

# NBLex: emotion prediction in Kannada-English code-switch text using naïve bayes lexicon approach

Ramesh Chundi<sup>1</sup>, Vishwanath R. Hulipalled<sup>2</sup>, Jay Bharthish Simha<sup>3</sup>

<sup>1</sup>School of Computer Science and Applications, REVA University, Bangalore, India

<sup>2</sup>School of Computing and Information Technology, REVA University, Bangalore, India

<sup>3</sup>Abiba Systems, CTO, and RACE Labs, REVA University, Bangalore, India

## Article Info

### Article history:

Received Apr 13, 2022

Revised Sep 21, 2022

Accepted Oct 14, 2022

### Keywords:

Code-switch text

Emotion analysis

Emotion prediction

Lexicon based approach

One-vs-rest

Text analytics

## ABSTRACT

Emotion analysis is a process of identifying the human emotions derived from the various data sources. Emotions can be expressed either in monolingual text or code-switch text. Emotion prediction can be performed through machine learning (ML), or deep learning (DL), or lexicon-based approach. ML and DL approaches are computationally expensive and require training data. Whereas, the lexicon-based approach does not require any training data and it takes very less time to predict the emotions in comparison with ML and DL. In this paper, we proposed a lexicon-based method called non-binding lower extremity exoskeleton (NBLex) to predict the emotions associated with Kannada-English code-switch text that no one has addressed till now. We applied the One-vs-Rest approach to generate the scores for lexicon and also to predict the emotions from the code-switch text. The accuracy of the proposed model NBLex (87.9%) is better than naïve bayes (NB) (85.8%) and bidirectional long short-term memory neural network (BiLSTM) (84.7%) and for true positive rate (TPR), the NBLex (50.6%) is better than NB (37.0%) and BiLSTM (42.2%). From our approach, it is observed that a simple additive model (lexicon approach) can also be an alternative model to predict the emotions in code-switch text.

*This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.*



## Corresponding Author:

Ramesh Chundi

School of Computer Science and Applications, REVA University

Rukmini Knowledge Park, Kattigenahalli, Srinivasa Nagar, Yelahanka, Bangalore-560064, India

Email: chundiramesh@gmail.com

## 1. INTRODUCTION

Many social networking websites like YouTube, Facebook, and Twitter, are available for users to express emotional text content in online every day. Prediction of emotions from these social media text will help us to understand the human emotional behavior towards trends and public events. Emotion prediction is to identifying the emotions like anger, disgust, and joy, from text content. According to Plutchik's wheel of emotions, there are eight basic emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, and trust) and two sentiment classes (positive and negative) [1].

In the recent past lot of work has been carried out on the monolingual text to predict the emotions by several researchers [2], [3] However, most of the internet users in India are multilingual (can read and write more than two languages) or bilingual (at least they can read and write two languages). According to the leading newspaper The times of India report (Nov 7<sup>th</sup>, 2018), 52% of Indians are bilingual and 18% are multilingual [4]. This multilingualism makes the internet users to use different languages while writing the text comments. This type of substituting the languages is known as code-switching or code-mixing [5]. Code-switch is a very common occurrence [6]–[8] and users can write the comments either in monolingual text or

code-switch text in social networking websites. To illustrate, Ex1 refers to monolingual text, where both source and scripting languages are same (English) and Ex2 refers to code-switch text, where the source and scripting languages are different (source is Kannada language and scripting is in the English language).

- Ex1: “*Apj Abdul kalam is the ever best & talented President of India*”
- Ex2: “Avrge enu artha haguthe, bidu guru”
- Translation: “What they can understand, leave it, sir”

Emotions can be expressed either in monolingual text or code-switch text. In Ex1 (monolingual text), user expressed joy emotion and in Ex2 (code-switch text), user expressed sad emotion. Hence, many people nowadays use code-switch text in various situations which is convenient for them. The main advantage of using code-switch text in social networking websites is that, people no need to follow any linguistic and grammatical rules.

Figure 1 shows the different approaches for emotion prediction. There are three approaches to predict the emotions, namely machine learning (ML) approach, the deep learning (DL) approach and lexicon-based approach. Both ML and DL approaches are to perform the prediction task by learning the patterns from the training dataset and test the performance on the test dataset. Support vector machine (SVM) and naïve bayes (NB) are the examples of ML approach and similarly, long short-term memory (LSTM) and bidirectional long short-term memory neural network (BiLSTM) are examples for the DL approaches. Whereas, the lexicon approach performs the prediction task by checking the words in the lexicon, if the words appear in the lexicon, then take the values of the words and add them to the score.

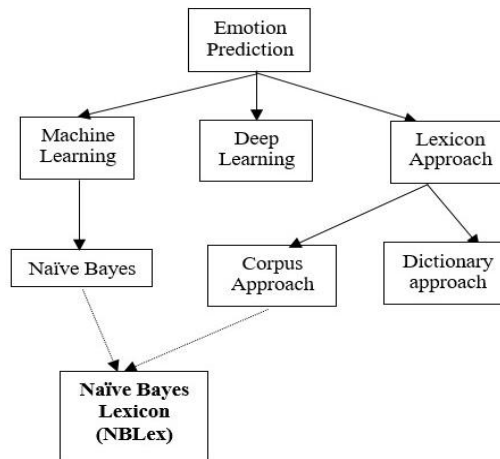


Figure 1. Different approaches for emotion prediction

Predicting the emotions expressed in code-switch text is a challenging task in comparison with predicting the emotions in the monolingual text since both source and scripting languages are not the same. Also, the out-of-vocabulary (OOV) words are more common on social networking websites [9]. Some of the researchers tried to predict the emotions in code-switch text using ML and DL techniques [10]–[13]. However, there are some limitations with ML and DL techniques as they need training data, they are computationally expensive, and also incapable to handle OOV words. One of the alternative approaches is machine translation, which translates the code-switch text into monolingual text using machine translation techniques. To translate this type of text requires more additional efforts to improve the performance of the translation process and sometimes the translation is incorrect. For example, considering Ex3, Google translator has given an incorrect translation.

- Ex3: “*avlee kannada baralla avl enu kannada kalstale*”
- Translation: “She doesn’t know Kannada how she can teach Kannada”
- Google Translation: “What is Kannada?”

To address the above limitations, few researchers are trying to use the lexicon-based approach for emotion prediction [14]. Here, the lexicon-based approach does not require any training data and these are computationally not expensive in comparison with ML and DL. The lexicon-based approach uses a list of pre-labeled words that are classified into various emotions [15]. The lexicon-based approach classified into two types, the first one is corpus-based approach and the second one is dictionary-based approach. In the corpus-based approach, lexicon will be built based on the co-occurrence of a word (statistical) or semantic of a word. In dictionary-based approach, first will collect a few words (seed words) with emotional labels. Next

step, we will use the Bootstrap method to find all synonyms of seed words from the dictionary and add these synonyms to the seed list. Some of the well-known emotion lexicons are affective norms for English words (ANEW) [16], linguistic norms and word count (LIWC) [17], WordNet-affect dictionary emotion lexicon [18], National Research Council Canada (NRC) word-emotion association lexicon [19], and DepecheMood emotional lexicon [20]. The dictionary-based approach is not a better approach for code-switch text since it has limited vocabulary at the same time social media text is OOV. In the recent past, the relevant literature on this topic has been carried out by many researchers who worked on monolingual text to predict the emotions due to the availability of large-scale monolingual data in social media networks and also it is easy to implement in comparison with code-switch text.

Wang *et al.* [10] proposed a method called BLSTM-multiple classifiers (BLSTM-MC) to study the association between various emotions in each code-switch post and predicts all the emotions conveyed by each post. Tang *et al.* [11] examined how to fine-tune the bidirectional encoder representations from transformers (BERT) model for multi-label sentiment (emotions) analysis in code-switch text. This approach contains a collection of pre-trained models and the fine-tuning techniques of BERT. Data augmentation, under sampling, and ensemble learning are used to get a balanced sample from an unbalanced sample. Mohammad and Turney [19] discussed emotion prediction in the paper titled “Crowdsourcing a word-emotion association lexicon”. In this, they created a huge and high-quality word-emotion, and word-polarity association lexicon quickly and inexpensively. They also used 10,000 Word-Sense pairs, and each pair of word-sense is associated with 8 basic emotions. However, this work can be enhanced to create lexicons in other languages.

Staiano and Guerini [20] build an emotional lexicon called DepecheMood with approximately 37,000 words marked with emotion values from a social news network. This lexicon shows substantial progress over the state-of-art unsupervised approach with high coverage and high precision. In this work, there is a scope to enhance for testing the unusual value decomposition on the word-by-document matrices. Purver and Battersby [21] proposed the cross-convention method for multi-class emotion detection in Twitter messages with no manual interference using automatically labeled data. This approach has given a reasonable performance at individual emotion prediction. However, this work is better for few of the emotions such as happy, sad, and anger in comparison with other emotions such as fear, surprise, and disgust. Wen and Wan [22] proposed an approach called class sequential rules to predict the emotions on a Chinese benchmark dataset. This approach has shown greater performance in comparison with other approaches. The performance of this approach can be improved by using additional social network information and discourse information.

Hasan *et al.* [23] used the Circumplex model to classify the individual emotional positions. This model has outperformed with 90% Accuracy in comparison with k-nearest neighbors (KNN), decision tree (DT), SVM, and NB. Further, this approach can be extended to examine the sequential nature of the emotions and how they change over the period. Mohammad and Kiritchenko [24] proposed that emotion-word Hashtags are good manual labels of emotions in tweets and also proposed a method to generate the large lexicon of word emotion. This is the first lexicon with real-valued word emotion scores. This work can be extend to analyzing the difference in emotions associated with different morphological forms of the emotion words. Al-Aziz *et al.* [25] proposed the model that combines the lexicon approach and multi-criteria decision-making approach (MCDM). This model can be used to classify the text into different emotions and also classify Arabic text with mixed emotions. The dataset which is used in this work is very small (200 tweets) and which needs to be tested on the large dataset to know the performance.

Gupta *et al.* [26] proposed a method to identify the emotions such as happy, sad, and angry using the deep learning approach LSTM. In this method, they combined the advantages of semantic and sentiment-based embedding. This method can be extend to train the context-aware models. Schuff *et al.* [27] explores the relationship between emotion annotation and the other annotation such as sentiment as well as stance layers. Here, BiLSTM method performed better than all other approaches. Mohammad and Bravo-Marquez [28] used a method called best-worst scaling (BWS) to improve annotation consistency and find the related fine-grained emotion scores. They identified that Emotion-Word Hashtags will carry more intense emotions.

Rabeya *et al.* [29] proposed an emotion prediction model to find the emotions from Bengali text. A lexicon-based Backtracking approach is used to find the sentiments of sentences to prove how often users express emotion at the end of the sentence. The corpus used in this work is very small (351 lexicons only) and which is not enough to get decent accuracy. Batbaatar *et al.* [30] proposed a method called semantic-emotion neural network (SENN) by adding both emotional information and semantic/syntactic information. This model is mainly divided into two sub-networks. First, BiLSTM which finds the contextual information and focuses on a semantic relationship. Secondly, convolutional neural networks (CNN) are to extract emotional information and focuses on emotional relationships. Further, SENN performance can be improved by using the higher emotion word embeddings.

Currently, most of the emotion prediction research work has been done on monolingual text but very few researchers have performed emotion prediction on code-switch text due to the complexity of the text. Lee and Wang [31] proposed a multiple-classifier based automatic approach to predict the emotions in the Chinese-English code-switch corpus. This approach shows that both English text and Chinese text are active to predict the emotions. Wang *et al.* [32] used a term-document Bipartite Graph to integrate both bilingual and sentimental information and then used a propagation-based method to acquire and predict the emotions. This work further can be extended to discover the association between emotions and caused languages for identifying the emotions. Lee and Wang [33] proposed a multi-view learning structure to study and predict the emotions through both monolingual and bilingual views. The investigational studies validate that the proposed approach significantly outperforms various strong baselines. Wang *et al.* [34] proposed a joint factor graph model to predict the emotions in code-switching text. Here, the Factor function is used to discover the association between different emotions, and a belief propagation algorithm is used to study and predict the model.

When it comes to emotion prediction in code-switching text of Indian languages, very few researchers are tried to predict the emotions. Vijay *et al.* [12] proposed a supervised classification approach that uses different machine learning methods SVM to find the emotion associated with Hindi-English code-mixed data using character level, word level, and lexicon based features. To achieve better results, the corpus can be annotated with part-of-speech (POS) tags at the word level. Appidi *et al.* [13] performed an experiment on Kannada-English code-mixed data to predict the emotions. They used SVM and deep learning method LSTM. The accuracy of SVM is 30% and the accuracy of LSTM is 32%. However, the corpus needs to be annotated with POS tags at the word level.

The corpus-based approach will handle the above limitation since there is no limit to vocabulary. In this paper, we proposed a corpus-based lexicon approach called NBLex to predict the emotions in Kannada-English code-switch text. We used NB classifier to build the proposed method.

The structure of the remaining paper is as follows. Section 2 we proposed the lexicon generation and emotion prediction. Comparative result analysis is discussed in section 3. Finally, the conclusion and future works are discussed in section 4.

## 2. METHOD

### 2.1. Lexicon generation

In this section, we discuss Kannada-English code-switch (KECS) corpus creation and annotation. Further, we will discuss the process for creating the NBLex lexicon. In order to develop KECS corpus, we assume only those comments which are having primarily code-switch nature. We consider the comments as code-switch, even if one word differs from the monolingual condition (i.e., source and scripting languages must be the same). Here, all the words in our corpus are English script only.

To create the corpus KECS, initially, we gathered approximately 7526 Kannada-English code-switch comments from YouTube.com based on various topics such as politics, movies, celebrities, and social events. We performed the pre-processing task to eliminate noisy data and the pre-processing task includes the removal of special characters, symbols, and digits as they will not have any importance in the process of lexicon creation since these words will not carry any emotional meaning. The next step is to carry out the process of annotation. In this process manual labeling of emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, and trust) is made, since we do not have any automatic labeling methods for Kannada-English code-switch text.

In lexicon creation, NB classifier is used to generate scores for each word along with the one-vs-rest approach. In this approach, every time we consider one category of emotion as positive (true or 1) and all other categories of emotions as *Not\_Positive* (False or 0) e.g., to generate the score for the anger emotion category, we consider all emotion under anger emotion category as 1 and rest of the emotions as 0. lexicon creation starts with building the frequency dictionary where, Keys tuples (word, label) and values are frequencies. The label value is either 1 or 0 and frequency is an integer value (the number of times that word and label tuple present in the data).

The next step in lexicon creation is to find the positive and *Not\_Positive* frequency of the word from the frequency dictionary. Once frequencies are computed, then we calculate the total number of positive words, the total number of *Not\_Positive* words and also, we calculate the total number of words from the data. After doing all these, we need to find positive and *Not\_Positive* probability for each of the word. We used 10-fold cross-validation while generating the scores for each word in the dictionary to make stronger values. To compute the positive probability of the word we are using (1).

$$P(pw) = \frac{pf + 1}{Np + T} \quad (1)$$

where,  $P(pw)$  is the positive probability of the word,  $pf$  is a positive frequency of the word,  $Np$  is the total number of positive words in corpus, and  $T$  is the total number of words in the corpus. To compute the *Not\_Positive* probability of the word we are using (2).

$$P(nw) = \frac{nf + 1}{Nn + T} \quad (2)$$

where,  $P(nw)$  is the *Not\_Positive* probability of the word,  $nf$  is the *Not\_Positive* frequency of the word,  $Nn$  is the total number of *Not\_Positive* words in the corpus. Finally, we compute the score for every word in the dictionary by using (3).

$$\text{Score} = \log \frac{P(pw)}{P(nw)} \quad (3)$$

where score is the total value of the sentence.

After generating the scores for each of the words in the dictionary, manually we removed one-letter words and English stop words since they will not have any emotional meaning. Also, we removed nouns (names) and other non-emotional words from the lexicon since they will not have emotional significance in the process of emotion prediction. This entire process is used to generate the scores for one emotion. The same process is followed for the rest of the emotions to generate the scores.

Once the final lexicon is ready, to find the dependency among the emotions, we need to check the relationship between different emotions. To perform this, we applied the correlation technique on lexicon and found that fear emotion has a positive correlation with surprise and anticipation emotions. One of the reasons is that the number of comments under these categories (anticipation, fear, and surprise) is very less in comparison with other categories of emotions.

Table 1 shows the correlation between the various emotions. It is observed that the positive correlation between fear and surprise as well as fear and anticipation both have the 0.7. And also, it is observed that the rest of the emotions are either independent or negatively correlated with other emotions.

Table 1. Correlation between emotions

	Anger	Anticipation	Disgust	Fear	Joy	Sad	Surprise	Trust
Anger	1	0.1	0.1	0.2	-0.3	0.0	0.1	-0.2
Anticipation	0.1	1	0.2	0.7	0.1	0.3	0.5	0.2
Disgust	0.1	0.2	1	0.3	-0.2	0.0	0.2	-0.1
Fear	0.2	0.7	0.3	1	0.2	0.4	0.7	0.3
Joy	-0.3	0.1	-0.2	0.2	1	-0.1	0.2	0.3
Sad	0.0	0.3	0.0	0.4	-0.1	1	0.3	0.0
Surprise	0.1	0.5	0.2	0.7	0.2	0.3	1	0.2
Trust	-0.2	0.2	-0.1	0.3	0.3	0.0	0.2	1

## 2.2. Emotion prediction

Figure 2 shows the complete steps for emotion prediction process. To perform emotion prediction using a lexicon-based approach, we gathered total of 1882 Kannada-English code-switch text comments from YouTube.com based on different topics. The pre-processing task is carried out to improve the accuracy of the emotion prediction task by eliminating noisy data such as symbols, digits, and other special characters. As such there is no automated labeling system for Kannada-English code-switch text, the annotation has been done for emotions by a linguistic expert manually. Further, the tokenization task is carried out to split the entire text into tokens to calculate the sentence score. To predict the emotions, first we need to calculate the score for each comment using (4).

$$SS = \sum_{i=0}^n \text{Score}(w) \quad (4)$$

where, SS refers sentence score.

The detailed steps of emotion prediction are shown in algorithm 1. Here, prediction is performed based upon one emotion at a time. e.g., first, predict the Anger emotion by assuming all other emotions are non-anger. Next, the anticipation emotion is predicted by assuming all other emotions are non-anticipation. Like this, the same approach is followed for the rest of the 6 emotions. Once score is calculated, now we need to perform emotion prediction using (5).

$$\text{Predict} = \begin{cases} 1 \text{ (Positive), if } SS > 0 \\ 0 \text{ (Not\_Positive), otherwise} \end{cases} \quad (5)$$

#### Algorithm 1. Emotion prediction

```

Input: t, Score /* t - Input text
           Score - Set of words (Lexicon) with values */
Output: Emotion /* Positive or Not_Positive */
           /*Calculating score of t (Text comment) */
           Cal_Score (t, Score)
    {
1.   SS = 0 /* SS is the temp variable to store the score of t */
2.   for each w in t /* For each word w from t compute the score */
3.     if w in Score then
4.       SS += Score.get(w)
5.     end if
6.   end for
7.   return SS
8. }

/* Predicting the Emotion of each text comment in test_data */
Emotion_Prediction (test_data)
{
9.   Predict = [] /* is a empty list */
10.  for each t in test_data /* for each comment (t) in the test_data compute the score */
11.    if Cal_Score(t, Score) > 0 then
12.      Predict_i = 1 (Positive)
13.    else
14.      Predict_i = 0 (Not-Positive)
15.    end if
16.  end for
17.  Predict.append(Predict_i)
18. }

```

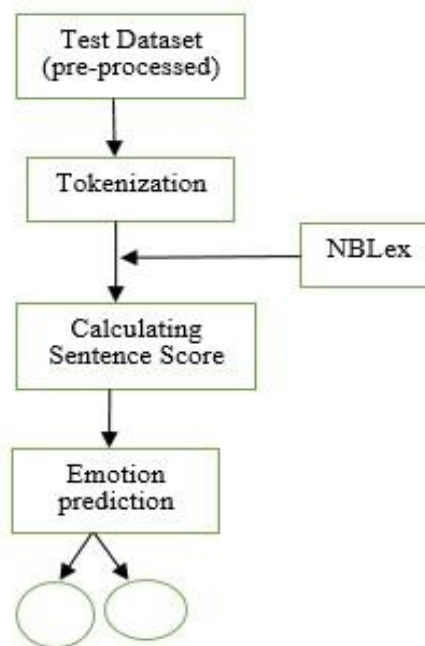


Figure 2. Emotion prediction process

### 3. RESULTS AND DISCUSSION

The results of the proposed NBLex lexicon method is compared with the ML (NB) and DLg (BiLSTM) methods. There are 7526 Kannada-English code-switch text comments used to train both the models separately and applied 10-fold cross-validation while training both the models. All these code-switch comments are manually labeled with eight basic emotions (anger, anticipation, disgust, fear, joy, sad, surprise, and trust). Totally, 1882 code-switch comments are used as test dataset.

### 3.1. Naïve Bayes

Here, two types of vectorizations such as bag-of-words (BOW) and term frequency inverse document frequency (TF-IDF) are performed and then predict the emotions on the test dataset. Considering only accuracy, the BOW method is showing the better results for some of the emotions (anger and joy) and TF-IDF showing the good results for rest of the emotions (anticipation, disgust, fear, sad, surprise and trust). Considering both accuracy and true positives, the BOW method performance is much better than TF-IDF as observed in the Table 2.

The BOW method produces better results for anger emotion (78.4%). The BOW and TF-IDF are almost the same for joy (84.7% and 84.6%) and for trust (78.9% and 79%) emotions. For the rest of the emotions (anticipation, disgust, fear, sad, and surprise) TF-IDF is much better. However, when considering the true positives, BOW provides very good results in comparison with TF-IDF. Surprising that the true positives are 0 for emotions like anticipation, fear, and sad in TF-IDF, it is due to sample bias and Distribution bias which are the main reasons for getting 0 true positives in TF-IDF. Also, the score of TF-IDF is always between 0 and 1 whereas, the score of BOW is a natural number. The BOW approach is compared with BiLSTM and lexicon. TF-IDF is not a suitable method since the performance of any method will not depend only on accuracy but also it depends upon both the accuracy and true positives.

Table 2. Accuracy and true positives of BOW and TF-IDF

Emotions	Accuracy		True Positives	
	BOW	TF-IDF	BOW	TF-IDF
Anger	78.4	74.1	289	68
Anticipation	91.3	95.9	18	0
Disgust	78.9	81.8	155	12
Fear	97.8	99.4	1	0
Joy	84.7	84.6	489	321
Sad	83.9	87.2	70	0
Surprise	92.5	97	4	1
Trust	78.9	79	231	69

### 3.2. Bidirectional long-short term memory

To predict the emotions, the BiLSTM approach is applied to the test dataset. The BiLSTM approach outperforms for fear (98.8%) emotion, but the total number of comments is very less (12). Generally, the DL models work well for large datasets. However, in this scenario, the BiLSTM model performs well on a small dataset. This can be observed in Table 3 which shows the accuracy and true positives.

### 3.3. NBLex lexicon

NBLex lexicon is applied on test dataset to predict emotions. Table 4 shows the accuracy and true positives results obtained for various emotions. The fear emotion achieved better accuracy (94%) in comparison with all other emotions for total of 12 comments. The reason for this condition is based on the emotion dependency.

Table 3. Accuracy and true positives of BiLSTM

Emotions	Accuracy	True Positives
Anger	75.6	294
Anticipation	88.2	28
Disgust	81.1	145
Fear	98.8	2
Joy	83.3	440
Sad	81.9	85
Surprise	91.7	14
Trust	77.2	261

Table 4. Accuracy and true positives of NBLex lexicon

Emotions	Accuracy	True Positives
Anger	83.2	332
Anticipation	92.3	32
Disgust	85.8	207
Fear	94	3
Joy	87.5	510
Sad	84.5	102
Surprise	92.4	17
Trust	83.8	284

From Table 5, it is observed that the NBLex is outperformed in terms of accuracy for most of the emotions (anger, anticipation, disgust, joy, sad, and trust) in comparison with other approaches like NB and BiLSTM. For fear emotion, BiLSTM produces remarkable results in comparison with other approaches like NB and NBLex. For the surprise emotion, both NB and NBLex are almost showing the same results (92.5% and 92.4%). From Table 6, it is observed that the results obtained from proposed model NBLex is better than NB and BiLSTM for all emotions in terms of true positives why because it is due to there is no limit for vocabulary and it can handle OOV words.

Figure 3 shows the accuracy of the various emotions for proposed model NBLex and other models NB, BiLSTM. From Figure 3, it is observed that the NBLex performs well for most of the emotions (6 emotions out of 8 emotions) in terms of accuracy in comparison with the NB and BiLSTM. Figure 4 shows the true positive rate (TPR) of the various emotions for proposed model NBLex and other models NB, BiLSTM. From Figure 4, it is observed that the NBLex performs well for all the emotions in terms of TPR in comparison with the NB and BiLSTM approaches.

Table 5. Accuracy comparison of NB, BiLSTM, and NBLex

Emotions	NB	BiLSTM	NBLex
Anger	78.4	75.6	83.2
Anticipation	91.3	88.2	92.3
Disgust	78.9	81.1	85.8
Fear	97.8	98.8	94
Joy	84.7	83.3	87.5
Sad	83.9	81.9	84.5
Surprise	92.5	91.7	92.4
Trust	78.9	77.2	83.8
Average Accuracy	85.8	84.7	87.9

Table 6. True positives rate of NB, BiLSTM, and NBLex

Emotions	NB	BiLSTM	NBLex
Anger	52.2	53.1	60.0
Anticipation	23.0	35.8	41.0
Disgust	43.7	40.9	58.4
Fear	8.3	16.6	25.0
Joy	82.8	74.5	86.4
Sad	29.0	35.2	42.3
Surprise	7.0	24.5	29.8
Trust	50.3	56.8	61.8
Average TPR	37.0	42.2	50.6

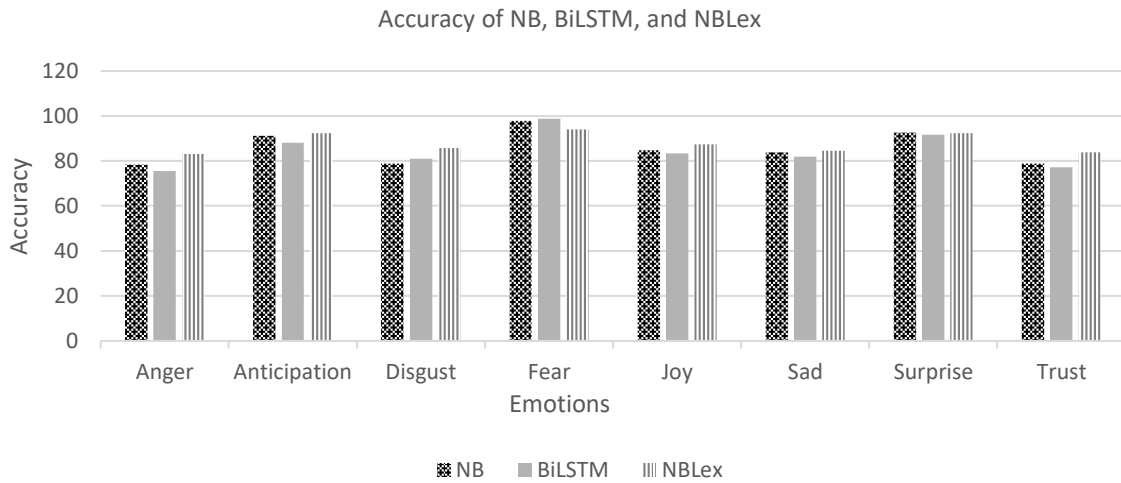


Figure 3. Accuracy of various emotions

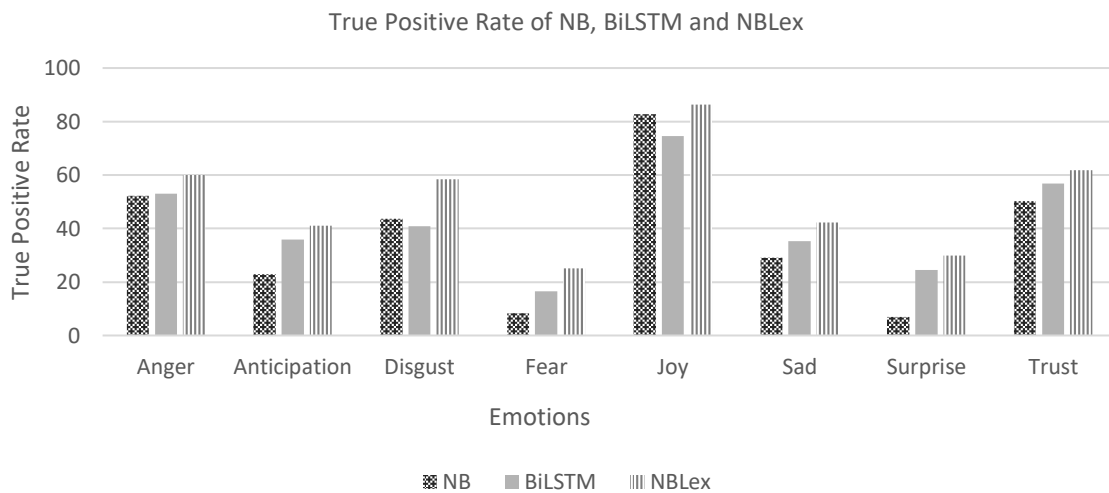


Figure 4. True positive rate of various emoticons



#### 4. CONCLUSION

In this paper, we contributed NBLex for emotion prediction in Kannada-English code-switch text. The proposed approach NBLex provides optimal results (87.9%) in terms of accuracy in comparison with other traditional NB (85.8%) and BiLSTM (84.7%) techniques. Also, in terms of TPR, the NBLex produces 50.6% which is better than NB (37.0%) and BiLSTM (42.2%). From our contribution, it shows that a simple additive model (i.e., lexicon) is also be an alternative approach to predict the emotions in Kannada-English code-switch text. This is the first work that aims to predict the emotions in Kannada-English code-switch text by using lexicon approach.

Further, in future we want to enhance our work by addressing following three points: i) adding additional emotional data and increasing the accuracy as well as TPR, ii) contextual detection of emotions from the corpus, and iii) AB testing for engagement activities like pacification of anger customers or increasing net profit score (NPR) of happy customers, as the proposed model detects these two prominent emotions effectively.

#### ACKNOWLEDGEMENTS

The authors want to thank the REVA University management for their support of this research activity.




#### REFERENCES

- [1] R. Plutchik, "A psychoevolutionary theory of emotions," *Social Science Information*, vol. 21, no. 4–5, pp. 529–553, Jul. 1982, doi: 10.1177/053901882021004003.
- [2] C. O. Alm, D. Roth, and R. Sproat, "Emotions from text: machine learning for text-based emotion prediction," in *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, 2005, pp. 579–586.
- [3] Y. Chen, S. Y. M. Lee, S. Li, and C.-R. Huang, "Emotion cause detection with linguistic constructions," in *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, 2010, pp. 179–187.
- [4] P. R. Nagarajan, "52% of India's urban youth are now bilingual, 18% speak three languages'," *The Times of India*, 2018. <https://timesofindia.indiatimes.com/india/52-of-indias-urban-youth-are-now-bilingual-18-speak-three-languages/articleshow/66530958.cms> (accessed Jun 16, 2022).
- [5] J. Lipski, "Code-switching and the problem of bilingual competence," *Aspects of bilingualism*. Columbia: Hornbeam Press Inc, pp. 250–264, 1978.
- [6] M. Gysels, "French in urban Lubumbashi Swahili: Codeswitching, borrowing, or both?," *Journal of Multilingual and Multicultural Development*, vol. 13, no. 1–2, pp. 41–55, Jan. 1992, doi: 10.1080/01434632.1992.9994482.
- [7] C. Myers-Scotton, *Duelling languages: Grammatical structure in codeswitching*. Oxford University Press, 1997.
- [8] P. Muysken and Others, *Bilingual speech: A typology of code-mixing*. Cambridge University Press, 2000.
- [9] T. Baldwin, P. Cook, M. Lui, A. MacKinlay, and L. Wang, "How noisy social media text, How diffrent social media sources?," in *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, Oct. 2013, pp. 356–364.
- [10] T. Wang, X. Yang, C. Ouyang, A. Guo, Y. Liu, and Z. Li, "A multi-emotion classification method based on BLSTM-MC in code-switching Text," in *CCF International Conference on Natural Language Processing and Chinese Computing*, 2018, pp. 190–199, doi: 10.1007/978-3-319-99501-4\_16.
- [11] T. Tang, X. Tang, and T. Yuan, "Fine-tuning BERT for multi-label sentiment analysis in unbalanced code-switching text," *IEEE Access*, vol. 8, pp. 193248–193256, 2020, doi: 10.1109/ACCESS.2020.3030468.
- [12] D. Vijay, A. Bohra, V. Singh, S. S. Akhtar, and M. Shrivastava, "Corpus creation and emotion prediction for Hindi-English code-mixed social media text," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, 2018, pp. 128–135, doi: 10.18653/v1/N18-4018.
- [13] A. R. Appidi, V. K. Srirangam, D. Suhas, and M. Shrivastava, "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*, 2020, pp. 6703–6709, doi: 10.18653/v1/2020.coling-main.587.
- [14] R. Kamal, M. A. Shah, C. Maple, M. Masood, A. Wahid, and A. Mehmood, "Emotion classification and crowd source sensing; A lexicon based approach," *IEEE Access*, vol. 7, pp. 27124–27134, 2019, doi: 10.1109/ACCESS.2019.2892624.
- [15] S. Mohammad, "From once upon a time to happily ever after: Tracking emotions in mail and books," *Decision Support Systems*, vol. 53, no. 4, pp. 730–741, 2012.
- [16] M. M. Bradley and P. J. Lang, "Affective norms for English words (ANEW): Instruction manual and affective ratings," Technical Report C-1, The Center for Research in Psychophysiology, University of Florida, 1999.
- [17] J. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic Inquiry and Word Count (LIWC): LIWC2001," *Mahway: Lawrence Erlbaum Associates*, 2001.
- [18] C. Strapparava and A. Valitutti, "WordNet-affect: an affective extension of WordNet," *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, May 2004.
- [19] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word-emotion association lexicon," *Computational Intelligence*, vol. 29, no. 3, pp. 436–465, Aug. 2013, doi: 10.1111/j.1467-8640.2012.00460.x.
- [20] J. Staiano and M. Guerini, "Depeche mood: a Lexicon for emotion analysis from crowd annotated news," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Jun. 2014, pp. 427–433, doi: 10.3115/v1/P14-2070.
- [21] M. Purver and S. Battersby, "Experimenting with distant supervision for emotion classification," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Apr. 2012, pp. 482–491.
- [22] S. Wen and X. Wan, "Emotion classification in microblog texts using class sequential rules," in *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014, pp. 187–193.
- [23] M. Hasan, E. A. Rundensteiner, and E. O. Agu, "EMOTEX: Detecting emotions in twitter messages," in *2014 ASE*




- Bigdata/Socialcom/Cybersecurity Conference*, 2014, pp. 1–10.
- [24] S. M. Mohammad and S. Kiritchenko, "Using hashtags to capture fine emotion categories from tweets," *Computational Intelligence*, vol. 31, no. 2, pp. 301–326, 2015.
- [25] A. M. Abd Al-Aziz, M. Gheith, and A. S. Eldin, "Lexicon based and multi-criteria decision making (MCDM) approach for detecting emotions from Arabic microblog text," in *2015 First International Conference on Arabic Computational Linguistics (ACLing)*, Apr. 2015, pp. 100–105, doi: 10.1109/ACLing.2015.21.
- [26] U. Gupta, A. Chatterjee, R. Srikanth, and P. Agrawal, "A sentiment-and-semantics-based approach for emotion detection in textual conversations," *arXiv:1707.06996*. arXiv, Jul. 2017.
- [27] H. Schuff, J. Barnes, J. Mohme, S. Padó, and R. Klinger, "Annotation, modelling and analysis of fine-grained emotions on a stance and sentiment detection corpus," in *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2017, pp. 13–23, doi: 10.18653/v1/W17-5203.
- [28] S. Mohammad and F. Bravo-Marquez, "Emotion intensities in tweets," in *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (SEM 2017)*, Aug. 2017, pp. 65–77, doi: 10.18653/v1/S17-1007.
- [29] T. Rabeya, S. Ferdous, H. S. Ali, and N. R. Chakraborty, "A survey on emotion detection: A lexicon based backtracking approach for detecting emotion from Bengali text," *2017 20th International Conference of Computer and Information Technology (ICCIT)*, pp. 1–7, Dec. 2017, doi: 10.1109/ICCITECHN.2017.8281855.
- [30] E. Batbaatar, M. Li, and K. H. Ryu, "Semantic-emotion neural network for emotion recognition from text," *IEEE Access*, vol. 7, pp. 111866–111878, 2019, doi: 10.1109/ACCESS.2019.2934529.
- [31] S. Lee and Z. Wang, "Emotion in code-switching texts: Corpus construction and analysis," in *Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing*, Jul. 2015, pp. 91–99, doi: 10.18653/v1/W15-3116.
- [32] Z. Wang, S. Lee, S. Li, and G. Zhou, "Emotion detection in code-switching texts via bilingual and sentimental information," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Jul. 2015, pp. 763–768, doi: 10.3115/v1/P15-2125.
- [33] S. Y. M. Lee and Z. Wang, "Multi-view learning for emotion detection in code-switching texts," in *2015 International Conference on Asian Language Processing (IALP)*, Oct. 2015, pp. 90–93, doi: 10.1109/IALP.2015.7451539.
- [34] Z. Wang, S. Y. M. Lee, S. Li, and G. Zhou, "Emotion analysis in code-switching text with joint factor graph model," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 469–480, Mar. 2017, doi: 10.1109/TASLP.2016.2637280.

## BIOGRAPHIES OF AUTHORS






**Ramesh Chundi**    received the B.Sc degree in computer science and MCA degree from Sri Venkateswara University, India, in 2004 and 2007, respectively. Currently pursuing Ph.D degree in Computer Science and Applications from REVA University, India. His research interests include natural language processing (NLP), artificial intelligence, ML, DL, data analytics, and data mining. He can be contacted at email: chundiramesh@gmail.com.



**Vishwanath R. Hulipalled**    is a Professor in the School of Computing and IT, REVA University, Bangalore, Karnataka, India. He completed BE, ME and Ph.D. in Computer Science and Engineering. His area of Interests includes Machine Learning, Natural Language Processing, Data Analytics and Time Series Mining. He has more than 24 years of academic experience and research. He authored more than 50 research articles in reputed journals and conference proceedings. He can be contacted at email: vishwanth.rh@reva.edu.in.



**Jay Bharthish Simha**    is the CTO of ABIBA Systems and Chief Mentor at RACE Labs, REVA University. He completed his BE (Mech), M.Tech (Mech), M.Phil.(CS) and Ph.D. (AI). His area of interest includes fuzzy logic, soft computing, machine learning, deep learning, and applications. He has more than 20 years of industrial experience and 4 years of academic experience. He has authored/co-authored more than 50 journal and conference publications. He can be contacted at email: jay.b.simha@reva.edu.in and jay.b.simha@abibasystems.com.