International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-4, November 2019

Prediction of Emoji from News Headlines using Machine Learning Techniques

Soumya S., K. V. Pramod

Abstract— Emojis are generated from the news headlines of Malayalam language using Naive Bayes (NB) and Support Vector Machine(SVM) classifiers. The human brain processes visual data faster than text data. Assigning emoji to the text helps the people easily categorize the news based on the emotion without reading the entire sentence. Six different emojis are assigned to the news headlines based on the emotional contents of the text. These emojis are used for representing emotions like sad, angry, fear, happy, love, and neutral. The dataset contains 3111 sentences which are retrieved from the tweets of Manorama Online. Both Bag-of-Words (BOW) and Term Frequency versus Inverse Document Frequency (TFIDF) features are used for feature vector formation of the dataset. The SVM shows better accuracy compared with the NB classifier.

Index Terms—emotional analysis, Malayalam tweets, natural language processing, SVM

I. INTRODUCTION

Sentiment Analysis (SA) is the computational study that analyses people's opinions from the text and classified as positive, negative, and neutral [1]. Emotional Analysis extracts people's emotions expressed in the text. The main emotions are sad, angry, fear, happy, love, and neutral. The negative sentiment oriented sentences are classified as sad, angry, and fear, whereas positive sentiment oriented sentences are categorized into happy and love based on the emotional words present in the text. The news that does not have any emotional words are assigned with neutral emoji. Emotion analysis is the deeper level of Analysis of text beyond sentiment analysis.

Human brain processes visual data 60,000 times faster than text data. Emojis are actual pictures created inthelate1990sby NTT DoCoMo, which are widely used insocialmediacontentssuchasfacebook,twitter,messages,etc. From2010onwards emojis are encoded into Unicode standard. People can easily identify the emotion behind the text from the emojis without reading the content of text. They can easily access the data based on their interest without reading the entire sentence.

Manuscript published on November 30, 2019.
* Correspondence Author
Soumya S*.,Department of Computer Applications, Cochin University of Science and Technology, Kochi-22, India. Email: ps.soumya02@gmail.com
K. V. Pramod, Department of Computer Applications, Cochin University of Science and Technology, Kochi-22, India. Email: pramodkv4@gmail.com

Prediction of emojis from the text is achallengingtaskinNaturalLanguageProcessing(NLP).Thenewsheadlinesof Malayalam are retrieved from the tweets ofManoramaOnline.Sixdifferentemojisareassignedtothedata setbasedontheemotion of each sentence. The positive sentiment oriented news are classified as happy and love and assigning two different emojis like thumbs up (👍) and smiling face with heart-eyes (😍).

The negative news are classified as sad, fear and anger and the corresponding emojis are disappointed face(😞), fearful face(😨), and angry face(😡) respectively. The news which belongs to neither of these categories is classified as a neutral face and represented it using the emoji with a neutral face(😐) [11].

In this paper, the NB classifier and SVM are used for predicting the Malayalam news headlines as six different emojis based on emotional words in the sentence. The news headlines are retrieved from Twitter. The Manorama online tweeted the news updates at regular intervals. These news headlines are retrieved using twitter API. The major challenge in emoji generation is the unavailability of the labeled dataset. Due to the unavailability of a labeled corpus, we have created a corpus contains 3111 Malayalam news headlines tweets that are manually tagged. After the preprocessing steps, the retrieved tweets are manually verified and assigned with six different emojis based on emotional content.

The remainder of this paper is organized as follows: Section II explains the major works done in this area. Whereas Section III describes the proposed method for emotion analysis. Section IV briefing machine learning models. Section V illustrates the experimental setup we have done. Section VI discusses the results of different machine learning models. Finally, Section VII concludes the paper.

II. RELATEDWORKS

The works done in emotion analysis and emoji prediction of English languages, as well as in Indian languages, are mentioned here.

Purver et al. (2012) classified the tweets as happy, sad, and anger using SVM on a manually annotated dataset, which contained 1000 sentences [3].

Hasan et al. (2014) proposed emotion detection in twitter messages using different classifiers like NB, SVM, decision tree, and KNN [2].

Emoji prediction of English and Spanish language was made by coltekin et al. (2018). They used both SVM and RNN for Emoji prediction and showed that SVM performs better than RNN[4].

Emoji usages in the tweet content, tweet structure and user demographics proposed by Peijun Zhao et al. (2018).

They proposed a multitask multimodality gated recurrent unit (mmGRU) model to predict the categories and positions of emojis[5].

Pre-trained Deep Moji model was proposed by Felbo et al. (2017) [6].

Multilingual Emoji Prediction of Hindi, Bengali, and Telungu was made by Choudhary et al. (2018). They have identified the top 20 emojis used in tweets of each language and predicted the percentage of occurrence of emojis in the languages [7].

Tissa Tony proposed Faith and Emotional Intelligence in 2019. They analyzed various emotions and studied how to transform the stressful lives into peaceful life [10].

III. PROPOSED METHOD

Jack Dorsey created Twitter in 2006, and tweets were restrictedto140charactersuntil2017.Now it is 280 characters long. The proposed work is the first attempt towards the emoji prediction of Malayalam text. Malayalam, the mother tongue of Keralites, is the most commonly used language to express their opinion through twitter. Fig. 1 shows the proposed architecture.

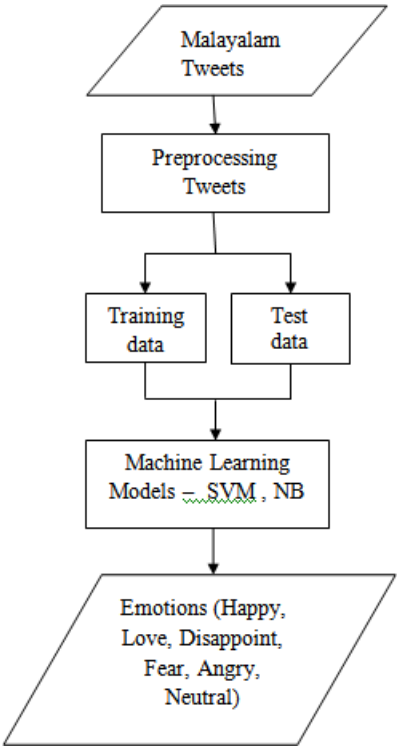

Fig. 1. Proposed architecture for emoji prediction ofMalayalam tweets

A. Dataset

The data set of news headlines is collected from the tweets of Manorama Online. The retrieved news headlines are classified as six emotions and manually assigned six emojis based on the emotional words present in the sentence. The dataset contains 3111 sentences. Table I and Fig. 2 show the number of news headlines under each category.

B. Preprocessing

The retrieved tweets contain 3111 tweets. As the first step of preprocessing, punctuations, special characters, hyperlinks, etc. are removed using regular expression in Python language. The retrieved news headlines are manually tagged with six different emojis. The tagged corpus is tokenized, and the word vector is created using BOW and TF-IDF features. Then the datasets are split in the ratio 70: 30 where 70% of the entire dataset is used for training purposes and 30% for testing. 10- fold cross validation is applied in this work.

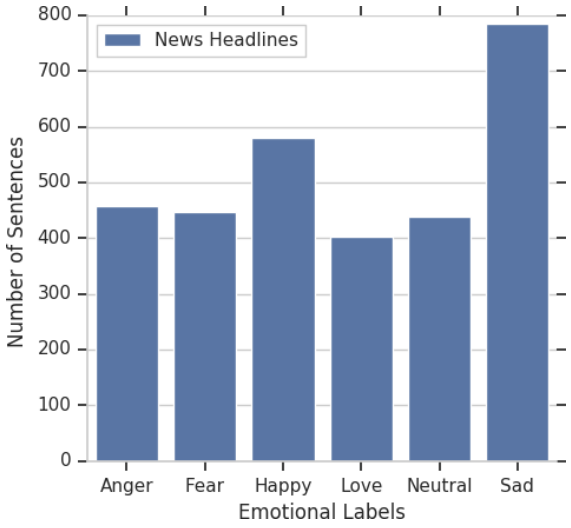

Fig.2. Number of Sentences in each Category

IV. MODELSDESCRIPTION

Emoji prediction of news headlines is done using NB and SVM classifiers. Multinomial NB, and linear SVM are used for predicting Emojis.

Features:Bag of Word (BOW) and Term Frequency vs. Inverse Document Frequency (TF-IDF) features are used for feature matrix formation of the dataset.

Bag of Word (BOW): In BOW, the text is transformed into a bag of words where each entry corresponds to the number of occurrences of a particular term in the sentence. The feature matrix is created with m * n dimension where m is the number of sentences and n is the number of unique words in the corpus.

Term Frequency vs. Inverse Document Frequency (TF-IDF): TF-IDF is the statistical measure to evaluate the significance of a particular term in a corpus. It helps to remove stop words from the dataset. Stop words are less information-oriented words, but it frequently appears in sentences. Term frequency(tf_t) is the ratio of the number of occurrences of particular term t to the total number of words in the corpus.

Inverse document frequency (idf_t) = *log* (*total number of sentences in the corpus / number of sentences which contains the term t*)

$tf - idf = tf_t * idf_t$

Naive Bayes Classifier:Naive Bayes classifier predicts the emotions expressed in a text-based on the emotional word present in the dataset. This classification is done based on Bayes theorem [8]. Before applying the Naive Bayes classifier, the datasets are converted to feature vectors (m * n) using BOW and TF-IDF features.

The output label is an m * 1 matrix. In this work, the Multinomial Naive Bayes classifier is used for the learning process. Add one smoothing technique issuedtoavoidzeroprobability.

Table – I: Number of sentences in each emotional category

Emotions	Emoji	Number of Sentences	Sentiment of Sentence
Happy	👍	581	Positive
Love	😍	402	Positive
Disappointed Face	😞	785	Negative
Fearful Face	😨	447	Negative
Angry Face	😠	457	Negative
Neutral	😐	439	Neutral

Support Vector Machine (SVM): Support Vector Machine is a supervised machine learning algorithm proposed by Vapnik in 1992 [9]. The linear model of SVM is created for predicting the emotions of tweets. The input dataset is vectorized using BOW and TF−IDF features. The input vector is an m * n matrix, which is mapped into a high dimensional feature space, and SVM finds the linear separator with maximum marginal distance in the high dimensional space using support vectors.

V. EXPERIMENTAL SETUP

Two different machine learning models such as NB and SVM are applied on manually created datasets of Malayalam news headlines for predicting emojis of various emotions like sad, angry, fear, happy, love, and neutral. SVM with the TFIDF feature got higher accuracy compared with the other models. Here, the analysis is done on 3111 Malayalam tweets containing 581 happy, 402 love, 785 disappoint 447 fear, 457 angry, and 439 neutral emotional tweets. The confusion matrix, precision, recall, and F−measure are noted in each model. Precision, recall, and F−measure of NB and SVM classifiers with BOW and TF-IDF features are shown in Table II and Table III, respectively. Finally, Fig. 3, the bar chart represents the comparative study of the accuracy of NB and SVM modelswith different features.

VI. C ONCLUSION

We have shown how machine learning models like NB and SVM can be used in emoji prediction of Malayalam tweets. As the major challenge in the emotional analysis of Malayalam is the scarcity of annotated corpus, we have created a manually labeled corpus containing 3111 Malayalam tweets. Emotional analysis was done using NB and SVM taking into account BOW and TF-IDF features.Experimental results have shown that using SVM with TF−IDF feature has the better performance when compared with NB model. We have obtained an accuracy of 74.3% in SVM when the feature selected was TF−IDF.

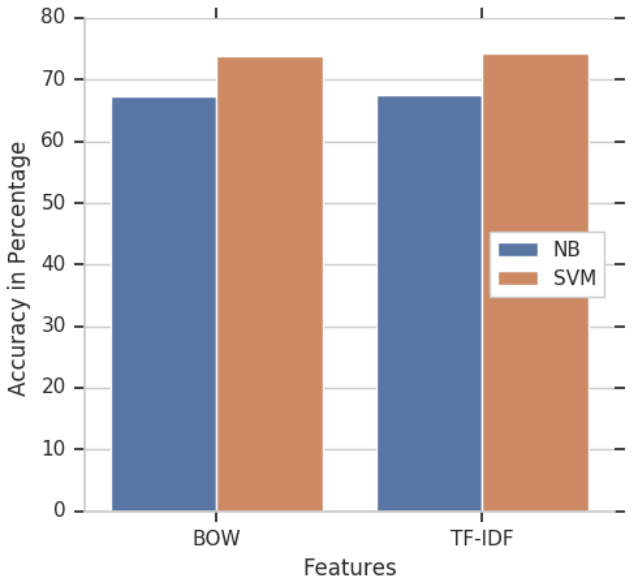

Fig. 3. Comparing Accuracy of Test Dataset with NB and SVM models with BOW and TF-IDF features.

Table- II: Precision, Recall and F-Measure of NB and SVM classifiers with BOW feature

Emotional Labels	NB			SVM		
	Precision	Recall	F-Measure	Precision	Recall	F-Measure
Happy	69.73	60.92	65.03	76.73	70.11	73.27
Love	57.97	66.11	61.78	69.74	68.6	69.17
Sad	63.2	75	68.6	71.74	69.91	70.81
Fear	77.34	73.88	75.57	89.26	80.6	84.71

Anger	76.61	69.34	72.8	72.14	73.72	72.92
Neutral	64.29	54.55	59.02	67.27	84.01	74.75

Table- III: Precision, Recall and F-Measure of NB and SVM classifiers with TF-IDF feature

Emotional Labels	NB			SVM		
	Precision	Recall	F-Measure	Precision	Recall	F-Measure
Happy	68.64	61.29	64.75	72.12	68.39	70.21
Love	57.83	67.42	62.25	72.17	68.6	70.34
Sad	65.21	75	69.76	71.9	73.73	72.8
Fear	76.84	74.73	75.77	90.52	78.36	84
Anger	78.9	71.25	74.88	76.47	75.91	76.19
Neutral	65.28	52.79	58.37	68.12	82.58	74.66

Table- IV: Cross Validation and Test data Accuracy

Classifiers	Validation Accuracy	Test data Accuracy
NB with BOW	66.77	67.34
NB with TF - IDF	67.12	67.63
SVM with BOW	76.84	73.88
SVM with TF - IDF	76.68	74.3

REFERENCES

1. Pang, Bo and Lee: Thumbs up?: sentiment classification using machinelearning techniques, in Proceedings of the ACL-02 conference onEmpirical methods in natural language processing-Volume 10 pp. 79-86 (2002).
2. Hasan, M., Rundensteiner, E., Agu, E. (2014). Emotex: Detectingemotions in twitter messages. ASE BIG DATA/ SOCIALCOM/ CYBER SECURITY Conference, Stanford University, May 27-31, 2014.
3. Purver, M., Battersby, S. (2012, April). Experimenting with distantsupervision for emotion classification. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (pp. 482-491). Association for Computational Linguistics.
4. ltekin, Rama, T. (2018, June). Tbingen-Oslo at SemEval-2018 Task 2: SVMs perform better than RNNs in emoji prediction. In Proceedings of the 12th International Workshop on Semantic Evaluation (pp. 34-38).
5. Zhao, P., Jia, J., An, Y., Liang, J., Xie, L., Luo, J. (2018, April).Analyzing and predicting emoji usages in social media. In CompanionProceedings of the The Web Conference 2018 (pp. 327-334).International World Wide Web Conferences Steering Committee.
6. Felbo, B., Mislove, A., Sgaard, A., Rahwan, I., Lehmann, S. (2017).Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. arXiv preprintarXiv:1708.00524.
7. Choudhary, N., Singh, R., Rao, V. A., Shrivastava, M. (2018, August).Twitter corpus of resource-scarce languages for sentiment analysis and multilingual emoji prediction. In Proceedings of the 27th International Conference on Computational Linguistics (pp. 1570-1577).
8. Jurafsky, Daniel and Martin, James H 2015 Classi- fication: Naive Bayes, Logistic Regression, Sentiment Speech and Language Processing.
9. Cortes, Corinna and Vapnik, Vladimir 1995. Support- vector networksMachine Learning vol. 20, pp. 273 - 297 Kluwer Academic Publishers,Boston. Manufactured in The Netherlands.
10. Tissaa Tony 2019 Faith and Emotional Intelligence International Journalof Recent Technology and Engineering (IJRTE) ISSN: 2277-3878,Volume-7 Issue-5S, January 2019.
11. https://unicode.org/emoji/charts/full-emoji-list.html

AUTHORS PROFILE

Soumya S. is a Research Scholar in Department of Computer Applications at Cochin University of Science and Technology. The research work is based on Sentiment Analysis and Emotion Analysis. The areas of interest include Natural Language Processing andMachine Learning. She has 15 years of teaching experience in Department of Computer Science, College of Engineering Munnar.

K. V. Pramod is an Emeritus Professor in Department of Computer Applications, Cochin University of Science and Technology. He is the research Guide in Faculty of Science and Faculty of Technology. His area of interest includes Language Computing, Cyber Forensics, Image Processing, and Mathematical Morphology. He has published several papers in international journal/conferences.