Proceedings of SIGUL2022 @LREC2022, pages 33–40
Marseille, 24-25 June 2022

Corpus Creation for Sentiment Analysis in Code-Mixed Tulu Text

Asha Hegde[1 a]
Mudoor Devadas Anusha[1 b]
Sharal Coelho[1 c]
Hosahalli Lakshmaiah Shashirekha[1 d]
Bharathi Raja Chakravarthi[2 e]
[1]Department of Computer Science, Mangalore University, Mangalore, India
[2]National University of Ireland Galway, Ireland
[e]bharathi.raja@insight-centre.org
{[a]hegdekasha, [b]anugowda251, [c]sharalmucs, ,[d]hlsrekha}@gmail.com

Abstract

Sentiment Analysis (SA) employing code-mixed data from social media helps in getting insights to the data and decision making for various applications. One such application is to analyze users' emotions from comments of videos on YouTube. Social media comments do not adhere to the grammatical norms of any language and they often comprise a mix of languages and scripts. The lack of annotated code-mixed data for SA in a low-resource language like Tulu makes the SA a challenging task. To address the lack of annotated code-mixed Tulu data for SA, a gold standard trlingual code-mixed Tulu annotated corpus of 7,171 YouTube comments is created. Further, Machine Learning (ML) algorithms are employed as baseline models to evaluate the developed dataset and the performance of the ML algorithms are found to be encouraging.

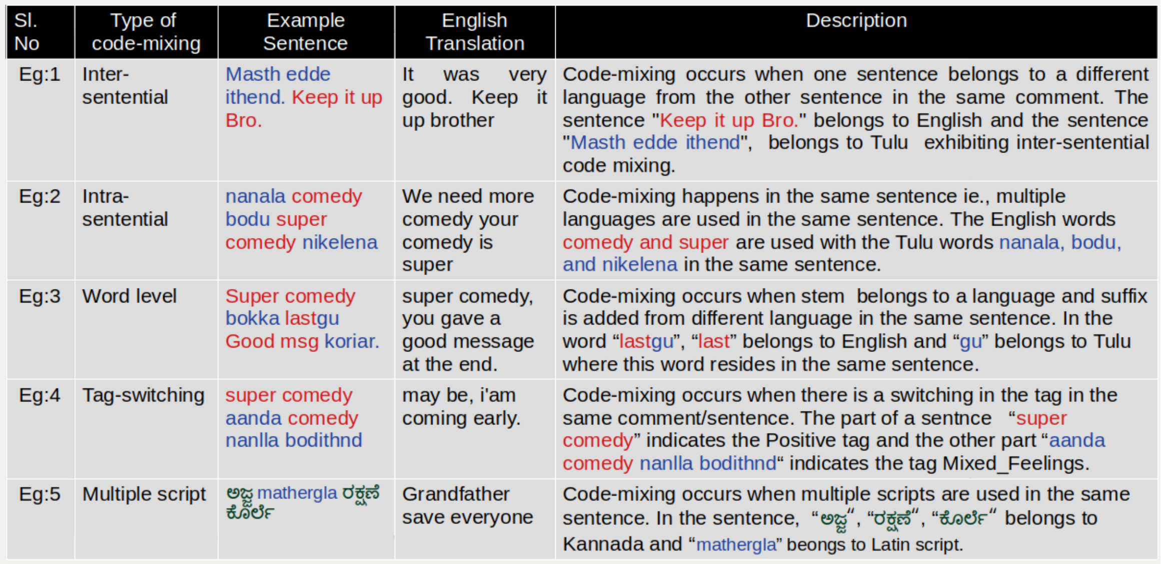
Keywords: Tulu, Code-mixed, Trilingual, Corpus creation, Sentiment Analysis

1. Introduction

Internet-enabled users express their thoughts on any topic through reviews, posts or comments on social media like YouTube, Facebook, Twitter, etc. Users knowing more than one language usually post their impressions about a topic in more than one language as there are no restrictions on the use of the languages or the grammar of any language (Scotton, 1982; Suryawanshi et al., 2020). Mixing multiple languages at different levels such as sentence, word, sub-word in the same text is referred to as code-mixing (Chakravarthi et al., 2019). Despite the fact that many languages have their own scripts, social media users in some parts of the world like India usually use non-native script to pen their comments (Bali et al., 2014). Due to the ease of entering the text in Latin and the usage of common English words, users usually enter the comments combining Latin and native scripts or only in Latin script.

The welcoming nature of online platforms encourages users from various social strata to express their thoughts/feelings on any topic. These thoughts/feelings can be extracted and used for many applications like SA. SA has recently gained popularity as a business strategy that can benefit from the insights gained from user opinions about a product or subject of interest. However, there hasn't been much effort put into analysing the sentiments of code-mixed content in many low-resourced Indian languages (Priyadharshini et al., 2021). These languages face more challenges for SA tasks due to the lack of text processing tools and annotated corpora in those languages. Some Indian languages such as Tulu, Konkani and Kashmiri are rarely explored for SA tasks.

Tulu belongs to the Dravidian language family, with over three million speakers known as Tuluvas in Karnataka, India. The majority of Tuluvas are found in Dakshina Kannada and Udupi, in the state of Karnataka and some in Mumbai, Maharashtra and in Gulf countries. Tulu is also spoken by some people in Kasargod, in the state of Kerala and it has its own script called Tigalari. The earliest written evidence of Tulu dates back to the 17th century AD, although now it exists only as a spoken language and has lost its script over time (Shetty, 2004). Despite the loss of script, Tulu is still a widely spoken language in the Southern part of Karnataka and Kannada script is prominently used to write Tulu. As Tulu is the regional language and Kannada is the official language of Karnataka, Tuluvas usually know both Tulu and Kannada languages fluently. In addition to this, many Kannada words are used in Tulu language. Further, English is predominantly known by many Tulu speaking people, especially those who are active on social media platforms. Tulu songs, videos, movies, comedy programs, skits are popular on social media. The comments posted by Tulu users for Tulu programs on social media will usually be a code-mix of Tulu, Kannada, and English. This has generated a lot of trilingual code-mixed data which is rarely explored for research purposes. In view of the availability of large volume of YouTube comments/posts in code-mixed Tulu, this study gathered comments from various YouTube Tulu songs, movies, comedy programs, skits, and serials to create a code-mixed Tulu dataset for SA. Sample comments from the proposed code-mixed



Sl. No	Type of code-mixing	Example Sentence	English Translation	Description
Eg:1	Inter-sentential	Masth edde ithend. Keep it up Bro.	It was very good. Keep it up brother	Code-mixing occurs when one sentence belongs to a different language from the other sentence in the same comment. The sentence "Keep it up Bro." belongs to English and the sentence "Masth edde ithend", belongs to Tulu exhibiting inter-sentential code mixing.
Eg:2	Intra-sentential	nanala comedy bodu super comedy nikelena	We need more comedy your comedy is super	Code-mixing happens in the same sentence ie., multiple languages are used in the same sentence. The English words comedy and super are used with the Tulu words nanala, bodu, and nikelena in the same sentence.
Eg:3	Word level	Super comedy bokka lastgu Good msg koriar.	super comedy, you gave a good message at the end.	Code-mixing occurs when stem belongs to a language and suffix is added from different language in the same sentence. In the word "lastgu", "last" belongs to English and "gu" belongs to Tulu where this word resides in the same sentence.
Eg:4	Tag-switching	super comedy aanda comedy nanlla bodithnd	may be, i'am coming early.	Code-mixing occurs when there is a switching in the tag in the same comment/sentence. The part of a sentnce "super comedy" indicates the Positive tag and the other part "aanda comedy nanlla bodithnd" indicates the tag Mixed_Feelings.
Eg:5	Multiple script	ಅಜ್ಜ mathergla ರಕ್ಷಣೆ ಕೊರ್ಲೆ	Grandfather save everyone	Code-mixing occurs when multiple scripts are used in the same sentence. In the sentence, "ಅಜ್ಜ", "ರಕ್ಷಣೆ", "ಕೊರ್ಲೆ" belongs to Kannada and "mathergla" beongs to Latin script.

Table 1: Sample code-mixed Tulu comments in the corpus

Tulu dataset along with the type of code-mixing are shown in Table 1.

In view of the lack of annotated code-mixed Tulu dataset for SA, this paper contributes by releasing the gold-standard trilingual code-mixed Tulu dataset to perform SA and presents the comprehensive results of using traditional ML classification methods to set the benchmark for the dataset. In most of the cases usually code-mixing includes two languages. However, the proposed dataset has code-mixing of Tulu, Kannada and English which makes it unique.

The rest of the paper is organized as follows: Section 2 throws light on SA in other Dravidian languages and Section 3 describes the procedure of corpus creation and annotation followed by the description of ML algorithms used to create baseline models in Section 4. Experiments and results are presented in Section 5 followed by the conclusion in Section 6.

2. Related Work

Due to the growth of social media, SA has become significantly important. Extensive research is being carried out on SA of monolingual corpora belonging to high-resource languages such as English, French, and Russian. However, only one work has been reported on SA in Tulu language and very less number of SA works are found for other Dravidian languages too. Some of the recent works on Dravidian languages using code-mixed text are described below:

Chakravarthi et al. (Chakravarthi et al., 2020b) have created a Tamil-English code-mixed annotated corpus for SA of YouTube comments. The corpus contains 15,744 code-mixed comments and each comment in the dataset is annotated by a minimum of three annotators. They implemented traditional ML algorithms, namely: Support Vector Machine (SVM), Decision Trees (DT), Multinomial Naive Bayes (MNB), Logistic Regression (LR), k-Nearest Neighbor (kNN), and Random Forest (RF) using Term-Frequency-Inverse-Document-Frequency (TF-IDF) of word n-grams in the range n = (1, 3) as features. Further, they have implemented Deep Learning (DL) models, namely: 1D Convolutional Long Short Term Memory (1DConvLSTM) and LSTM using the Keras embedding[1] and Dynamic Meta Embedding (DME) respectively. Further, the authors also implemented a transformer based classifier with multilingual Bidirectional Encoder Representations from Transformers (mBERT) for SA of code-mixed Tamil-English language. Among all the models RF model obtained the highest macro F1-score of 0.65.

KanCMD, a Kannada code-mixed dataset was developed by Hande et al. (Hande et al., 2020) by scraping YouTube comments[2]. The comments were segmented into sentences and each sentence was annotated by 5 annotators at three levels. KanCMD consists of 7,671 comments released for multitask learning of Offensive Language Detection (OLD) and SA. Both the tasks were adressed using traditional ML algorithms (SVM, MNB, DT, LR, kNN and RF) and DL based models (1DConvLSTM and LSTM). TF-IDF values, Keras embedding and DME of words were used as features to train ML models, 1DconvLSTM model and LSTM model respectively. Further, they also implemented a transformer based classifier with mBERT to perform SA of KanCMD dataset. The LR model outperformed other models with macro F1-scores of 0.57 and 0.66 for SA and OLD respectively.

Reddy et al. (Appidi et al., 2020b) presented a

[1] https://keras.io/api/layers/core_layers/embedding/

[2] https://github.com/philbot9/youtube-comment-scraper-cli

code-mixed Kannada-English corpus which is a collection of tweets extracted from Twitter on topics like sports, trending, hashtags, politics, movies and events for Parts-Of-Speech (POS) tagging. Conditional Random Fields (CRF), Bidirectional LSTM (BiLSTM), and BiLSTM+CRF are implemented to tag POS for code-mixed Kannada-English corpus. TF-IDF of character n-grams and word n-grams in the range n = (1, 3) followed by the count of common symbols, capitalization of words and numbers are used as features to train their models. Among the three models, BiLSTM+CRF model achieved the best results with macro F1-score of 0.81. Reddy et al. (Appidi et al., 2020a) have adressed the problem of emotion prediction using Kannada-English code-mixed tweets annotated with emotions. The authors trained the SVM classifier using TF-IDF of character n-grams, word tri-grams, and count of English negative words[3], punctuation, capitalization, and repetitive characters as features. They used the Keras embedding to train LSTM model and the LSTM model outperformed the SVM model with an accuracy of 32%.

Kusampudi et al. (Kusampudi et al., 2021) presented Twitter and Blog datasets for code-mixed Telugu-English text to perform SA. The authors implemented traditional ML models (SVM, MNB, DT, LR, KNN and RF), DL models (Convolutional Neural Network, BiLSTM) and hybrid models (BiLSTM+CRF and BiLSTM+LSTM) to predict sentiments in code-mixed Telugu-English text. TF-IDF of character n-grams and word n-grams in the range n=(1,3) followed by hand picked features, namely, count of special characters, capital letters, and digits are used by the authors to train ML models. BiLSTM+LSTM model exhibited a better accuracy of 0.98 on Blog dataset and BiLSTM+CRF model achieved an accuracy of 0.99 on Twitter dataset.

Malayalam-English code-mixed annotated dataset for SA is created by Chakravarthi et al. (Chakravarthi et al., 2020a) by scraping the YouTube comments using YouTube comment-scraper[4] to extract the comments. These comments were annotated at three levels by 11 annotators. Further, the authors used Krippendorff's inter-annotator agreement to ensure the agreement between annotators. The annotated English-Malayalam dataset is used to implement traditional ML (LR, SVM, DT, RF, MNB, and kNN) and DL-based models (1DConvLSTM and LSTM) to perform SA. Authors have used TF-IDF of word tri-grams, Keras embeddings and DME as features to train ML, 1DConvLSTM, and LSTM models respectively. Further, they also implemented a transformer based classifier with mBERT and among all the models, mBERT outperformed with a F1-score of 0.75.

Kannadaguli (Kannadaguli, 2021) has created a Tulu-English code-mixed dataset of 5,536 comments for SA by scraping YouTube posts. During dataset construction, the author extracted only Tulu and Tulu-English code-mixed comments written in Latin script. Krippendorff's inter-annotator agreement was calculated to ensure the agreement between annotators. The annotated Tulu-English dataset was used to implement ML models (NB,LR, DT, k-NN, RF, SVM, and Principal Component Analysis), DL models (BiLSTM and Contextualized Dynamic Meta Embeddings), and transformer based classifier with BERT models. TF-IDF values and Keras embeddings are used as features for ML and DL models respectively. Among all the models, BiLSTM model outperformed with considerable F1-scores for all the classes.

Information of Annotators		# of Annotators
Gender	Male	2
	Female	13
Highest Education	Graduate	0
	Postgraduate	12
	Research student	3
Medium of Schooling	English	6
	Native	9
Total		15

Table 2: Details of annotators

From the literature, it is clear that the under-resourced Dravidian languages, namely, Tamil, Kannada, Malayalam, and Telugu have been rarely explored for SA. Further, to the best of our knowledge, there is only one work on SA of code-mixed Tulu text (Kannadaguli, 2021).

3. Corpus Creation and Annotation

The purpose of this work is to construct a code-mixed Tulu dataset for SA. YouTube contains a lot of videos on Tulu movies, movie trailers, skits, songs, and so on, and also the comments posted by users for these videos. These comments are used as corpus for the SA task. The corpus construction work begins by scraping the YouTube comments for the videos in Tulu using the YouTube-comment-scraper tool[5] and the comments collected are anonymized for the privacy of users. The raw data obtained from the scraper is split into sentences consisting of a single comment amounting to 48,000 comments. The comments are written entirely in English, Kannada, Tulu or in a combination of English, Tulu, and Kannada languages in Kannada/Latin script or in a combination of Kannada and Latin scripts. Hence, comments which are entirely in English language written in Latin or Kannada script are filtered out retaining the rest. It may be noted that, after filtering, the comments consist of only code-mixed Tulu content written in either Kannada and/or Latin script. This data filtering is carried out manually as there are

[3] `http://sentiment.christopherpotts.net/lingstruc.html`

[4] `https://github.com/philbot9/`

[5] `https://github.com/g1mishra/Youtube_Comment_Scraper/`

no tools/libraries to identify text in Tulu language. The comments consisting less than 3 words and longer than 15 words were removed as it is difficult to comprehend the sentiments. Further, all the emojis were removed as the majority of the comments contain only emojis without any text. Additionally, duplicate sentences are removed. This process resulted in 7,171 code-mixed comments which are subjected to annotation for SA.

3.1. Annotation Setup

Annotation scheme proposed by Mohammad et al. (Mohammad, 2016) is adopted to annotate the code-mixed Tulu data. Each comment is annotated by a minimum of 3 annotators according to the following guidelines provided to each annotator:

- **Positive :** The text provides an explicit or implicit hint that the speaker is in a positive mood.
  Ex: Masth edde ithend. Keep it up Bro.
  English translation: It was very good. Keep it up brother.

- **Negative :** The comment contains explicit or implicit clues that suggest the speaker is in a negative mood.
  Ex: Ponnu edde ijjal.
  English translation: The girl is not good.

- **Mixed-Feelings :**The text indicates both positive and negative feelings experienced by the speaker.
  Ex: Paniyere aavandina naataka
  English translation: A drama that could not be explained.

- **Neutral :** There is no indication of the speaker's emotional state. For eg: asking for likes or subscriptions, questions about the release date and conveying information etc. This state is considered as neutral state.
  Ex: Yel ganteg sari battnd.
  English translation: It became correct at 7 o'clock.

- **Not_Tulu :** These are the comments that do not contain Tulu content written in Kannada or Latin script. The entire comment may consist of English words written in Kannada script or Kannada words written in Latin and/or Kannada script.
  Ex: tulu artha agaala
  English translation: Do not understand Tulu.

The annotation process involved 15 native Tulu speakers with diversity in gender, medium of education in their schooling, and educational level, as volunteers. Table 2 shows the information about annotators involved in this work. A demonstration was given to the volunteers regarding the annotations and sample sheets with 200 comments were sent to them. If the quality of the sample annotation was good only then that annotator was selected for the annotation of the code-mixed Tulu corpus. Each volunteer was allowed to annotate as many comments from the corpus as they wish. Annotators were notified that the annotations they were going to do will be recorded and they could opt-out at any time during the annotation process. The annotation setup has two phases: (i) blind annotation where each comment is annotated by two annotators and the annotators were not allowed to discuss regarding the annotations, and (ii) verification of comments and their annotations by an annotator who did not participate in the first phase. If both the annotators in the first phase have tagged the same label for the comment then that label is considered as the final label for that comment. If there is any conflict in the labels assigned by the first two annotators, the third annotator will annotate that comment and that label will be considered as the label of that comment.

Languages	Tulu
Number of Tokens	82,763
Vocabulary Size	24,006
Number of comments	7,171
Average number of Tokens per comment	11

Table 3: Statistics of code-mixed Tulu corpus

Classes	**# of Comments**
Positive	3,164
Mixed-Feelings	1,212
Neutral	1,201
Negative	670
Not_Tulu	924

Table 4: Class-wise distribution of code-mixed Tulu annotated corpus

3.2. Inter-annotator Agreement

During annotation, the annotator has to select only one of the categories to which the comment belongs adhering to the guidelines supplied. Since multiple annotators were given the task of annotating the same piece of data, a metric is required to compare the annotation qualities. This motivates the use of inter-annotator agreement which measures how well the annotations were carried out by many annotators on the same dataset. It also indicates the degree of agreement about a category among the annotators, but not whether the annotations are accurate. In other words, high inter-annotator agreement implies that guidelines are clear and interpretations are accurate.

Krippendorff's alpha (α) - a popular inter-annotator agreement algorithm is employed to measure the degree of agreement between annotators, despite its computational complexity (Krippendorff, 2011). This agreement is more relevant as it is not affected by miss-

Classes	Train set	Test set
Positive	2,501	663
Mixed-Feelings	953	248
Neutral	984	228
Negative	548	122
Not_Tulu	750	174

Table 5: Details of Train and Test set

ing data, varying sample sizes, categories, or number of annotators and can be applied to any type of measurements, including nominal, ordinal, interval, and ratio. Since the annotation work was carried out by more than two persons and the same person did not annotate all of the comments, Krippendorff's alpha (α) fits better (Artstein, 2017). The range of α must be 0 to 1 and α=1 implies a perfect agreement between annotators. The annotation for code-mixed Tulu corpus produced a nominal metric agreement of 0.6832.

3.3. Difficult Examples

During annotation, it was found that as some of the comments were ambiguous, it was difficult to find out the right feelings of the users who posted those comments. Annotation of such comments seemed difficult and some of such comments are described below:

1. Yes maaatha kadetla inchina jana ippuveru, hilarious show
   -*Yes from all the places like this people are there, hilarious show*
   Because of using the word 'hilarious show' the comment becomes ambiguous whether the speaker has 'Positive' sentiment or sarcastically giving the comment.

2. Valtaranna erege daye bodu Ladaye?
   - *Valter brother why you want fighting?*
   The comment conveys in a positive way that fighting is not good. However, the annotator cannot decide whether the comment has 'Positive' sentiment or 'Mixed-Feelings' as there are no explicit clues to identify 'Positive' sentiment.

3. Yappa devare ivaru yalli avaru marre
   -*My God from where he is?*
   In the comment, the words 'Yappa devare' and 'marre' belong to both Kannada and Tulu. Hence, difficult to decide whether it belongs to 'Not_Tulu' or 'Mixed-Feelings' class.

4. Comedy jaasti uppad. Family emotion drama maata maltar da flop aapundu.
   -*Need more comedy. If you add more family sentiments and drama then it will flop.*
   From the comment, it is difficult to decide whether the speaker liked the comedy or disliked it.

According to the instructions given to the annotators, the comment which has explicit clues are utilized for annotations. However, some examples have subtle sentiments which are different than the sentiments that can be decided from the explicit clues. Hence, some comments have shown disagreement between the annotators.

3.4. Dataset

Corpus statistics are given in Table 3 and class-wise distribution of the annotated corpus is shown in Table 4. The comments are categorized into five groups: Positive, Negative, Neutral, Mixed-Feelings, and Not_Tulu. Among 7,171 comments, 3,164 comments have a Positive polarity which is the most common category. Since there are only a few YouTube channels in Tulu language compared to other languages, the majority of the viewers encourage such channels with positive comments. The second common categories in this corpus are Mixed-Feelings and Neutral with 1,212 and 1,201 comments respectively. Because, most of the comments collected from YouTube are from Tulu songs, movies, movie trailers and skits, the users show either the ambiguity in their emotion or they just convey some information. Further, Not_Tulu and Negative categories have fewer comments compared to the other categories with 924 and 670 comments respectively. This is because, Tulu channels attract specially Tuluvas and there is least possibility that they post negative comments on the video/work of someone who belongs to their region or community. The dataset will be made available to the research community for exploring different models for SA.

4. Baseline Classifiers

Traditional ML algorithms are implemented using TF-IDF of word bigrams and trigrams as features to predict emotions in code-mixed Tulu data in order to provide baseline. The brief description of ML algorithms along with the hyper-parameters used are given below:

4.1. Multinomial Naive Bayes

Naive-Bayes classifier is a probabilistic model developed from the Bayes theorem that determines the probability of hypothesis activity based on the evidence (Xu et al., 2017). alpha - smoothing parameter value is set to 1 for MNB.

4.2. Logistic Regression

LR algorithm predicts the probability of a target variable using L2 regularization which is the default value for the penalty (Genkin et al., 2007) and the same is used in the baseline LR classifier.

4.3. Support Vector Machine

SVM is an algorithm that determines the best decision boundary between the vectors that belong to a given group (or category) and those which do not belong to

Classes	Classifiers					
	MNB			RF		
	Precision	Recall	F1 score	Precision	Recall	F1 score
Mixed-Feelings	0.39	0.04	0.07	0.53	0.19	0.28
Negative	0.83	0.04	0.08	0.46	0.17	0.25
Neutral	0.71	0.18	0.29	0.35	0.70	0.46
Not_Tulu	1.00	0.17	0.29	0.83	0.28	0.42
Positive	0.50	1.00	0.67	0.72	0.84	0.77
Macro Average	0.69	0.28	0.28	0.58	0.44	0.44
Weighted Average	0.60	0.52	0.41	0.62	0.58	0.55
	LR			SVM		
	Precision	Recall	F1 score	Precision	Recall	F1 score
Mixed-Feelings	0.47	0.25	0.33	0.41	0.29	0.34
Negative	0.49	0.17	0.25	0.45	0.33	0.38
Neutral	0.54	0.40	0.46	0.49	0.43	0.46
Not_Tulu	0.90	0.44	0.59	0.82	0.57	0.68
Positive	0.63	0.96	0.76	0.69	0.89	0.78
Macro Average	0.61	0.44	0.48	0.57	0.50	0.53
Weighted Average	0.61	0.62	0.57	0.61	0.63	**0.60**
	DT			KNN		
	Precision	Recall	F1 score	Precision	Recall	F1 score
Mixed-Feelings	0.35	0.23	0.28	0.28	0.33	0.30
Negative	0.31	0.22	0.26	0.35	0.29	0.32
Neutral	0.32	0.54	0.40	0.40	0.34	0.37
Not_Tulu	0.57	0.32	0.41	0.78	0.42	0.54
Positive	0.72	0.75	0.73	0.71	0.81	0.76
Macro Average	0.45	0.41	0.42	0.50	0.44	0.46
Weighted Average	0.54	0.53	0.52	0.56	0.56	0.55
	MLP			Cross validation		
	Precision	Recall	F1 score	Precision	Recall	F1 score
Mixed-Feelings	0.41	0.36	0.38	0.36	0.47	0.41
Negative	0.43	0.29	0.34	0.50	0.28	0.36
Neutral	0.43	0.46	0.45	0.43	0.56	0.49
Not_Tulu	0.77	0.56	0.65	0.83	0.54	0.66
Positive	0.72	0.83	0.77	0.80	0.77	0.78
Macro Average	0.55	0.50	0.52	0.58	0.52	0.54
Weighted Average	0.60	0.61	**0.60**	0.64	0.61	**0.62**

Table 6: Performance measures of the benchmark systems

that (Tong and Koller, 2001) and is implemented with L2 regularization.

4.4. k Nearest Neighbor

kNN algorithm classifies data by finding the 'k' nearest neighbors in the training data and then predicting the label of the test set based on the labels of these neighbours using the majority voting (Cunningham and Delany, 2021) and the value of 'k' is set to 3.

4.5. Decision Tree

DT algorithm is a tree-structured classifier with internal nodes representing the features of a dataset, branches representing the decision rules, and leaf nodes representing the outcome. In this classifier, classification process begins with a root node and ends with a decision made by leaves based on features (Pranckevičius and Marcinkevičius, 2017). The baseline DT classifier is implemented with max_depth = None, min_samples_split = 2, and criterion = 'gini'.

4.6. Random Forest

RF model consists of a collection of decision trees, each of which is trained using a random subset of features, and the prediction is the result of the majority vote of trees. High-dimensional noisy data can be handled well by this classifier (Shah et al., 2020). RF is implemented with the same hyper-parameter values as in DT.

4.7. Multi-Layer Perceptron

MLP classifiers are widely used in ML models due to their simplicity. It is based on neural network that consists of three types of layers: the input layer, the output layer, and one or more hidden layers. Input layer holds the input features and weighted sums of the input features are calculated by the input function. An activation function is subsequently applied to the result of this computation in order to obtain the output (Bounabi et al., 2018). The MLP model is implemented with random_state = 1 and max_iter = 300.

5. Experiments and Results

Several experiments were conducted using traditional ML algorithms, namely: MNB, LR, SVM, kNN, DT, RF, and MLP. Details of the Train and Test set are shown in Table 5 and Table 6 shows the experimental results using different ML models for SA. Precision, Recall, F1-score, macro average, and weighted average metrics are considered for evaluating the models. A Macro-average computes Precision, Recall, and F1-score independently for each class and then takes the average. Thus, it treats all the classes equally. Weighted average takes metrics from each class similar to the macro average, but the contribution from each class to the average is weighted based on the number of examples available for it.

The results illustrate that all the classification algorithms performed moderately on code-mixed Tulu data. This may be due to the characteristics of the dataset. The scores for different sentiment classes appear to be consistent with the distribution of sentiments in the dataset. Across all the sentiment classes, MLP and SVM classifiers performed comparatively better with the same weighted average F1-score of 0.60. Further, the 5-fold cross validation for SVM classifier resulted in a weighted average F1-score of 0.62.

The dataset does not have a balanced distribution. Table 4 shows that out of 7,171 comments, 44% comments belong to the 'Positive' class while the other sentiment classes share 17%, 17%, 13% and 9% for 'Neutral', 'Mixed-Feelings', 'Not_Tulu' and 'Negative' classes respectively. The Precision, Recall, and F1-score for 'Positive' class are higher than those for other classes. Further, 'Not_Tulu' and 'Negative' are the classes with lowest comments which leads to the poor results. In addition to their low distribution in the dataset, some comments are difficult to annotate even by human annotators, as mentioned in Section 3.3. Comparatively, the 'Negative' and 'Not_Tulu' classes are easy to annotate by human annotators. However, the lack of examples belonging to these classes moderates the performance of the models. Surprisingly in SVM, LR, and MLP models, the 'Negative' and 'Not_Tulu' classes obtained higher F1-scores than the 'Neutral' and 'Mixed-Feelings' classes which have more support data. This is due to more explicit clues for 'Negative' and 'Not_Tulu' words. However, the proposed code-mixed Tulu dataset is imbalanced with more support data for 'Positive' class. This resource could serve as a starting point for further research in SA of code-mixed Tulu data. There is considerable room for exploring code-mixed research with this dataset. Further, the proposed Tulu dataset has three languages and rarely explored for SA ensuring the scope for trilingual code-mixing in SA tasks.

6. Conclusion

In this paper, we have presented code-mixed Tulu dataset construction using YouTube comments for SA. Kripendorff's inter-annotator agreement is used to analyze the agreement between the annotators. Traditional ML algorithms are evaluated using TF-IDF of bi-grams and tri-grams on this code-mixed Tulu annotated corpus to provide baseline results. As the proposed work intends researchers to develop models for SA using this dataset, the dataset will be made available to the research community.

7. Acknowledgements

We thank our Masters students for helping us in creating code-mixed Tulu corpus.

8. References

Appidi, A. R., Srirangam, V. K., Suhas, D., and Shrivastava, M. (2020a). Creation of Corpus and Analysis in Code-Mixed Kannada-English Social Media Data for POS Tagging. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 101–107.

Appidi, A. R., Srirangam, V. K., Suhas, D., and Shrivastava, M. (2020b). Creation of Corpus and Analysis in Code-Mixed Kannada-English Twitter Data for Emotion Prediction. In *Proceedings of the 28th international conference on computational linguistics*, pages 6703–6709.

Artstein, R. (2017). Inter-annotator Agreement. In *Handbook of linguistic annotation*, pages 297–313.

Bali, K., Sharma, J., Choudhury, M., and Vyas, Y. (2014). I am Borrowing ya Mixing? an Analysis of English-Hindi Code Mixing in Facebook. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 116–126.

Bounabi, M., Moutaouakil, K. E., and Satori, K. (2018). A Probabilistic Vector Representation and Neural Network for Text Classification. In *International Conference on Big Data, Cloud and Applications*, pages 343–355.

Chakravarthi, B. R., Arcan, M., and McCrae, J. P. (2019). Comparison of Different Orthographies for Machine Translation of Under-resourced Dravidian Languages. In *Second Conference on Language, Data and Knowledge (LDK 2019)*, pages 6:1–6:14.

Chakravarthi, B. R., Jose, N., Suryawanshi, S., Sherly, E., and McCrae, J. P. (2020a). A Sentiment Analysis Dataset for Code-Mixed Malayalam-English. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 177–184.

Chakravarthi, B. R., Muralidaran, V., Priyadharshini, R., and McCrae, J. P. (2020b). Corpus Creation for Sentiment Analysis in Code-Mixed Tamil-English Text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 202–210.

Cunningham, P. and Delany, S. J. (2021). k-Nearest Neighbour Classifiers-A Tutorial. pages 1–25.

Genkin, A., Lewis, D. D., and Madigan, D. (2007). Large-scale Bayesian Logistic Regression for Text Categorization. pages 291–304.

Hande, A., Priyadharshini, R., and Chakravarthi, B. R. (2020). KanCMD: Kannada CodeMixed Dataset for Sentiment Analysis and Offensive Language Detection. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 54–63.

Kannadaguli, P. (2021). A Code-Diverse Tulu-English Dataset For NLP Based Sentiment Analysis Applications. In *2021 Advanced Communication Technologies and Signal Processing (ACTS)*, pages 1–6.

Krippendorff, K. (2011). Computing Krippendorff's alpha-reliability.

Kusampudi, S. S. V., Chaluvadi, A., and Mamidi, R. (2021). Corpus Creation and Language Identification in Low-Resource Code-Mixed Telugu-English Text. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 744–752.

Mohammad, S. (2016). A Practical Guide to Sentiment Annotation: Challenges and Solutions. In *Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 174–179.

Pranckevičius, T. and Marcinkevičius, V. (2017). Comparison of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification. pages 209–221.

Priyadharshini, R., Chakravarthi, B. R., Thavareesan, S., Chinnappa, D., Thenmozhi, D., and Ponnusamy, R. (2021). Overview of the DravidianCodeMix 2021 Shared Task on Sentiment Detection in Tamil, Malayalam, and Kannada. In *Forum for Information Retrieval Evaluation*, pages 4–6.

Scotton, C. M. (1982). The Possibility of Code-Switching: Motivation for Maintaining Multilingualism. pages 432–444.

Shah, K., Patel, H., Sanghvi, D., and Shah, M. (2020). A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification. *Augmented Human Research*, pages 1–16.

Shetty, M. (2004). Language Contact and the Maintenance of the Tulu Language in South India.

Suryawanshi, S., Chakravarthi, B. R., Verma, P., Arcan, M., McCrae, J. P., and Buitelaar, P. (2020). A Dataset for Troll Classification of TamilMemes. In *Proceedings of the WILDRE5–5th workshop on indian language data: resources and evaluation*, pages 7–13.

Tong, S. and Koller, D. (2001). Support Vector Machine Active Learning with Applications to Text Classification. pages 45–66.

Xu, S., Li, Y., and Wang, Z. (2017). Bayesian Multinomial Naïve Bayes Classifier to Text Classification. In *Advanced multimedia and ubiquitous engineering*, pages 347–352.