Twitter and YouTube dataset

Count Attribute	Twitter	YouTube
Mean Word Count	14	20
Word Count with high frequency (Mode)	4	3
Mean Character Count	92	137

4 METHODOLOGY

Multilingual Large Language models were utilised for the task of text classification on various categories mentioned above. All models are downloaded from the Hugging Face API. Three models were considered for the research, all of which are MLMs (Masked Language Models). This section describes the architecture and functionality of these models along with the ensemble methods used in detail.

4.1 Models

4.1.1 Bert-base-multilingual-cased. Bert-base-multilingual cased is a transformer-based model developed by Google. Their paper [4] introduces bidirectional training and attention mechanisms to capture complex sentence contextual relationships. This overcomes the limitations of unidirectional training done by models such as GPT2 that only allow predictions based on previous tokens and not subsequent tokens. It is one of the first models to implement MLM architecture for pre-training and uses WordPiece tokenization. A simple flow of masked language modelling is depicted in figure 2.

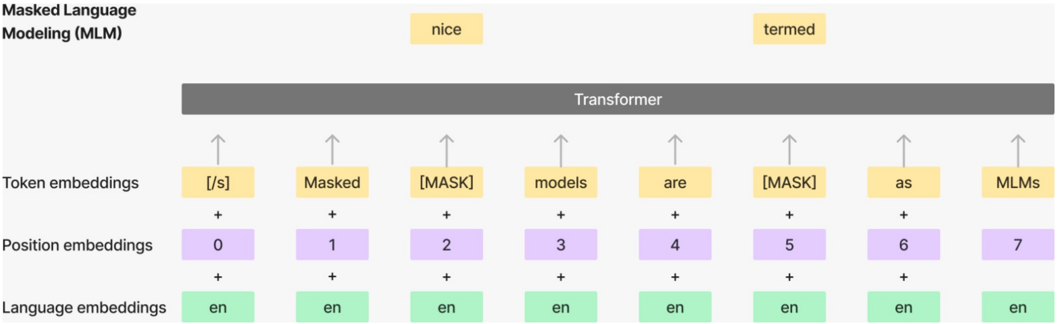

Fig. 2. Depiction of Masked Language Modeling

4.1.2 XLM-MLM-100-1280. Developers at Facebook AI first showed the effectiveness of training a model in multilingual data with their XLM-100. They suggest a few improvements over monolingual training that could overall enhance the model performance on multilingual data. The paper elaborates on how BPE (Bit Pair Encoding) tokenization of the words helps better align embedding spaces across different languages that share similar letter symbols or nouns. The encoding scheme splits on the concatenation of data sampled using a multinomial distribution, which eases the bias towards high-resource languages and increases the number of tokens associated with low-resource languages.

Although there are similar models that use CLM, and CLM + MLM schemes in parallel, a specific version of this model labelled in Hugging Face as XLM-MLM-100-1280 (Standing for XLM Masked Language Model, with 1280 hidden units) is used as it is the best publicly available model trained with multiple languages.

4.1.3 XLM-Roberta-Large. XLM-Roberta-Large is an extension of the BERT architecture from Facebook AI. This model is specifically optimised for cross-lingual tasks, making it adept at capturing nuanced representations for text classification. It is claimed to outperform mBERT by 23% on low-resource languages over various benchmarks. The creators also suggest that the previous XLM and mBERT models are under-tuned.

One of the key differences in their training methodology is using the CommonCrawl corpus for training, as they suggest Wikipedia lacks proper resources for low-resource languages. It is to be noted that the Wikipedia corpus is the data on which mBERT and XLM-100 models are trained.

They also provide a table of tokens used for all their languages. Their dataset is a cumulation of 97 languages, including their Romanized counterparts (Such as Tamil and Hindi typed with English symbols), and including such languages is an improvement to their previous XLM-100 model. However, in the CC-100 Corpus tokens, 500 M is from Tamil (The language we consider in this paper for evaluation) and only 36 M is from Romanized Tamil, compared to 55608M of English.

XLM-R utilizes SPM (Sentence Piece Model) for tokenization of its data. Both mBERT and XLM-100 use language-specific embedding schemes such as WordPiece and BPM. Such schemes fail to provide performance on languages that have unique syntaxes. For example, certain languages do not use spaces to represent the separation of words, and WordPiece embedding cannot make a difference in words on raw data. Table 5 displays various attributes for each of the three models used for classification.

Table 5. Basic description of the models used for classification

Model	Parameters	Hidden Units	#lc	Dataset	Tokenization
BERT-Base-Multilingual	110M	768	104	Wikipedia	WordPiece
XLM-Roberta-Large	355M	1024	100	CommonCrawl	SPM
XLM-MLM-100-1280	570M	1280	97	Wikipedia	BPE

4.2 Ensemble Voting

Two different voting methods were used to ensemble the predictions from the three models. The majority voting method chooses the final output as the majority prediction of the three models. For the attributes with three classes, if all three models provide differing results, then the output of the xlm-mlm-100-1280 model is considered final as it has provided the best individual performance for most of the attributes. Single-class voting is a voting method that works in favour of one particular class. This type of voting is used to address the class imbalance issue by giving priority to the class with the lower number of samples. In this voting, if a single model predicts the output as the lower class, then it is considered the final output. For the final output to be any one of the other classes, all the models must output that class as the result.

5 EXPERIMENTAL SETUP

This section explains all the implementation details of the experiment including the hyperparameter tuning, the hardware and software used, the cloud platforms used for training and the performance metrics used for the evaluation of the proposed models.

5.1 Experimental Platform

All experiments were conducted on the Kaggle kernel using the GPU P100 accelerator. Python version 3.10.12 was used. The NLTK and Matplotlib libraries are used for all the data preprocessing, analysis and visualization tasks. The PyTorch library of version 2.0.0 was used for creating the dataloaders and fine-tuning of the models.

5.2 Hyperparamaters

Three individual models were trained independently with varying hyperparameters. Certain hyperparameters such as the batch size, loss function and optimizer are common for all three models, whereas the number of epochs differs between the models.

The training, validation and testing batch sizes were set to 32 to achieve faster training speed within the limits of GPU available. The learning rate is 0.1. Five-split cross-validation was used with the number of epochs per split set to 30, 15 and 10 for the mBert, XLM-Roberta-large and XLM-MLM-100-1280 models respectively. The difference in the number of epochs is due to the change in the total number of parameters in each of the models. Increasing the epochs above the mentioned values leads to overfitting.

The loss function used is BCEWithLogitsLoss. This loss function returns a single value between 0 and 1 for binary classification tasks, whereas a vector with the probability scores for each class is returned for multi-class classification tasks. The AdamW optimizer, a slightly modified form of the Adam optimizer for addressing the issue of weight decay is used for optimization of the models.

5.3 Training and Validation

The Twitter and YouTube datasets are merged to obtain a large dataset with 3480 comments. The merged data was split into 75 and 25 percent for training and testing respectively. Sklearn library's train_test_split function is used for this purpose. All three models are independently trained on the training dataset containing about 2610 tweets and comments. The training for each attribute is done independently with separate models. The consolidated training time for all the models was around sixty hours. The tweets and comments are the only input given to the model for training. K-fold cross-validation is performed using five splits for the training and validation of the models. Cross-validation ensures a more robust assessment of the model's performance by evaluating its generalization across different subsets of the text data. This approach reduces the possibility of overfitting and underfitting to an extent.

5.4 Performance Metrics

The metrics used for the evaluation of the performance of the different models are accuracy, precision, recall, and F1 score. These are the most common metrics used for evaluation purposes in a classification task. The definition for each metric is given here.

Accuracy is the number of texts classified correctly as belonging to a particular attribute divided by the total number of texts in that attribute, as represented in Eq. (1).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

where TP is true positive (the number of correctly classified texts for each attribute), TN is true negative (the number of correctly classified texts in other attributes except the correct attribute), FP is false positive (the number of texts misclassified in other attributes except the correct attribute), and FN is false negative (the number of texts misclassified in the relevant attribute).

The number of texts correctly categorized as a certain attribute out of the total number of actual texts in that attribute is defined as recall (also known as true positive rate or sensitivity) and is computed using Eq. (2).

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

Precision is known as the number of texts correctly categorized as a particular attribute out of the total number of texts categorized as that attribute and is given by Eq. (3).

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

F1-score is the harmonic average of precision and recall. It is also known as the weighted average of precision and recall. It is calculated as in Eq. (4).

$$F1 - score = \frac{2 - precision - recall}{precision + recall} \tag{4}$$

6 RESULTS AND INFERENCES

This section discusses the results obtained from testing with the three models for the classification of nine different attributes. Each of the models is separately tested for each attribute. Various voting-based ensembling techniques were used for testing purposes. Downsampling and upsampling techniques were also incorporated to overcome class imbalance issues in the dataset.

6.1 Results for multilingual-Bert

Table 6 shows the results obtained from the multilingual Bert model. The model provides an average accuracy with better f1-scores for the values with large data. The f1-scores for responding to tone and relevancy are the lowest.

Table 6. Performance results of the Bert-base-multilingual-cased model

Attribute	Values	Precision	Recall	F1-score	Accuracy
Quality	High	0.49	0.33	0.39	0.63
	Med	0.78	0.70	0.74	
	Low	0.37	0.61	0.46	
Stance with respect to topic	For	0.42	0.24	0.31	0.55
	Against	0.49	0.55	0.52	
	Undetermined	0.62	0.65	0.64	
Stance with respect to content	For	0.40	0.44	0.42	0.66
	Against	0.35	0.24	0.29	
	Undetermined	0.77	0.79	0.78	
Relevancy	Relevant	0.92	0.99	0.95	0.90
	Irrelevant	0.15	0.03	0.05	
Categories based on the hierarchy of disagreement					
Argument	Yes	0.50	0.44	0.47	0.82
	No	0.88	0.90	0.89	
Comment	Yes	0.80	0.85	0.82	0.73
	No	0.42	0.34	0.38	
Responding to tone	Yes	0.25	0.01	0.02	0.90
	No	0.90	1.00	0.95	
Writer characteristics	Yes	0.36	0.24	0.29	0.82
	No	0.87	0.93	0.90	
Remarks	Yes	0.47	0.47	0.47	0.71
	No	0.80	0.80	0.80	

6.2 Results for XLM-Roberta

Table 7 depicts the results obtained from the xlm-roberta-large model. This model clearly overfits to the classes with high samples. As a result, the prediction scores obtained from the model for the classes with low samples such as irrelevant class and tone class is zero.

Table 7. Performance results of the XLM-Roberta-Large model

Attribute	**Values**	**Precision**	**Recall**	**F1-score**	**Accuracy**
Quality	High	0.42	0.69	0.53	0.66
	Med	0.74	0.83	0.78	
	Low	0.00	0.00	0.00	
Stance with respect to topic	For	0.00	0.00	0.00	0.51
	Against	0.62	0.05	0.09	
	Undetermined	0.51	0.99	0.68	
Stance with respect to content	For	0.53	0.29	0.38	0.72
	Against	0.38	0.05	0.08	
	Undetermined	0.74	0.95	0.83	
Relevancy	Relevant	0.91	0.99	0.95	0.91
	Irrelevant	0.00	0.00	0.00	
Categories based on the hierarchy of disagreement					
Argument	Yes	0.89	0.92	0.91	0.84
	No	0.58	0.48	0.53	
Comment	Yes	0.54	0.26	0.35	0.77
	No	0.80	0.93	0.86	
Responding to tone	Yes	0.90	1.00	0.95	0.90
	No	0.00	0.00	0.00	
Writer characteristics	Yes	0.85	1.00	0.92	0.84
	No	0.00	0.00	0.00	
Remarks	Yes	0.74	0.94	0.83	0.72
	No	0.46	0.13	0.20	

6.3 Results for xlm-mlm-100-1280 model

The overall results obatined from the xlm-mlm-100-1280 model is given in table 8. This model provides the best results of all the three models. The precision of this model is better to the classes with minimal number of samples.

Table 8. Performance results of the XLM-MLM-100-1280 model

Attribute	Values	Precision	Recall	F1-score	Accuracy
Quality	High	0.44	0.40	0.42	0.65
	Med	0.74	0.77	0.76	
	Low	0.40	0.37	0.39	
Stance with respect to topic	For	0.37	0.45	0.41	0.53
	Against	0.50	0.34	0.41	
	Undetermined	0.60	0.69	0.64	
Stance with respect to content	For	0.48	0.36	0.41	0.63
	Against	0.21	0.31	0.25	
	Undetermined	0.77	0.76	0.77	
Relevancy	Relevant	0.92	0.95	0.94	0.86
	Irrelevant	0.24	0.16	0.19	
Categories based on the hierarchy of disagreement					
Argument	Yes	0.87	0.93	0.90	0.83
	No	0.56	0.38	0.45	
Comment	Yes	0.41	0.26	0.32	0.73
	No	0.79	0.88	0.83	
Responding to tone	Yes	0.90	0.98	0.94	0.89
	No	0.28	0.06	0.10	
Writer characteristics	Yes	0.85	0.96	0.91	0.83
	No	0.27	0.08	0.12	
Remarks	Yes	0.77	0.91	0.83	0.73
	No	0.52	0.27	0.36	

6.4 Results with downsampling and upsampling

Random upsampling was performed to increase the size of the minimal samples to twice the original size. In this sampling, a data value is chosen at random from the original data and duplicated. Upsampling is done for three attributes namely responding to tone, writer characteristics and relevancy as they have the largest class imbalance. The relevancy attribute has a very poor class ratio of 1:10 for relevant and irrelevant samples respectively. Similarly, the tone and the writer's attributes have a very low number of positive samples.

Table 9 portrays the validation and testing scores for these three attributes obtained with the xlm-mlm-100-1280 model. The validation metrics for all the attributes have shown a tremendous increase whereas no significant change is visible in the test metrics. This proves the overfitting of data and makes the upsampling technique unsuitable to increase the scores.

Table 9. Validation and Test scores for the xlm-mlm-100-1280 model with upsampling

Attributes	Values	Validation Scores				Test Scores			
		Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
Responding to tone	Yes (1)	0.94	0.93	0.93	0.98	0.10	0.05	0.06	0.86
	No (0)	0.98	0.99	0.99		0.90	0.95	0.93	
Writer characteristics	Yes (1)	0.97	0.93	0.95	0.98	0.22	0.11	0.14	0.81
	No (0)	0.98	0.99	0.98		0.86	0.93	0.89	
Relevancy	Relevant	0.98	1.00	0.99	0.98	0.26	0.08	0.12	0.90
	Irrelevant	0.99	0.91	0.95		0.92	0.98	0.95	

Downsampling was experimented with the argument attribute using the xlm-mlm-100-1280 model. The negative class of the argument attribute has a large proportion and it was downsampled to about 25% of its original size to match the size of the positive class. Figure 3 demonstrates the testing results for both the positive and negative classes of the argument attribute with and without downsampling. Precision for the positive class has fallen whereas the recall rate has increased. On the other hand, the precision has gone up and the recall rate has fallen for the negative class. The increase in the overall f1-score of the positive class is not very significant and at the same time, the overall f1-score has dropped for the negative class. Therefore, downsampling is also not improving the results of the model.

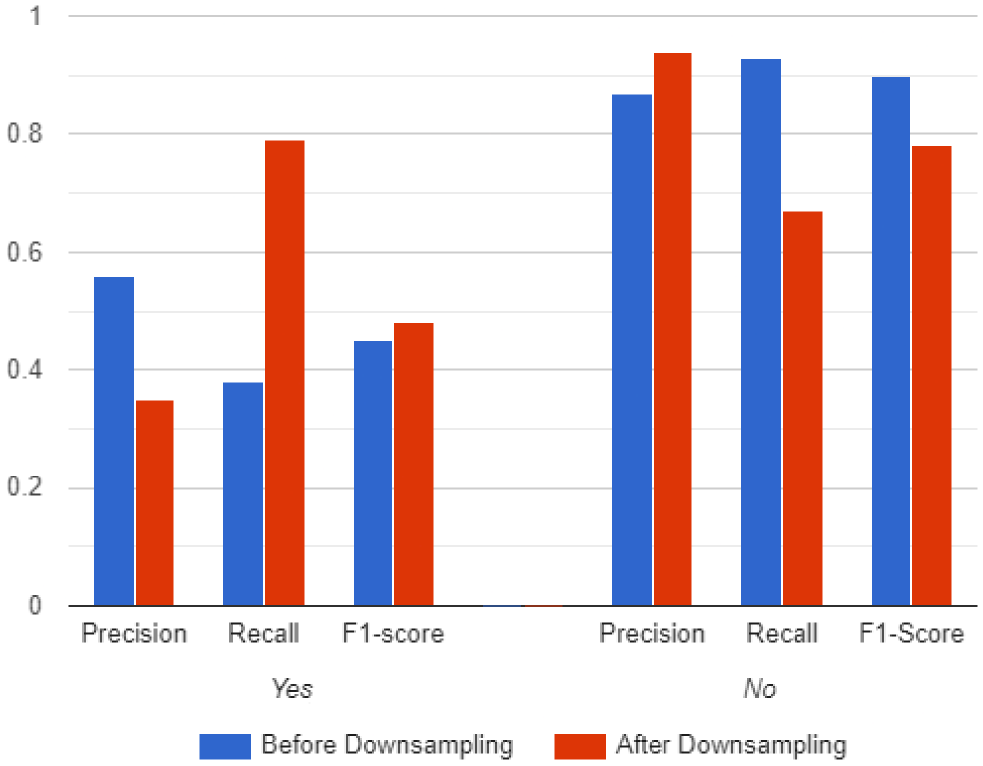

Fig. 3. Metrics for the positive and negative class of argument attribute before and after downsampling

6.5 Language specific results

Although all three chosen models support Tamil and its Romanized form, the size of the dataset used for training in Tamil is much lower when compared to the size of the English corpus. This impacts the performance of the model in this type of data. Figure 4 demonstrates the weighted f1-score of the three models classified based on language for the attribute quality. The overall performance is much better in English than in Tamil and romanized Tamil. Both the XLM models depict better performance in code-mixed and romanized Tamil as compared to pure Tamil data.

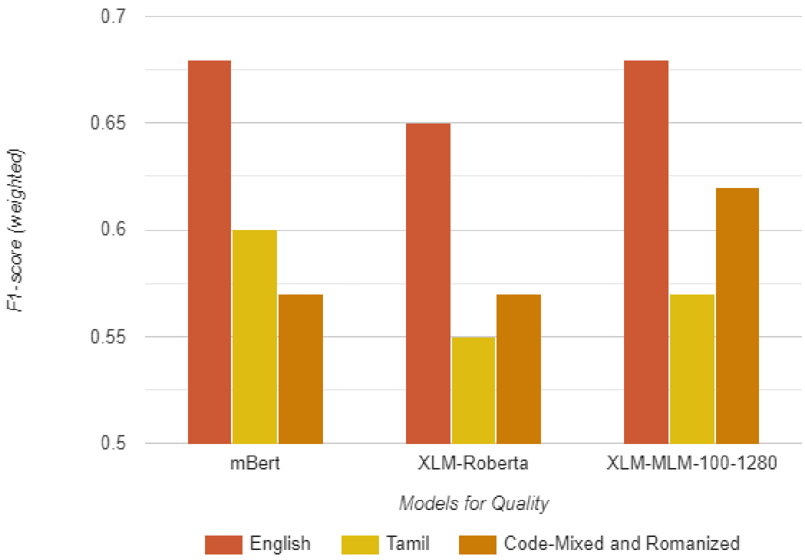

Fig. 4. Metrics for the positive and negative class of argument attribute before and after downsampling

6.6 Results with ensemble voting

The weighted F1-score for all the attributes is depicted in figure 5. The individual scores of the models as well as the scores obtained after ensembling are present. The majority voting method has provided the best score for most of the attributes. The score from single-class voting is lower than the score of individual models.

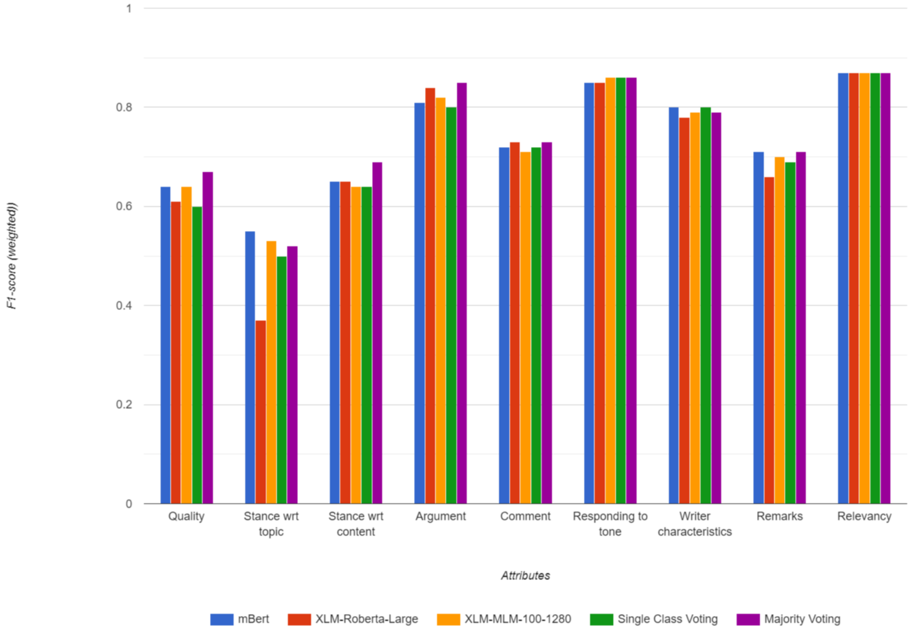

Fig. 5. Weighted F1-score for the attributes using different ensemble voting

6.7 Inference on overall quality

The proportion of comments based on overall quality, grouped by topic is depicted in figure 6. According to the distribution, Twitter has more lower quality comments than YouTube for most of the topics. NEET is the topic with the least number of low-quality comments on both platforms, whereas Jallikattu and Alcohol have had the most. The reason for this is the parent tweet and the video which has harsh content against the common people of the society. As the people are attacked, the comments and replies are derogatory. Vaccination is a topic with a good number of high-quality comments. Irrespective of the topic, medium-quality comments have the highest proportion.

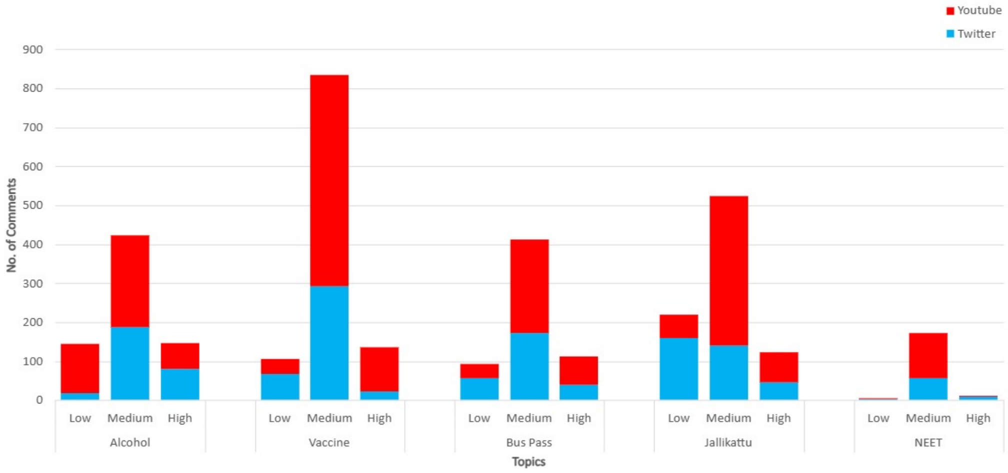

Fig. 6. Distribution of the overall quality of comments with respect to the topics

Similar to the topic, the overall quality of comments has a relation also with the language. Figure 7 shows the proportion of comments based on overall quality with respect to the language used, grouped by topics. The code-mixed or romanized language tends to have a large number of low-quality comments. This proves that the users tend to use more derogatory and unstructured words while writing in code-mixed languages. Most of the high-quality comments are in pure English followed by pure Tamil. The proportion of code-mixed comments with high quality is very low. The distribution for the medium-quality comments is almost equal in all languages.

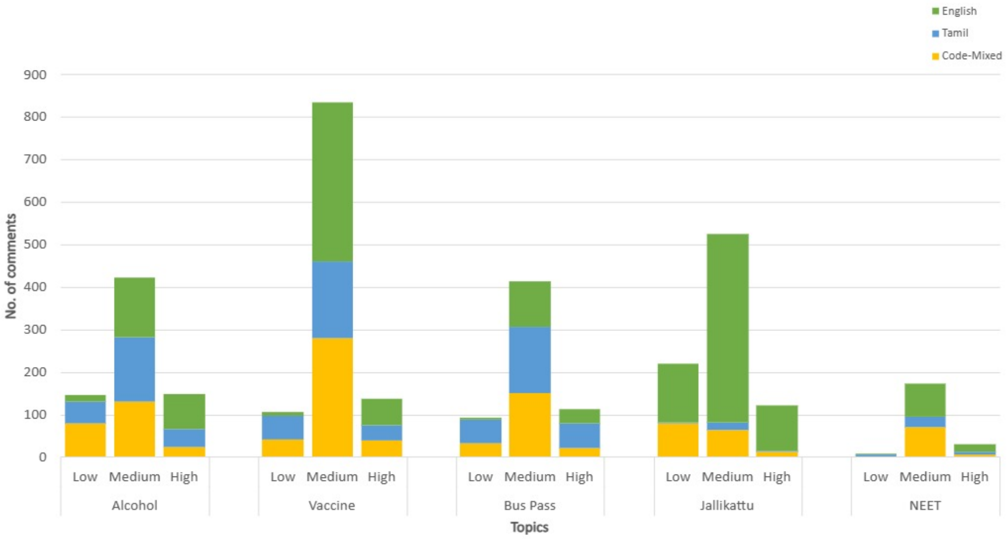

Fig. 7. Distribution of the overall quality of comments with respect to language

6.8 Inference on the Stance of the users

Figure 8 depicts the proportion of the stance of comments with respect to topic. YouTube has a very high number of comments with the stance being undetermined, whereas, on Twitter, more users tend to take a stance (For/ Against). The word length restriction on Twitter allows users to convey their messages in short and succinct sentences. However, on YouTube, huge paragraphs of stories and lamentations are found with no clear explanation or stance. Twitter being a more user opinion-oriented platform, also had comments targeted at the users rather than the topic.

The proportion for the 'against' stance is higher than that for the 'for' stance. This suggests that the users need to reply to a tweet or a video only when they disagree with the content portrayed rather than when they agree with the content. Jallikattu was a huge issue in 2017 and a high proportion of people from the state of Tamilnadu were against the ban of Jallikattu. Similarly, there is also a huge number of comments with stance taken for or against Covid vaccination. Thus, whenever an issue persists for a longer period, people develop a clear idea about it leading them to take a stance on the issue.

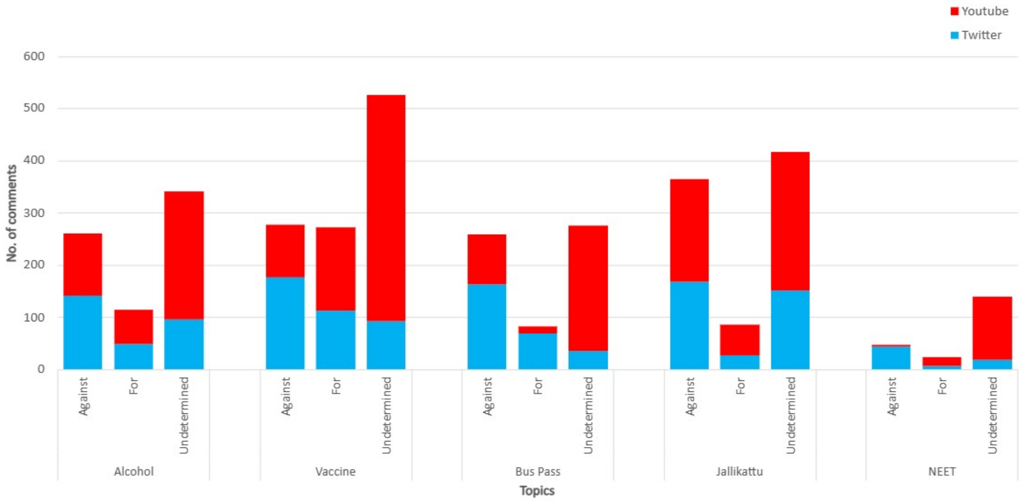

Fig. 8. Distribution of the stance of comments with respect to topic

7 CONCLUSION AND FUTURE WORKS

Various kinds of arguments and comments from two major social platforms were successfully analysed for the identification and classification of nine attributes using three different models. Two different ensemble voting techniques were used and the majority-based voting method provided the best results. The XLM-MLM-100-1280 model performed better compared to the rest of the two models. The downsampling and upsampling techniques used do not provide any significant changes to the final results of the models. The performance of the model also slightly decreases concerning Romanized and code-mixed languages.

The models proposed in the paper can further be improved or new state-of-the-art models can be developed in the future that provide better accuracy and f1-score for the classification of the given texts. All state-of-the-art models found at present are trained on a huge dataset for the English language only, whereas the dataset size for low-resource languages such as Tamil is very low. Thus, in future, models must be developed and trained with large amounts of code-mixed and romanized data in other languages too. This research can also be extended to more languages in the future rather than just English and Tamil as the models mentioned support about 100 languages in total. Similarly, other social platforms such as Instagram can be included to analyse the arguments and comments from teenagers as they constitute a large portion of the users of that platform. This allows us to track the views and opinions of the younger generation.

ACKNOWLEDGEMENTS

We extend our sincere gratitude to Dr Rajalakshmi R, Professor at VIT Chennai, and Shibani Antonette, Lecturer at the University of Technology, Sydney, for their invaluable guidance, insightful feedback, and unwavering support throughout this research. Their expertise and insights have significantly enriched the depth and overall quality of the study. Their mentor ship has been a source of inspiration, and we are deeply grateful for their commitment to fostering academic excellence.

REFERENCES

[1] Staphord Bengesi, Timothy Oladunni, Ruth Olusegun, and Halima Audu. 2023. A Machine Learning-Sentiment Analysis on Monkeypox Outbreak: An Extensive Dataset to Show the Polarity of Public Opinion From Twitter Tweets. *IEEE Access* 11 (2023), 11811–11826. https://doi.org/10.1109/ACCESS.2023.3242290

[2] Maryum Bibi, Wajid Aziz, Majid Almaraashi, Imtiaz Hussain Khan, Malik Sajjad Ahmed Nadeem, and Nazneen Habib. 2020. A Cooperative Binary-Clustering Framework Based on Majority Voting for Twitter Sentiment Analysis. *IEEE Access* 8 (2020), 68580–68592. https://doi.org/10.1109/ACCESS.2020.2983859

[3] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. arXiv:1911.02116 [cs.CL]

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL]

[5] Brendan Eagan, Jais Brohinsky, Jingyi Wang, and David Williamson Shaffer. 2020. Testing the Reliability of Inter-Rater Reliability. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge* (Frankfurt, Germany) *(LAK '20)*. Association for Computing Machinery, New York, NY, USA, 454–461. https://doi.org/10.1145/3375462.3375508

[6] Paul Graham. 2008. How to Disagree. https://www.paulgraham.com/disagree.html

[7] Pervaiz Iqbal Khan, Imran Razzak, Andreas Dengel, and Sheraz Ahmed. 2023. Performance Comparison of Transformer-Based Models on Twitter Health Mention Classification. *IEEE Transactions on Computational Social Systems* 10, 3 (2023), 1140–1149. https://doi.org/10.1109/TCSS.2022.3143768

[8] Guillaume Lample and Alexis Conneau. 2019. Cross-lingual Language Model Pretraining. arXiv:1901.07291 [cs.CL]

[9] Ratnavel Rajalakshmi, Ankita Duraphe, and Antonette Shibani. 2022. DLRG@DravidianLangTech-ACL2022: Abusive Comment Detection in Tamil using Multilingual Transformer Models. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, Bharathi Raja Chakravarthi, Ruba Priyadharshini, Anand Kumar Madasamy, Parameswari Krishnamurthy, Elizabeth Sherly, and Sinnathamby Mahesan (Eds.). Association for Computational Linguistics, Dublin, Ireland, 207–213. https://doi.org/10.18653/v1/2022.dravidianlangtech-1.32

[10] Ratnavel Rajalakshmi, Srivarshan Selvaraj, Faerie Mattins R., Pavitra Vasudevan, and Anand Kumar M. 2023. HOTTEST: Hate and Offensive content identification in Tamil using Transformers and Enhanced STemming. *Computer Speech Language* 78 (2023), 101464. https://doi.org/10.1016/j.csl.2022.101464

[11] Tulika Saha, Srivatsa Ramesh Jayashree, Sriparna Saha, and Pushpak Bhattacharyya. 2020. BERT-Caps: A Transformer-Based Capsule Network for Tweet Act Classification. *IEEE Transactions on Computational Social Systems* 7, 5 (2020), 1168–1179. https://doi.org/10.1109/TCSS.2020.3014128

[12] Kogilavani Shanmugavadivel, Sai Haritha Sampath, Pramod Nandhakumar, Prasath Mahalingam, Malliga Subramanian, Prasanna Kumar Kumaresan, and Ruba Priyadharshini. 2022. An analysis of machine learning models for sentiment analysis of Tamil code-mixed data. *Computer Speech Language* 76 (2022), 101407. https://doi.org/10.1016/j.csl.2022.101407

[13] Malliga Subramanian, Rahul Ponnusamy, Sean Benhur, Kogilavani Shanmugavadivel, Adhithiya Ganesan, Deepti Ravi, Gowtham Krishnan Shanmugasundaram, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2022. Offensive language detection in Tamil YouTube comments by adapters and cross-domain knowledge transfer. *Computer Speech Language* 76 (2022), 101404. https://doi.org/10.1016/j.csl.2022.101404

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009