TADS@Dravidian-CodeMix-FIRE2020: Sentiment Analysis on CodeMix Dravidian Language

Deepesh Sharma[a]

[a] IIIT Kottayam, Kerela, India

Abstract

Sentimental analysis on Social Media has received much attention in research recently. Social Media will be the biggest source of big data in the upcoming years. Hence, the sentiment analysis of social media contents very important to regularize it. The FIRE 2020 organizers provided participants with annotated data-sets containing comments on YouTube videos in Malayalam and Tamil(including code-mixing). Approached the problem using classic machine learning algorithms for classification i.e. SVM, Perceptron, and Logistic classifier.

Keywords

Sentiment Analysis, Dravidian language, Text Classification,

1. Introduction

We are exploring the field of natural language processing, which is the broad study of how computers and machines can understand human to human communication and how texts are analyzed based on contextual information by machines.

Code-Mixing is a phenomenon where speakers switch between multiple languages in a single utterance[1, 2]. Code-Mixing is common in multilingual countries such as India[3, 4]. There is an increasing demand for sentiment analysis on social media texts which are largely code-mixed [5, 6]. Sentiment analysis is the interpretation and classification of emotions (positive, negative, and neutral) within text data using text analysis techniques. The machine learning model based on monolingual data fails on code-mixed data. As the usage of the internet growing amount of code-mix multilingual data is increasing. The mixing of the scripts in the code-mixing makes it even more complicated to use the model trained on monolingual corpora[7, 8, 9]. In this paper, I provide a classic Machine learning algorithms trained on code mixed multilingual data.

In this paper, we present a model which can be used to find the sentiment of a given text. We want to classify the text as 'Positive ', 'Negative ', 'Mixed feelings ', 'unknown state ', 'not-Tamil'

2. Data-set

[5, 6] Malayalam is one of the Dravidian languages spoken in the the southern region of India with nearly 38 million Malayalam speakers in India and other countries.

FIRE 2020: Forum for Information Retrieval Evaluation, December 16-20, 2020, Hyderabad, India

deepeshsharma2017@iiitkottayam.ac.in (D. Sharma)

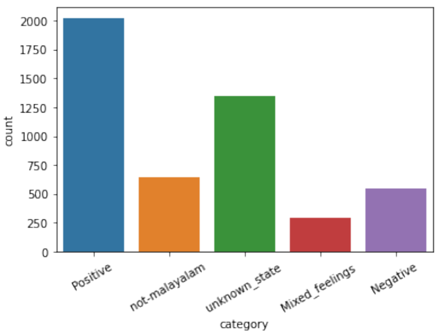

Figure 1: Malayalam data.

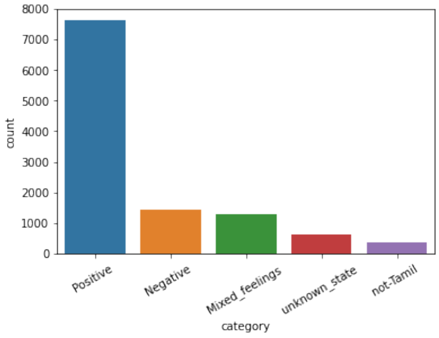

Figure 2: Tamil data.

Tamil, is a Dravidian language natively spoken by the Tamil people of India and Sri Lanka. Tamil is the official language of the South Indian state of Tamil Nadu, as well as two sovereign states, Sri Lanka and Singapore. For this shared task, we have been provided with a new gold standard corpus by the organizers for sentiment analysis of code-mixed text in Dravidian languages (Malayalam-English and Tamil-English). The data-set consists of YouTube comments which are then marked as one of the following.' Positive', 'Negative', 'Mixed feelings', 'unknown state', 'not-Tamil'.

The distribution of the dataset is below.

As we can see from the data the very skewed a simpler machine learning approach will be more generalized.

3. Task Description

This is a message-level polarity classification task[10, 11]. Given a YouTube comment, systems have to classify it into positive, negative, neutral, mixed emotions, or not in the intended languages.

4. Experiment Setup

We experimented with broadly three kinds of classic systems - an SVM classifier, a logistic classifier, and a Perceptron. We used the sci-kit learn implementation of SVM, Logistic Regression, and Perceptron. Support Vector Machines are one of the most successful classic machine learning models used for various kinds of text classification tasks. Used logistic regression with a multi-class variable as 'ovr' for multi-class classification. Perceptron is a single layer neural network and a multi-layer perceptron is called Neural Networks. We used a grid search for finding the best parameters for SVM algorithms.

For text to vector conversion, we used sklearn CountVectorizer. The CountVectorizer provides a simple way to both tokenize a collection of text documents and build a vocabulary of known words, but also to encode new documents using that vocabulary. I trained models separately for the Tamil dataset and the Malayalam dataset.

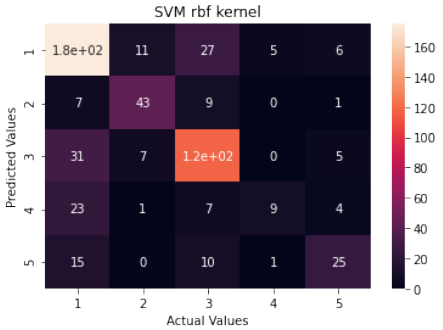

Figure 3: Malayalam test confusion matrix.

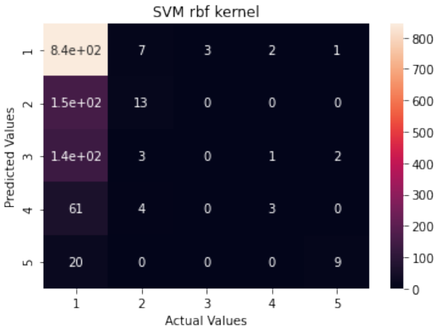

Figure 4: Tamil test confusion matrix.

Best hyperparameter for the Models after grid search.

Model	SVM	Logistic Reg	Perceptron
Hyper Parameters	C= 1, gamma= 0.01	C=1, max iter=100	alpha=0.0001

5. Results Analysis

This section presents the results of the evaluation of the three architectures. We compare the performance of the above machine learning architectures to select submissions for each language. Classification Accuracy is what we usually mean when we use the term accuracy. It is the ratio of the number of correct predictions to the total number of input samples.

Model	SVM	Logistic Reg	Perceptron
Accuracy Score	0.63	0.677	0.614

For further analysis, I used the confusion matrix. A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. It is extremely useful for measuring Recall, Precision, Specificity, Accuracy, and most importantly AUC-ROC Curve.

In confusion matrix figure 3 and figure 4, we can see that due to an unbalanced data-set many test cases were classified as negative 1.

6. Conclusion

In this paper, we have described how I trained machine learning algorithms for classification. Simple machine learning algorithms were fast to train and set the base for further research. For, future work we can train complex deep learning algorithms but we will need a more balanced dataset for complex deep learning algorithms.

References

[1] N. Jose, B. R. Chakravarthi, S. Suryawanshi, E. Sherly, J. P. McCrae, A survey of current datasets for code-switching research, in: 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), IEEE, 2020, pp. 136–141.

[2] R. Priyadharshini, B. R. Chakravarthi, M. Vegupatti, J. P. McCrae, Named entity recognition for code-mixed indian corpus using meta embedding, in: 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), IEEE, 2020, pp. 68–72.

[3] B. R. Chakravarthi, R. Priyadharshini, B. Stearns, A. Jayapal, S. S, M. Arcan, M. Zarrouk, J. P. McCrae, Multilingual multimodal machine translation for Dravidian languages utilizing phonetic transcription, in: Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages, European Association for Machine Translation, Dublin, Ireland, 2019, pp. 56–63. URL: https://www.aclweb.org/anthology/W19-6809.

[4] B. R. Chakravarthi, M. Arcan, J. P. McCrae, 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, European Association for Machine Translation, Dublin, Ireland, 2019, pp. 1–7. URL: https://www.aclweb.org/anthology/W19-7101.

[5] B. R. Chakravarthi, N. Jose, S. Suryawanshi, E. Sherly, J. P. McCrae, A sentiment analysis dataset for code-mixed Malayalam-English, in: Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL), European Language Resources association, Marseille, France, 2020, pp. 177–184. URL: https://www.aclweb.org/anthology/2020.sltu-1.25.

[6] B. R. Chakravarthi, V. Muralidaran, R. Priyadharshini, J. P. McCrae, Corpus creation for sentiment analysis in code-mixed Tamil-English text, in: Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL), European Language Resources association, Marseille, France, 2020, pp. 202–210. URL: https://www.aclweb.org/anthology/2020.sltu-1.28.

[7] B. R. Chakravarthi, M. Arcan, J. P. McCrae, 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, 2019.

[8] B. R. Chakravarthi, Leveraging orthographic information to improve machine translation of under-resourced languages, Ph.D. thesis, NUI Galway, 2020.

[9] B. R. Chakravarthi, P. Rani, M. Arcan, J. P. McCrae, A survey of orthographic information in machine translation, arXiv preprint arXiv:2008.01391 (2020).

[10] B. R. Chakravarthi, R. Priyadharshini, V. Muralidaran, S. Suryawanshi, N. Jose, E. Sherly, J. P. McCrae, Overview of the track on Sentiment Analysis for Dravidian Languages in Code-Mixed Text, in: Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2020). CEUR Workshop Proceedings. In: CEUR-WS. org, Hyderabad, India, 2020.

[11] B. R. Chakravarthi, R. Priyadharshini, V. Muralidaran, S. Suryawanshi, N. Jose, E. Sherly, J. P. McCrae, Overview of the track on Sentiment Analysis for Dravidian Languages in Code-Mixed Text, in: Proceedings of the 12th Forum for Information Retrieval Evaluation,

FIRE ’20, 2020.