Creation of Corpus and analysis in Code-Mixed Kannada-English Twitter data for Emotion Prediction

Appidi Abhinav Reddy, Vamshi Krishna Srirangam,
Suhas Darsi and Manish Shrivastava
Language Technologies Research Centre (LTRC)
Kohli Centre on Intelligent Systems(KCIS)
International Institute of Information Technology, Hyderabad, India.
(abhinav.appidi, v.srirangam, darsi.suhas)@research.iiit.ac.in
m.shrivastava@iiit.ac.in

Abstract

Emotion prediction is a critical task in the field of Natural Language Processing (NLP). There has been a significant amount of work done in emotion prediction for resource-rich languages. There has been work done on code-mixed social media corpus but not on emotion prediction of Kannada-English code-mixed Twitter data. In this paper, we analyze the problem of emotion prediction on corpus obtained from code-mixed Kannada-English extracted from Twitter annotated with their respective 'Emotion' for each tweet. We experimented with machine learning prediction models using features like Character N-Grams, Word N-Grams, Repetitive characters, and others on SVM and LSTM on our corpus, which resulted in an accuracy of 30% and 32%, respectively.

1 Introduction

Identification and analysis of emotions in user-generated data in social media like Twitter, Facebook, Reddit, etc., is essential in understanding the daily trends and human behavior. Emotion prediction aims at identifying and analyzing such emotions like 'Happy, 'Sad,' 'Angry,' 'Fear,' 'Surprise,' and 'Disgust' types present in the text. Original works were focussed more on monolingual text (Alm et al., 2005; Chen et al., 2010) due to the large-scale availability of monolingual texts.

India has twenty-three significant languages with over seven hundred and twenty dialects. The majority of people are multilingual, and they tend to mix words from different languages in speech and written text. This method of interchanging languages is commonly addressed by terms 'Code-switching' and 'Code-mixing' as described by Lipski (1978). Code-mixing refers to the use of words from different languages in the same sentence. Code-switching refers to the use of words or phrases from different languages within the same speech context.

We can understand the difference between code-switching and code-mixing from the positions of altered elements. Code-mixing refers to the intrasentential modification of codes, whereas code-switching refers to the intersentential modification of codes. We observe code-switching and code-mixing frequently on social media platforms. Since the available resources are limited for Kannada-English code-mixed text, we primarily focus on the creation of corpus and annotating the code-mixed tweets with their respective emotions, in this paper.

Here are some examples from a corpus of code-mixed Kannada-English generated from Twitter data and its translation in English.

T1: *"Nam placement officer helidda ee thara helidre ond company lu kelsa sigalla anta...I had 2 offers before I left college, ondu IT innondu core..."*
Translation: "Our placement officer said,'if you talk like this, you wont get a single job.' I had 2 offers before i left college. One was IT, the other was core."

Proceedings of the 28th International Conference on Computational Linguistics, pages 6703–6709
Barcelona, Spain (Online), December 8-13, 2020

T2: *"Eshwarappa avarey neevu petrol bunk ge hogilla ansuthe. me nimmannu karkondu hogthini"*
Translation: "Eshwarappa, it looks like you did not go to the petrol bunk. I will take you there."

2 Background and Related Work

There has been a plethora of research done on Emotion prediction in resource-rich languages. The same is not true for the Kannada-English code-mixed corpus.

Shambhavi (2012) has done the work on the Kannada POS tagger with probabilistic classifiers. Ketan Kumar (2018) presented work on the Kannada POS tagger using machine learning models, and Amarappa (2013) worked on NER and classification in the Kannada language. The following are some works in code-mixed Indian languages. Antony (2010) worked on a kernel-based POS tagger for Kannada. Lakshmi (2017) presented an automatic identification system for code-mixed Kannada-English Social media text, and Shalini (2018) worked on sentiment analysis for Code-Mixed Kannada-English Social Media Text. Rohini (2016) worked on domain-based Sentiment Analysis in the regional language which is Kannada. Kumar (2015) has worked on the analysis of users' sentiments from Kannada Web Documents. When it comes to Emotion Prediction, we believe the corpus we created is the first Kannada-English code-mixed corpus with Emotion tags.

3 Corpus Creation and Annotation

The corpus created consists of Kannada-English code mixed tweets gathered from twitter using twintproject[1]-an opensource twitter intelligence tool. We have collected tweets from the past 5-8 years based on various topics such as movies, sports, celebrities, politics, trending hashtags, social events, not limited to a particular domain. We can find the topics list in the appendices section of this paper. We have done extensive pre-processing of tweets and retrieved them in JSON format. This JSON formatted data includes metadata like URLs, usernames, retweets, tweet IDs, likes, full names, and others.

Pre-processing:

Below are the steps followed by two annotators for the pre-processing of tweets. The two annotators have a linguistic background and are proficient in both Kannada and English languages.

- Tweets that contain linguistic units from both English and Kannada are considered.
- We removed tweets that contain words only in Kannada or only in English.
- Tweets that consist of a minimum of five or above words only are considered.
- We replaced URLs and links with the 'URL' word and removed multiple spaces as they do not contribute towards emotions in the tweet.
- We removed tweets that do not depict the code-mixing nature predominantly. We deleted tweets that contain only one or two linguistic units like affixes, suffixes, etc. from a different language.

3.1 Annotation and Inter Annotator Agreement

We annotated the Kannada-English code-mixed tweets using six emotions 'Happy,' 'Angry,' 'Sad,' 'Fear,' 'Disgust,' 'Surprise,' and a 'Multiple Emotion' tag if the tweet contains one or more emotions. Two people with linguistic background manually did the annotations of the data for Emotion Prediction, both proficient in Kannada and English. The quality of the annotation is validated using the Inter Annotator Agreement (IAA) between the sets of 6396 tweets using Cohen's Kappa coefficient Hallgren (2012). The agreement is significantly high. Refer to Table 1.

A few examples of Kannada-English tweets depicting the emotions are as follows.

[1]https://github.com/twintproject/twint

T3: *"@VikramBK @acharya2 picture allirodanna emoticon alli tOrsbiTyallappaa.. dhanyanaade!!"*
Translation: "Whatever is in picture..you have depicted in emoticon..I'm blessed!!"

T4: *"appa thande ninu adhe kelsa madapa..national issue adhre nanu donald trump kelabeka..State issues na state nalli mathadabekkappa..Ninage yake ashtu sittu..hucchu gichhu heidare madalu doctor beda sidda na hatira hogu yenne kodisthane.."*
Translation: "Do the given work.. should I ask Donald Trump for a national issue?.. state issues must be spoken in state only.. why are you so hesitant.. If you are doing mad things, doctor is not needed, to go a sidda, he'll give you some oil.."

T5: *"Adre esto kade signaller sigodilla? Complaint madidre bcz of forest area antare!! Landline work agalla ... Kelsa madoke staff iralla !! En madodu"*
Translation: "But at many places we don't even get signal? if we complain they say forest area!! landline doesn't work... no staff to work!! ..what to do"

T6: *"Sir marappa layout side nim beat police avre barola nice underpass thumba danger place agidhe adhu"*
Translation: "Sir towards marappa layout side your beat police only will not come.. nice underpass has become a very dangerous place"

T7: *"3-4 years tym tagondu kelsa madinu promotion madade Nim movie haalumadkotideera guy's ..... seriously promotion madi sariyagi....Dabang nodi 3 months inda promotion madtidare...."*
Translation: "even after taking 3-4 years time to complete your movie...you have spoilt it by not doing its promotion... seriously do the promotion...look at dabang 3 from three months they are promoting"

T8: *"@Suharsh2512 oho, idyaavdo brilliant facility. Nanna phone alli sound barutte.. ondond sala baralla. Hyaage nodu..."*
Translation: "Oho...this is some brilliant facility....in my phone there is sound ..once there is no sound....see how it is"

T9: *"He doesn't represent us.Ond site iskond bittu deshane marbidtira neevu, avamana kanro neevu namge, nachke agutte helkolloke ache. Avara makle hogi bekadre @mepratap vote haktare modi goskara, nim antavaru site duddge yenta neecha kelsa bekadru madtira."*
Translation: "He doesn't represent us. For one site he has sold the entire country. You are an insult to us. I feel ashamed that you're our representative, His children will give @mepratap their vote so that Modi can win. You people will do anything however low for money and land."

The above examples contain both Kannada and English texts. Example **T3** expresses happy through the words 'dhanyanaade' which means 'I'm blessed' and **T4** expresses angry through the phrase 'Ninage yake ashtu sittu' which translates to 'why are you so hesitant?'. Sad is expressed in **T5** through Kannada phrase 'En madodu'. Similarly, Fear can be seen in **T6** with the statement 'thumba danger placeagidhe adhu,' which means 'ice underpass has become a very dangerous place' in English. In **T7**, we can see the emotion Disgust from the context of the given an example and also through the phrase 'Nim movie haalumadkotideera' in Kannada and **T8** depicts Surprise through 'oho, idyaavdo brilliant facility', the word Oho here expresses the emotion in the statement. Multiple emotions can be seen expressed in **T9**, like Disgust and Sad. Disgust is expressed through 'avamana kanro neevunamge' and Sad can be seen through the phrase 'nim antavaru site duddge yenta neecha kelsa bekadru madtira'.

As very few resources are available for code-mixed Kannada-English text, our primary focus in this paper is creating the corpus and annotating associated emotions to the code-mixed tweets. We believe our efforts in creating the annotated corpus will provide extreme value to the researchers working in a similar field.

	Cohen Kappa
Happy	0.92
Angry	0.89
Sad	0.84
Fear	0.83
Disgust	0.82
Surprise	0.89
Multiple Emotion	0.93

Table 1: Inter Annotator Agreement.

Emotion	Sentences
Happy	1257
Angry	1400
Sad	817
Fear	64
Disgust	955
Surprise	89
Multiple Emotions	1814
Total	6396

Table 2: Data Distribution

4 Corpus Statistics

We have collected more than 3,34,600 tweets from Twitter using TwintProject. We obtained 6396 Kannada-English code-mixed tweets after extensive cleaning of the corpus. We made sure that all the words in the corpus are in Roman script. Table 2 shows the distribution of Emotion tags in the code-mixed corpus. We used hashtags related to politics, sports, social events, recent trends and words which depict emotions in Kannada like 'santhoshada', 'amodha' for happy, 'nirase', 'amodha' for sad etc., in collecting the corpus.

We have made language identification for each word to have a better understanding of the corpus, using the tool [2] from the research done by Bhat (2015). We have shown the distribution of words present in the corpus between Kannada and English languages in Table 3 and Table 4, which helps us for a better understanding of code-mixing nature.

Language	Word Count
English	49202
Kannada	106798
Total	156000

Table 3: Total Word Distribution

Language	Unique Words
English	13193
Kannada	30192
Total	15899

Table 4: Unique Word Distribution

5 System Architecture

This section explains the emotion prediction of the annotated corpus in the code-mixed Kannada-English tweets. We performed experiments using machine learning models to classify emotions into happy, angry, sad, fear, disgust, surprise and 'multiple emotion'.

5.1 Feature Identification and Extraction

Here, to train our supervised machine learning models, we have used the following feature vectors.

1. **Character N-Grams:** This is one of the crucial features for classifying texts and is language independent. Character N-Grams helps us in capturing the semantic information as social media texts contain misspellings and informal words that are different from standard English and Kannada words. We used Character N-Grams of size 2 and 3 in order to capture the information in the string.

2. **Word N-Grams:** We use Word N-Grams as a feature in our model, which helps us to capture emotion in a text. These are also called contextual features.

3. **Negation Words:** Negative words always alter the perceived emotion. 'Not happy' depicts sadness, even though it contains the word happy. We take a list of English negation words from Christopher

[2] https://github.com/irshadbhat/litcm

Pott's sentiment tutorial[3]. We make a count of all such terms and use them as a feature. Zhu (2014) worked on the effect of negation words on sentiment.

4. **Punctuation:** Multiple question marks and multiple exclamation marks are used to depict feelings of angry and astonishment, respectively. We count the occurrence of such, in a sentence, and use them as a feature.

5. **Emoticons:** In social media, we use emoticons to express emotions like ':)' for happiness and ':(' to express sadness. We use a list of Western emoticons from Wikipedia [4]. We use count of emoticons for each emotion in each tweet as a feature.

6. **Capitalization:** People often use capital letters to denote anger in social media. We use all such words, count them, and use them as a feature in our experiments.

7. **Repetitive Characters:** Words like 'yayyy,' 'partyyy,' 'lolll,' 'happyyy,' etc. are used in social media to stress an emotion or feeling. If particular characters were repeated more than two times in a row, we make a count of all such words and use it as a feature.

8. **Emotion Words:** From the corpus, we analyzed each emotion tweet and obtained a list of Kannada and English words and used the count of occurrence of each word as a feature. For example words like 'santhoshada', 'yaadha', 'santhushta', 'amodha' and other words are present in the list for 'HAPPY'. Similarly multiple words which depict the emotion for other tags are used.

9. **Intensifiers:** We use a list of intensifiers from Wikipedia[5]. We used this to emphasize emotion or sentiment. For example, in the following phrase, 'nice underpass thumba danger place agidhe adhu' means 'nice underpass has become a very dangerous place'. Here, 'very' is used to emphasize the fear in the statement. English intensifiers were transliterated into Kannada. Kannada words which were used as intensifiers in the corpus was also added to the list.

6 Results and Discussions

We experimented with prediction models, SVM and LSTM model on our corpus.

6.1 SVM

Support vector machines (SVM) are supervised learning models that analyze data used for classification and regression analysis. We performed several experiments using different parameters like RBF, linear kernels, gamma value, regularization parameter. We used SVM classifiers using RBF kernel as they perform efficiently with high dimensional feature vectors. We carried out 5-fold cross-validation. We have used scikit-learn for training our system classifier. With SVM, we had the best accuracy of 30% with RBF kernel and 100 iterations. Table 5 shows the results with the SVM classifier.

6.2 LSTM

Long Short Term Memory (LSTM) is an RNN architecture that is well suited for classification and making predictions based on time series data. LSTM is widely used in many natural language processing applications like classification and language modeling. In our problem of emotion prediction, which is a classification task, the input words are processed by LSTM networks sequentially, and the last output of the LSTM represents the meaning of the sentence.

We performed several experiments using different parameters in LSTM like dropout, loss function, optimizer, activation function, and number of epochs. In the experiments with LSTM, the best F1-score we had is 0.3 and the best accuracy of 32% using 'softmax' as activation function, and 'categorical_crossentropy' as loss function with a dropout of 0.2 for five epochs. The training, validation and testing splits are taken as 70%, 10%, 20% of total data. Table 6 shows the results of the LSTM on the

[3]http://sentiment.christopherpotts.net/lingstruc.html
[4]https://en.wikipedia.org/wiki/List of emoticons
[5]https://en.wikipedia.org/wiki/Intensifier

Feature	Accuracy
Char N-Grams	0.29
Word N-Grams	0.28
Negation words	0.26
Punctuation	0.24
Emoticons	0.23
Capitalization	0.28
Repetitive Characters	0.25
Emotion Words	0.22
Intensifiers	0.25

Table 5: Results of SVM on our corpus

Emotion	precision	recall	f1-score
Happy	0.32	0.44	0.37
Angry	0.33	0.49	0.39
Sad	0.29	0.16	0.20
Fear	0.01	0.01	0.01
Disgust	0.17	0.18	0.17
Surprise	0.05	0.02	0.03
Multiple	0.36	0.21	0.27

Table 6: LSTM Results on all emotions

corpus. The low results of precision, recall and f1-scores for the emotions 'Fear' and 'Surprise' are due to the less number of data samples 64 and 89 respectively.

7 Conclusion and Future Work

Our findings are as follows :

- Presented an annotated code-mixed Kannada-English corpus for Emotion Prediction. The corpus will be published online soon.
- We have experimented with the machine learning models SVM, LSTM, on our data, accuracy for which is 30%, 32%, respectively.
- We have proposed nine handcrafted features which helps us in the capturing of emotion in code-mixed Kannada-English text.
- We are introducing and addressing Emotion Prediction of Kannada-English code-mixed data as a research problem.

For future work, the corpus can be enriched by also giving the respective POS tags for the words. An increase in the size of the corpus helps in more applications of code-mixed Indian languages. We can adapt the problem for Emotion Prediction in code-mixed data containing more than two languages from multilingual societies.

References

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 579–586. Association for Computational Linguistics.

S Amarappa and SV Sathyanarayana. 2013. Named entity recognition and classification in kannada language. *International Journal of Electronics and Computer Science Engineering*, 2(1):281–289.

PJ Antony and KP Soman. 2010. Kernel based part of speech tagger for kannada. In *2010 International Conference on Machine Learning and Cybernetics*, volume 4, pages 2139–2144. IEEE.

Irshad Ahmad Bhat, Vandan Mujadia, Aniruddha Tammewar, Riyaz Ahmad Bhat, and Manish Shrivastava. 2015. Iiit-h system submission for fire2014 shared task on transliterated search. In *Proceedings of the Forum for Information Retrieval Evaluation*, FIRE '14, pages 48–53, New York, NY, USA. ACM.

Shambhavi BR and P Ramakanth Kumar. 2012. Kannada part-of-speech tagging with probabilistic classifiers. *international journal of computer applications*, 48(17):26–30.

Ying Chen, Sophia Yat Mei Lee, Shoushan Li, and Chu-Ren Huang. 2010. Emotion cause detection with linguistic constructions. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 179–187. Association for Computational Linguistics.

Kevin A Hallgren. 2012. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology*, 8(1):23.

KM Anil Kumar, N Rajasimha, Manovikas Reddy, A Rajanarayana, and Kewal Nadgir. 2015. Analysis of users' sentiments from kannada web documents. *Procedia Computer Science*, 54:247–256.

BS Sowmya Lakshmi and BR Shambhavi. 2017. An automatic language identification system for code-mixed english-kannada social media text. In *2017 2nd International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS)*, pages 1–5. IEEE.

John Lipski. 1978. Code-switching and the problem of bilingual competence. *Aspects of bilingualism*, 250:264.

V Rohini, Merin Thomas, and CA Latha. 2016. Domain based sentiment analysis in regional language-kannada using machine learning algorithm. In *2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, pages 503–507. IEEE.

K Shalini, HB Barathi Ganesh, M Anand Kumar, and KP Soman. 2018. Sentiment analysis for code-mixed indian social media text with distributed representation. In *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1126–1131. IEEE.

Ketan Kumar Todi, Pruthwik Mishra, and Dipti Misra Sharma. 2018. Building a kannada pos tagger using machine learning and neural network models. *arXiv preprint arXiv:1808.03175*.

Xiaodan Zhu, Hongyu Guo, Saif Mohammad, and Svetlana Kiritchenko. 2014. An empirical study on the effect of negation words on sentiment. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 304–313.