Proceedings of the 1st Joint SLTU and CCURL Workshop (SLTU-CCURL 2020), pages 202–210
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Corpus Creation for Sentiment Analysis in Code-Mixed Tamil-English Text

Bharathi Raja Chakravarthi[1]**,Vigneshwaran Muralidaran**[2]**,**
Ruba Priyadharshini[3]**, John P. McCrae**[1]
[1]Insight SFI Research Centre for Data Analytics, Data Science Institute,
National University of Ireland Galway, {bharathi.raja, john.mccrae}@insight-centre.org
[2]School of English, Communication and Philosophy, Cardiff University, muralidaranV@cardiff.ac.uk
[3]Saraswathi Narayanan College, Madurai, India, rubapriyadharshini.a@gmail.com

Abstract

Understanding the sentiment of a comment from a video or an image is an essential task in many applications. Sentiment analysis of a text can be useful for various decision-making processes. One such application is to analyse the popular sentiments of videos on social media based on viewer comments. However, comments from social media do not follow strict rules of grammar, and they contain mixing of more than one language, often written in non-native scripts. Non-availability of annotated code-mixed data for a low-resourced language like Tamil also adds difficulty to this problem. To overcome this, we created a gold standard Tamil-English code-switched, sentiment-annotated corpus containing 15,744 comment posts from YouTube. In this paper, we describe the process of creating the corpus and assigning polarities. We present inter-annotator agreement and show the results of sentiment analysis trained on this corpus as a benchmark.

Keywords: code mixed, Tamil, sentiment, corpus, dataset

1. Introduction

Sentiment analysis has become important in social media research (Yang and Eisenstein, 2017). Until recently these applications were created for high-resourced languages which analysed monolingual utterances. But social media in multilingual communities contains more code-mixed text (Barman et al., 2014; Chanda et al., 2016; Pratapa et al., 2018a; Winata et al., 2019a). Our study focuses on sentiment analysis in Tamil, which has little annotated data for code-mixed scenarios (Phani et al., 2016; Jose et al., 2020). Features based on the lexical properties such as a dictionary of words and parts of speech tagging have less performance compared to the supervised learning (Kannan et al., 2016) approaches using annotated data. However, an annotated corpus developed for monolingual data cannot deal with code-mixed usage and therefore it fails to yield good results (AlGhamdi et al., 2016; Aguilar et al., 2018) due to mixture of languages at different levels of linguistic analysis.

Code-mixing is common among speakers in a bilingual speech community. As English is seen as the language of prestige and education, the influence of lexicon, connectives and phrases from English language is common in spoken Tamil. It is largely observed in educated speakers although not completely absent amongst less educated and uneducated speakers (Krishnasamy, 2015). Due to their pervasiveness of English online, code-mixed Tamil-English (Tanglish) sentences are often typed in Roman script (Suryawanshi et al., 2020a; Suryawanshi et al., 2020b).

We present TamilMixSentiment [1], a dataset of YouTube video comments in Tanglish. TamilMixSentiment was developed with guidelines following the work of Mohammad (2016) and without annotating the word level language tag. The instructions enabled light and speedy annotation while maintaining consistency. The overall inter-annotator agreement in terms of Kripendorffs's α (Krippendorff, 1970) stands at 0.6. In total, 15,744 comments were annotated; this makes the largest general domain sentiment dataset for this relatively low-resource language with code-mixing phenomenon.

We observed all the three types of code-mixed sentences -- Inter-Sentential switch, Intra-Sentential switch and Tag switching. Most comments were written in Roman script with either Tamil grammar with English lexicon or English grammar with Tamil lexicon. Some comments were written in Tamil script with English expressions in between. The following examples illustrate the point.

- **Intha padam vantha piragu yellarum Thala ya kondaduvanga.** - *After the movie release, everybody will celebrate the hero.* Tamil words written in Roman script with no English switch.
- **Trailer late ah parthavanga like podunga.** - *Those who watched the trailer late, please like it.* Tag switching with English words.
- **Omg .. use head phones. Enna bgm da saami ..** -- *OMG! Use your headphones. Good Lord, What a background score!* Inter-sentential switch
- **I think sivakarthickku hero getup set aagala.** - *I think the hero role does not suit Sivakarthick.* Intra-sentential switch between clauses.

In this work we present our dataset, annotation scheme and investigate the properties and statistics of the dataset and information about the annotators. We also present baseline classification results on the new dataset with ten

[1]https://github.com/bharathichezhiyan/TamilMixSentiment

models to establish a baseline for future comparisons. The best results were achieved with models that use logistic regression and random forest.

The contribution of this paper is two-fold:

1. We present the first gold standard code-mixed Tamil-English dataset annotated for sentiment analysis.

2. We provide an experimental analysis of logistic regression, naive Bayes, decision tree, random forest, SVM, dynamic meta-embedding, contextualized dynamic meta-embedding, 1DConv-LSTM and BERT on our code-mixed data for sentiment classification.

2. Related Work

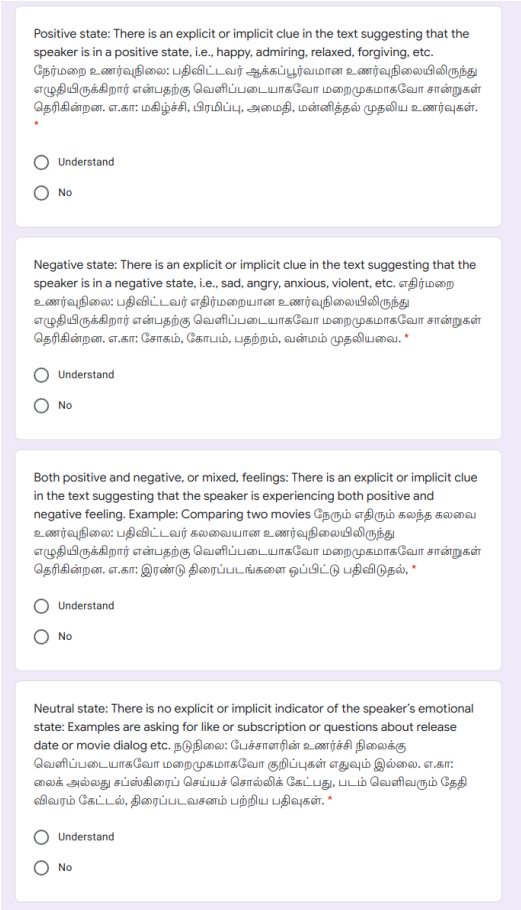
Recently, there has been a considerable amount of work and effort to collect resources for code-switched text. However, code-switched datasets and lexicons for sentiment analysis are still limited in number, size and availability. For monolingual analysis, there exist various corpora for English (Hu and Liu, 2004; Wiebe et al., 2005; Jiang et al., 2019), Russian (Rogers et al., 2018), German (Cieliebak et al., 2017), Norwegian (Mæhlum et al., 2019) and Indian languages (Agrawal et al., 2018; Rani et al., 2020).

When it comes to code-mixing, an English-Hindi corpus was created by (Sitaram et al., 2015; Joshi et al., 2016; Patra et al., 2018), an English-Spanish corpus was introduced by (Solorio et al., 2014; Vilares et al., 2015; Vilares et al., 2016), and a Chinese-English one (Lee and Wang, 2015) was collected from Weibo.com and English-Bengali data were released by Patra et al. (Patra et al., 2018).

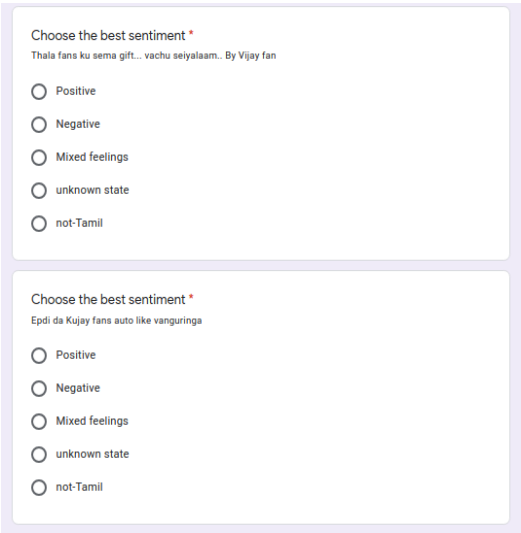
Tamil is a Dravidian language spoken by Tamil people in India, Sri Lanka and by the Tamil diaspora around the world, with official recognition in India, Sri Lanka and Singapore (Chakravarthi et al., 2018; Chakravarthi et al., 2019a; Chakravarthi et al., 2019b; Chakravarthi et al., 2019c). Several research activities on sentiment analysis in Tamil (Padmamala and Prema, 2017) and other Indian languages (Ranjan et al., 2016; Das and Bandyopadhyay, 2010; A.R. et al., 2012; Phani et al., 2016; Prasad et al., 2016; Priyadharshini et al., 2020; Chakravarthi et al., 2020) are happening because the sheer number of native speakers are a potential market for commercial NLP applications. However, sentiment analysis on Tamil-English code-mixed data (Patra et al., 2018) is under-developed and data tare not readily available for research.

Until recently, word-level annotations were used for research in code-mixed corpora. Almost all the previous systems proposed were based on data annotated at the word-level. This is not only time-consuming but also expensive to create. However, neural networks and meta-embeddings (Kiela et al., 2018) have shown great promise in code-switched research without the need for word-level annotation. In particular, work by Winata et al. (2019a) learns to utilise information from pre-trained embeddings without explicit word-level language tags. A recent work by Winata et al. (2019b) utilised the subword-level information from closely related languages to improve the performance on the code-mixed text.

As there was no previous dataset available for Tamil-English (Tanglish) sentiment annotation, we create a sentiment dataset for Tanglish with voluntary annotators. We also show the baseline results with a few models explained in Section 5.



(a) Example 1



(b) Example 2

Figure 1: Examples of Google Form.

3. Corpus Creation and Annotation

Our goal was to create a code-mixed dataset for Tamil to ensure that enough data are available for research purposes. We used the *YouTube Comment Scraper tool*[2] and collected 184,573 sentences for Tamil from YouTube comments. We collected the comments from the trailers of a movies released in 2019. Many of the them contained sentences that were either entirely written in English or code-mixed Tamil-English or fully written in Tamil. So we filtered out a non-code-mixed corpus based on language identification at comment level using the *langdetect library* [3]. Thus if the comment is written fully in Tamil or English, we discarded that comment since monolingual resources are available for these languages. We also identified if the sentences were written in other languages such as Hindi, Malayalam, Urdu, Telugu, and Kannada. We preprocessed the comments by removing the emoticons and applying a sentence length filter. We want to create a code-mixed corpus of reasonable size with sentences that have fairly defined sentiments which will be useful for future research. Thus our filter removed sentences with less than five words and more than 15 words after cleaning the data. In the end we got 15,744 Tanglish sentences.

3.1. Annotation Setup

For annotation, we adopted the approach taken by Mohammad (2016), and a minimum of three annotators annotated each sentence in the dataset according to the following schema shown in the Figure 1. We added new category **Other language:** If the sentence is written in some other language other than Tamil or English. Examples for this are the comments written in other Indian languages using the Roman script. The annotation guidelines are given in English and Tamil.

As we have collected data from YouTube we anonymized to keep the privacy of the users who commented on it. As the voluntary annotators' personal information were collected to know about the them, this gives rise to both ethical, privacy and legal concerns. Therefore, the annotators were informed in the beginning that their data is being recorded and they can choose to withdraw from the process at any stage of annotation. The annotators should actively agree to being recorded. We created Google Forms in which we collected the annotators' email addresses which we used to ensure that an annotator was allowed to label a given sentence only once. We collected the information on gender, education and medium of instruction in school to know the diversity of annotators. Each Google form has been set to contain a maximum of 100 sentences. Example of the Google form is given in the Figure 1. The annotators have to agree that they understood the scheme; otherwise, they cannot proceed further. Three steps complete the annotation setup. First, each sentence was annotated by two people. In the second step, the data were collected if both of them agreed. In the case of conflict, a third person annotated the sentence. In the third step, if all the three of them did not agree, then two more annotators annotated the sentences.

Gender	Male	9
	Female	2
Higher Education	Undegraduate	2
	Graduate	2
	Postgraduate	7
Medium of Schooling	English	6
	Tamil	5
Total		11

Table 1: Annotators

3.2. Annotators

To control the quality of annotation, we removed the annotator who did not annotate well in the first form. For example, if the annotators showed unreasonable delay in responding or if they labelled all sentences with the same sentiment or if more than fifty annotations in a form were wrong, we removed those contributions. Eleven volunteers were involved in the process. All of them were native speakers of Tamil with diversity in gender, educational level and medium of instruction in their school education. Table 1 shows information about the annotators. The volunteers were instructed to fill up the Google form, and 100 sentences were sent to them. If an annotator offers to volunteer more, the next Google form is sent to them with another set of 100 sentences and in this way each volunteer chooses to annotate as many sentences from the corpus as they want. We send the forms to an equal number of male and female annotators. However, from Table 1, we can see that only two female annotators volunteered to contribute.

3.3. Corpus Statistics

Corpus statistics is given in the Table 2. The distribution of released data is shown in Table 3. The entire dataset of 15,744 sentences was randomly shuffled and split into three parts as follows: 11,335 sentences were used for training, 1,260 sentences form the validation set and 3,149 sentences were used for testing. The machine learning models were applied to this subset of data rather than k-fold cross validation. The only other code-mixed dataset of reasonable size that we could find was an earlier work by Remmiya Devi et al. (2016) on code-mix entity extraction for Hindi-English and Tamil-English tweets, released as a part of the shared task in FIRE 2016. The dataset consisted of 3,200 Tanglish tweets used for training and 1,376 tweets for testing.

3.4. Inter Annotator Agreement

We used **Krippendorff's alpha** (α) (Krippendorff, 1970) to measure inter-annotator agreement because of the nature of our annotation setup. This is a robust statistical measure that accounts for incomplete data and, therefore, does not require every annotator to annotate every sentence. It is also a measure that takes into account the degree of disagreement between the predicted classes, which is crucial in our annotation scheme. For instance, if the annotators disagree

[2]https://github.com/philbot9/youtube-comment-scraper
[3]https://pypi.org/project/langdetect/

Language pair	Tamil-English
Number of Tokens	169,833
Vocabulary Size	30,898
Number of Posts	15,744
Number of Sentences	17,926
Average number of Tokens per post	10
Average number of sentences per post	1

Table 2: Corpus statistic of and Tamil-English

Class	Tamil-English
Positive	10,559
Negative	2,037
Mixed feelings	1,801
Neutral	850
Other language	497
Total	15,744

Table 3: Data Distribution

between **Positive** and **Negative** class, this disagreement is more serious than when they disagree between **Mixed feelings** and **Neutral**. α can handle such disagreements. α is defined as:

$$\alpha = 1 - \frac{D_o}{D_e} \quad (1)$$

D_o is the observed disagreement between sentiment labels by the annotators and D_e is the disagreement expected when the coding of sentiments can be attributed to chance rather than due to the inherent property of the sentiment itself.

$$D_o = \frac{1}{n} \sum_c \sum_k o_{ck\ metric}\ \delta^2_{ck} \quad (2)$$

$$D_e = \frac{1}{n(n-1)} \sum_c \sum_k n_c \,.\, n_{k\ metric}\ \delta^2_{ck} \quad (3)$$

Here o_{ck} n_c n_k and n refer to the frequencies of values in coincidence matrices and $metric$ refers to any metric or level of measurement such as nominal, ordinal, interval, ratio and others. Krippendorff's alpha applies to all these metrics. We used nominal and interval metric to calculate annotator agreement. The range of α is between 0 and 1, $1 \geq \alpha \geq 0$. When α is 1 there is perfect agreement between annotators and when 0 the agreement is entirely due to chance. Our annotation produced an agreement of 0.6585 using nominal metric and 0.6799 using interval metric.

4. Difficult Examples

In this section we talk about some examples that were difficult to annotate.

1. **Enakku iru mugan trailer gnabagam than varuthu** - *All it reminds me of is the trailer of the movie Irumugan*. Not sure whether the speaker enjoyed Irumugan trailer or disliked it or simply observed the similarities between the two trailers.

2. **Rajini ah vida akshay mass ah irukane** - *Akshay looks more amazing than Rajini*. Difficult to decide if it is a disappointment that the villain looks better than the hero or a positive appreciation for the villain actor.

3. **Ada dei nama sambatha da dei** - *I wonder, Is this our sampath? Hey!*. Conflict between neutral and positive.

4. **Lokesh kanagaraj movie naalae.... English Rap....Song vandurum** - *If it is a movie of Lokesh kanagaraj, it always has an English rap song*. Ambiguous sentiment.

According to the instructions, questions about music director, movie release date and remarks about when the speaker is watching the video should be treated as neutral. However the above examples show that some comments about the actors and movies can be ambiguously interpreted as neutral or positive or negative. We found annotator disagreements in such sentences.

5. Benchmark Systems

In order to provide a simple baseline, we applied various machine learning algorithms for determining the sentiments of YouTube posts in code-mixed Tamil-English language.

5.1. Experimental Settings

5.1.1. Logistic Regression (LR):

We evaluate the Logistic Regression model with L2 regularization. The input features are the Term Frequency Inverse Document Frequency (TF-IDF) values of up to 3 grams.

5.1.2. Support Vector Machine (SVM):

We evaluate the SVM model with L2 regularization. The features are the same as in LR. The purpose of SVM classification algorithm is to define optimal hyperplane in N dimensional space to separate the data points from each other.

5.1.3. K-Nearest Neighbour (K-NN):

We use KNN for classification with 3,4,5,and 9 neighbours by applying uniform weights.

5.1.4. Decision Tree (DT):

Decision trees have been previously used in NLP tasks for classification. In decision tree, the prediction is done by splitting the root training set into subsets as nodes, and each node contains output of the decision, label or condition. After sequentially choosing alternative decisions, each node

Classifier	Positive	Negative	Neutral	Mixed	Other language	Micro Avg	Macro Avg	Weighted Avg
KNN	0.70	0.23	0.35	0.16	0.06	0.45	0.30	0.53
Decision Tree	0.71	0.30	0.24	0.17	0.60	0.61	0.40	0.56
Random Forest	0.69	0.51	0.80	0.41	0.68	0.68	0.62	0.63
Logistic Regression	0.68	0.56	0.61	0.36	0.76	0.68	0.59	0.62
Naive Bayes	0.66	0.62	0.00	0.40	0.69	0.66	0.48	0.59
SVM	0.66	0.00	0.00	0.00	0.00	0.66	0.13	0.43
1DConv-LSTM	0.71	0.30	0.00	0.14	0.67	0.63	0.36	0.54
DME	0.68	0.34	0.31	0.29	0.71	0.67	0.46	0.57
CDME	0.67	0.56	0.56	0.20	0.68	0.67	0.53	0.59
BERT Multilingual	0.67	0.00	0.00	0.00	0.64	0.67	0.26	0.46

Table 4: Precision

Classifier	Positive	Negative	Neutral	Mixed	Other language	Micro Avg	Macro Avg	Weighted Avg
KNN	0.63	0.04	0.10	0.02	0.61	0.45	0.28	0.45
Decision Tree	0.83	0.21	0.13	0.12	0.54	0.61	0.36	0.61
Random Forest	0.98	0.18	0.09	0.04	0.55	0.68	0.32	0.68
Logistic Regression	0.98	0.13	0.06	0.01	0.32	0.68	0.30	0.68
Naive Bayes	1.00	0.01	0.00	0.01	0.18	0.66	0.24	0.67
SVM	1.00	0.00	0.00	0.00	0.00	0.66	0.20	0.66
1DConv-LSTM	0.91	0.11	0.00	0.10	0.28	0.63	0.28	0.63
DME	0.99	0.03	0.02	0.01	0.49	0.67	0.31	0.57
CDME	0.99	0.01	0.03	0.00	0.52	0.67	0.31	0.67
BERT Multilingual	0.99	0.00	0.00	0.00	0.58	0.67	0.31	0.46

Table 5: Recall

Classifier	Positive	Negative	Neutral	Mixed	Other language	Micro Avg	Macro Avg	Weighted Avg
KNN	0.66	0.06	0.15	0.04	0.10	0.45	0.29	0.50
Decision Tree	0.77	0.24	0.17	0.14	0.54	0.61	0.38	0.58
Random Forest	0.81	0.18	0.09	0.04	0.55	0.68	0.42	0.65
Logistic Regression	0.81	0.21	0.12	0.03	0.45	0.68	0.40	0.64
Naive Bayes	0.80	0.02	0.00	0.01	0.29	0.66	0.32	0.63
SVM	0.79	0.00	0.00	0.00	0.00	0.66	0.16	0.52
1DConv-LSTM	0.80	0.16	0.00	0.12	0.39	0.63	0.31	0.58
DME	0.80	0.05	0.04	0.01	0.58	0.67	0.37	0.57
CDME	0.80	0.02	0.05	0.01	0.59	0.67	0.39	0.63
BERT Multilingual	0.80	0.00	0.00	0.00	0.61	0.67	0.28	0.46

Table 6: F-score

recursively is split again and finally the classifier defines some rules to predict the result. We used it to classify the sentiments for baseline. Maximum depth was 800 and minimum sample splits were 5 for DT. The criterion were Gini and entropy.

5.1.5. Random Forest (RF):

In random forest, the classifier randomly generates trees without defining rules. We evaluate the RF model with same features as in DT.

5.1.6. Multinominal Naive Bayes (MNB):

Naive-Bayes classifier is a probabilistic model, which is derived from Bayes Theorem that finds the probability of hypothesis activity to the given evidence activity. We evaluate the MNB model with our data using α=1 with TF-IDF vectors.

5.1.7. 1DConv-LSTM:

The model we evaluated consists of Embedding layer, Dropout, 1DConv with activation ReLU, Max-pooling and LSTM. The embeddings are randomly initialized.

5.1.8. BERT-Multilingual:

Devlin et al. (2019) introduced a language representation model which is Bidirectional Encoder Representation from Transforms. It is designed to pre-train from unlabelled text and can be fine-tuned by adding last layer. BERT has been used for many text classification tasks (Tayyar Madabushi et al., 2019; Ma et al., 2019; Cohan et al., 2019). We explore classification of a code-mixed data into their corresponding sentiment categories.

5.1.9. DME and CDME:

We also implemented the Dynamic Meta Embedding (Kiela et al., 2018) to evaluate our model. As a first step, we used Word2Vec and FastText to train from our dataset since dy-

namic meta-embedding is an effective method for the supervised learning of embedding ensembles.

5.2. Experiment Results and Discussion

The experimental results of the sentiment classification task using different methods are shown in terms of precision in Table 4, recall in Table 5, and F-score in Table 6. We used *sklearn* [4] for evaluation. The micro-average is calculated by aggregating the contributions of all classes to compute the average metric. In a multi-class classification setup, micro-average is preferable if there are class imbalances. For instance in our data, we have many more examples of positive classes than other classes. A macro-average will compute the metrics (precision, recall, F-score) independently for each class and then take the average. Thus this metric treats all classes equally and it does not take imbalance into account. A weighted average takes the metrics from each class just like macro but the contribution of each class to the average is weighted by the number of examples available for it. For our test, positive is 2,075, negative is 424, neutral is 173, mixed feelings are 377, and non-Tamil is 100.

As shown in the tables, all the classification algorithms perform poorly on the code-mixed dataset. Logistic regression, random forest classifiers and decision trees were the ones that fared comparatively better across all sentiment classes. Surprisingly, the classification result by the SVM model has much worse diversity than the other methods. Applying deep learning methods also does not lead to higher scores on the three automatic metrics. We think this stems from the characteristics of the dataset. The classification scores for different sentiment classes appear to be in line with the distribution of sentiments in the dataset.

The dataset is not a balanced distribution. Table 3 shows that out of total 15,744 sentences 67% belong to *Positive* class while the other sentiment classes share 13%, 5% and 3% respectively. The precision, recall and F-measure scores are higher for the *Positive* class while the scores for *Neutral* and *Mixed feeling* classes were disastrous. Apart from their low distribution in the dataset, these two classes are difficult to annotate for even human annotators as discussed in Section 4. In comparison, the *Negative* and *Other language* classes were better. We suspect this is due to more explicit clues for negative and non-Tamil words and due to relatively higher distribution of negative comments in the data.

Since we collected the post from movie trailers, we got more positive sentiment than others as the people who watch trailers are more likely to be interested in movies and this skews the overall distribution. However, as the code-mixing phenomenon is not incorporated in the earlier models, this resource could be taken as a starting point for further research. There is significant room for improvement in code-mixed research with our dataset. In our experiments, we only utilized the machine learning methods, but more information such as linguistic information or hierarchical meta-embedding can be utilized. This dataset can be used to create a multilingual embedding for code-mixed data (Pratapa et al., 2018b).

[4] https://scikit-learn.org/

6. Conclusion

We presented, to the best of our knowledge, the most substantial corpus for under-resourced code-mixed Tanglish with annotations for sentiment polarity. We achieved a high inter-annotator agreement in terms of Krippendorff α from voluntary annotators on contributions collected using Google form. We created baselines with gold standard annotated data and presented our results for each class in Precision, Recall, and F-Score. We expect this resource will enable the researchers to address new and exciting problems in code-mixed research.

7. Acknowledgments

This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289 (Insight), SFI/12/RC/2289_P2 (Insight_2), co-funded by the European Regional Development Fund as well as by the EU H2020 programme under grant agreements 731015 (ELEXIS-European Lexical Infrastructure), 825182 (Prêt-à-LLOD), and Irish Research Council grant IRCLA/2017/129 (CARDAMOM-Comparative Deep Models of Language for Minority and Historical Languages).

8. Bibliographical References

Agrawal, R., Chenthil Kumar, V., Muralidharan, V., and Sharma, D. (2018). No more beating about the bush : A step towards idiom handling for Indian language NLP. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Gustavo Aguilar, et al., editors. (2018). *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, Melbourne, Australia, July. Association for Computational Linguistics.

AlGhamdi, F., Molina, G., Diab, M., Solorio, T., Hawwari, A., Soto, V., and Hirschberg, J. (2016). Part of speech tagging for code switched data. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 98–107, Austin, Texas, November. Association for Computational Linguistics.

A.R., B., Joshi, A., and Bhattacharyya, P. (2012). Cross-lingual sentiment analysis for Indian languages using linked WordNets. In *Proceedings of COLING 2012: Posters*, pages 73–82, Mumbai, India, December. The COLING 2012 Organizing Committee.

Barman, U., Das, A., Wagner, J., and Foster, J. (2014). Code mixing: A challenge for language identification in the language of social media. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 13–23, Doha, Qatar, October. Association for Computational Linguistics.

Chakravarthi, B. R., Arcan, M., and McCrae, J. P. (2018). Improving wordnets for under-resourced languages using machine translation. In *Proceedings of the 9th Global WordNet Conference (GWC 2018)*, page 78.

Chakravarthi, B. R., Arcan, M., and McCrae, J. P. (2019a). Comparison of different orthographies for machine translation of under-resourced dravidian languages. In *2nd Conference on Language, Data and Knowledge (LDK 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Chakravarthi, B. R., Arcan, M., and McCrae, J. P. (2019b). WordNet gloss translation for under-resourced languages using multilingual neural machine translation. In *Proceedings of the Second Workshop on Multilingualism at the Intersection of Knowledge Bases and Machine Translation*, pages 1–7, Dublin, Ireland, 19 August. European Association for Machine Translation.

Chakravarthi, B. R., Priyadharshini, R., Stearns, B., Jayapal, A., S, S., Arcan, M., Zarrouk, M., and McCrae, J. P. (2019c). Multilingual multimodal machine translation for Dravidian languages utilizing phonetic transcription. In *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*, pages 56–63, Dublin, Ireland, 20 August. European Association for Machine Translation.

Chakravarthi, B. R., Jose, N., Suryawanshi, S., Sherly, E., and McCrae, J. P. (2020). A sentiment analysis dataset for code-mixed Malayalam-English. In *Proceedings of the 1st Joint Workshop of SLTU (Spoken Language Technologies for Under-resourced languages) and CCURL (Collaboration and Computing for Under-Resourced Languages) (SLTU-CCURL 2020)*, Marseille, France, May. European Language Resources Association (ELRA).

Chanda, A., Das, D., and Mazumdar, C. (2016). Unraveling the English-Bengali code-mixing phenomenon. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 80–89, Austin, Texas, November. Association for Computational Linguistics.

Cieliebak, M., Deriu, J. M., Egger, D., and Uzdilli, F. (2017). A Twitter corpus and benchmark resources for German sentiment analysis. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 45–51, Valencia, Spain, April. Association for Computational Linguistics.

Cohan, A., Beltagy, I., King, D., Dalvi, B., and Weld, D. (2019). Pretrained language models for sequential sentence classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3693–3699, Hong Kong, China, November. Association for Computational Linguistics.

Das, A. and Bandyopadhyay, S. (2010). SentiWordNet for Indian languages. In *Proceedings of the Eighth Workshop on Asian Language Resouces*, pages 56–63, Beijing, China, August. Coling 2010 Organizing Committee.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, page 168–177, New York, NY, USA. Association for Computing Machinery.

Jiang, Q., Chen, L., Xu, R., Ao, X., and Yang, M. (2019). A challenge dataset and effective models for aspect-based sentiment analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6279–6284, Hong Kong, China, November. Association for Computational Linguistics.

Jose, N., Chakravarthi, B. R., Suryawanshi, S., Sherly, E., and McCrae, J. P. (2020). A survey of current datasets for code-switching research. In *2020 6th International Conference on Advanced Computing & Communication Systems (ICACCS)*.

Joshi, A., Prabhu, A., Shrivastava, M., and Varma, V. (2016). Towards sub-word level compositions for sentiment analysis of Hindi-English code mixed text. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2482–2491, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Kannan, A., Mohanty, G., and Mamidi, R. (2016). Towards building a SentiWordNet for Tamil. In *Proceedings of the 13th International Conference on Natural Language Processing*, pages 30–35, Varanasi, India, December. NLP Association of India.

Kiela, D., Wang, C., and Cho, K. (2018). Dynamic meta-embeddings for improved sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1466–1477, Brussels, Belgium, October-November. Association for Computational Linguistics.

Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30(1):61–70.

Krishnasamy, K. (2015). Code mixing among Tamil-English bilingual children. *International Journal of Social Science and Humanity*, 5(9):788.

Lee, S. and Wang, Z. (2015). Emotion in code-switching texts: Corpus construction and analysis. In *Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing*, pages 91–99, Beijing, China, July. Association for Computational Linguistics.

Ma, X., Xu, P., Wang, Z., Nallapati, R., and Xiang, B. (2019). Domain adaptation with BERT-based domain classification and data selection. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 76–83, Hong

Kong, China, November. Association for Computational Linguistics.

Mæhlum, P., Barnes, J., Øvrelid, L., and Velldal, E. (2019). Annotating evaluative sentences for sentiment analysis: a dataset for Norwegian. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 121–130, Turku, Finland, September–October. Linköping University Electronic Press.

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, San Diego, California, June. Association for Computational Linguistics.

Padmamala, R. and Prema, V. (2017). Sentiment analysis of online Tamil contents using recursive neural network models approach for Tamil language. In *2017 IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM)*, pages 28–31, Aug.

Patra, B. G., Das, D., and Das, A. (2018). Sentiment analysis of code-mixed indian languages: An overview of sail_code-mixed shared task@ icon-2017. *arXiv preprint arXiv:1803.06745*.

Phani, S., Lahiri, S., and Biswas, A. (2016). Sentiment analysis of Tweets in three Indian languages. In *Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016)*, pages 93–102, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Prasad, S. S., Kumar, J., Prabhakar, D. K., and Tripathi, S. (2016). Sentiment mining: An approach for Bengali and Tamil tweets. In *2016 Ninth International Conference on Contemporary Computing (IC3)*, pages 1–4, Aug.

Pratapa, A., Bhat, G., Choudhury, M., Sitaram, S., Dandapat, S., and Bali, K. (2018a). Language modeling for code-mixing: The role of linguistic theory based synthetic data. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1543–1553, Melbourne, Australia, July. Association for Computational Linguistics.

Pratapa, A., Choudhury, M., and Sitaram, S. (2018b). Word embeddings for code-mixed language processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3067–3072, Brussels, Belgium, October-November. Association for Computational Linguistics.

Priyadharshini, R., Chakravarthi, B. R., Vegupatti, M., and McCrae, J. P. (2020). Named entity recognition for code-mixed Indian corpus using meta embedding. In *2020 6th International Conference on Advanced Computing & Communication Systems (ICACCS)*.

Rani, P., Suryawanshi, S., Goswami, K., Chakravarthi, B. R., Fransen, T., and McCrae, J. P. (2020). A comparative study of different state-of-the-art hate speech detection methods for Hindi-English code-mixed data. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, Marseille, France, May. European Language Resources Association (ELRA).

Ranjan, P., Raja, B., Priyadharshini, R., and Balabantaray, R. C. (2016). A comparative study on code-mixed data of Indian social media vs formal text. In *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*, pages 608–611.

Remmiya Devi, G., Veena, P., Anand Kumar, M., and Soman, K. (2016). Amrita-cen@ fire 2016: Code-mix entity extraction for Hindi-English and Tamil-English tweets. In *CEUR workshop proceedings*, volume 1737, pages 304–308.

Rogers, A., Romanov, A., Rumshisky, A., Volkova, S., Gronas, M., and Gribov, A. (2018). RuSentiment: An enriched sentiment analysis dataset for social media in Russian. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 755–763, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Sitaram, D., Murthy, S., Ray, D., Sharma, D., and Dhar, K. (2015). Sentiment analysis of mixed language employing hindi-english code switching. In *2015 International Conference on Machine Learning and Cybernetics (ICMLC)*, volume 1, pages 271–276, July.

Solorio, T., Blair, E., Maharjan, S., Bethard, S., Diab, M., Ghoneim, M., Hawwari, A., AlGhamdi, F., Hirschberg, J., Chang, A., and Fung, P. (2014). Overview for the first shared task on language identification in code-switched data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 62–72, Doha, Qatar, October. Association for Computational Linguistics.

Suryawanshi, S., Chakravarthi, B. R., Arcan, M., and Buitelaar, P. (2020a). Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, Marseille, France, May. European Language Resources Association (ELRA).

Suryawanshi, S., Chakravarthi, B. R., Verma, P., Arcan, M., McCrae, J. P., and Buitelaar, P. (2020b). A dataset for troll classification of Tamil memes. In *Proceedings of the 5th Workshop on Indian Language Data Resource and Evaluation (WILDRE-5)*, Marseille, France, May. European Language Resources Association (ELRA).

Tayyar Madabushi, H., Kochkina, E., and Castelle, M. (2019). Cost-sensitive BERT for generalisable sentence classification on imbalanced data. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 125–134, Hong Kong, China, November. Association for Computational Linguistics.

Vilares, D., Alonso, M. A., and Gómez-Rodríguez, C. (2015). Sentiment analysis on monolingual, multilingual and code-switching Twitter corpora. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 2–8, Lisboa, Portugal, September. Association for Computational Linguistics.

Vilares, D., Alonso, M. A., and Gómez-Rodríguez, C. (2016). En-es-cs: An English-Spanish code-switching twitter corpus for multilingual sentiment analysis. In Nicoletta Calzolari (Conference Chair), et al., edi-

tors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Wiebe, J., Wilson, T., and Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2):165–210, May.

Winata, G. I., Lin, Z., and Fung, P. (2019a). Learning multilingual meta-embeddings for code-switching named entity recognition. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 181–186, Florence, Italy, August. Association for Computational Linguistics.

Winata, G. I., Lin, Z., Shin, J., Liu, Z., and Fung, P. (2019b). Hierarchical meta-embeddings for code-switching named entity recognition. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3532–3538, Hong Kong, China, November. Association for Computational Linguistics.

Yang, Y. and Eisenstein, J. (2017). Overcoming language variation in sentiment analysis with social attention. *Transactions of the Association for Computational Linguistics*, 5:295–307.