Sentiment Analysis on Tamil texts using K-means and K-Nearest Neighbor

Sajeetha Thavareesan, Sinnathamby Mahesan
sajeethas@esn.ac.lk, mahesans@univ.jfn.ac.lk

SLSWCS Student Workshop on Computer Science

① Introduction

Sentiment Analysis is an application of Natural Language Processing which identifies and categorises the opinions into positive or negative.
In our model,

Bag of Words (BoW) and fastText vectors are used to represent features. These features are clustered using K-means clustering and the cluster centers are used to build the Sentiment Analysis model using K-Nearest Neighbour (K-NN).

BoW is used to represent the number of times a word appears in a document.
fastText treats each word as composed of character ngram. The vector for a word is made of the sum of the character ngram. Each word is represented using a 300 dimension vector.

② Problem Specification

The aim of this research is to build a suitable model with less number of training samples to perform Sentiment Analysis in Tamil text.

③ Contribution

Constructed *UJ_Corpus_Opinions* corpus to tackle the inavailability of the opinion corpus, that contains 1518 positive and 1173 negative reviews and comments.
Proposed three models to perform Sentiment Analysis:

- *Model1*: *Sentiment classification using K-NN*
- *Model2*: *Sentiment classification using K-means clustering with K-NN*
- *Model3*: *Sentiment classification using class-wise K-means clustering with K-NN*

④ Methodology

Three models are built using two types of feature vectors: *BoW* and *fastText*. *UJ_Corpus_Opinions* corpus is used to train and test these three models.

Model1: In this model K-NN is used as the classifier. K-NN is trained and tested on *UJ_Corpus_Opinions* corpus. Accuracy of this model is evaluated for different number of neighbours `Kn` in K-NN.

Model2: In this model feature vectors of training set are clustered using K-means with various number if clusters `Km` and the cluster centers are used to train K-NN.

Model3: Training set is split into groups based on class label, and these feature vectors of these groups are created and clustered these groups separatly using K-means clustering. We have tested this approach with different values of `Kn` and `Km`.

The general structure of *Model2* and *Model3* is described in Figure 1.

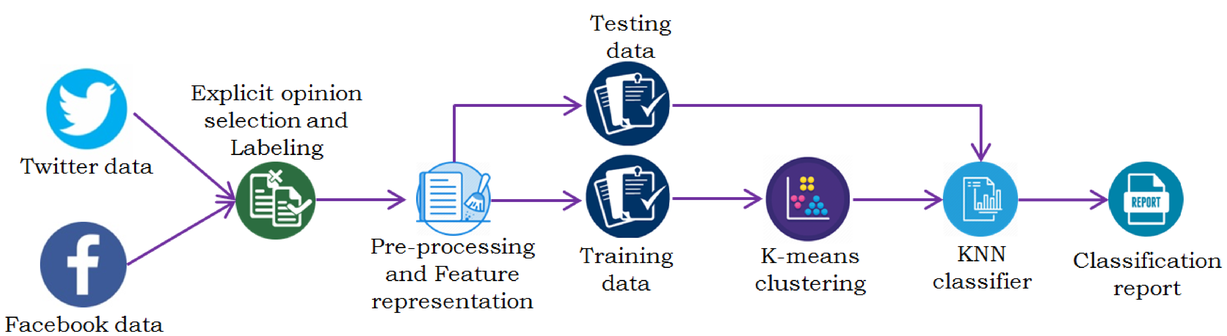
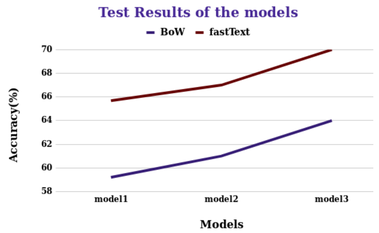


Figure 1: Structure of Model2 and Model3

⑤ Results

Table1: Tests results of the models

Model	BoW			fastText		
	Accuracy	Kn	Km	Accuracy	Kn	Km
Model1	59	1	-	66	1	-
Model2	61	1	10	67	1	10
Model3	64	1	6	70	1	16



Test results of three models are listed in Table1. 70 as the highest accuracy is found for *Model3*.

⑥ Discussion and Conclusion

- We considered *Model1* as our base model and obtained 59% and 66% of accuracies for *BoW* and *fastText* feature vectors.
- We tested the models using different values of `Km` to check their influence in the accuracy and noticed that the accuracy increases with the values of `Km`.
- In *Model2* and *Model3* we used centroids as training set for K-NN and obtained better results compared with *Model1*. We obtained 61% and 67% of accuracies for *Model2* as we used centroids as training set of K-NN.
- *Model3* outperformed other two models as we used centroids of class-wice K-means clustering to train K-NN. It shows that class-wise clustering performs better than global clustering. Highest accuracy is found for *Model3* for both features *BoW* (64%) and *fastText* (70%) .
- High accuracy occurred with `Kn=1` in K-NN for all three models.
- *fastText* as features gives better results than BoW for all `Km`.
- Furthermore, *Model3* outperformed the other models for all `Km`. Thus, *fastText* and class-wise clustering with increased number of clusters can be used to classify the sentiments expressed in the Tamil text.

⑦ References

[1] E. Nivedhitha, S. P. Sanjay, M. Anand Kumar, and K. P. Soman. Unsupervised word embedding based polarity detection for tamil tweets. International Journal of Computer Technology and Applications (IJCTA), 9(10):4631–4638, 2016.
[2] Shanta Phani, Shibamouli Lahiri, and Arindam Biswas. 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, 2016.
[3] Braja Gopal Patra, Dipankar Das, Amitava Das, and Rajendra Prasath. Shared task on sentiment analysis in indian languages (sail) tweets- an overview. In International Conference on Mining Intelligence and Knowledge Exploration, pages 650–655. Springer, 2015.
[4] N. Ravishankar and R. Shriram. Corpus based sentiment classification of tamil movie tweets using syntactic patterns. IIOAB Journal: A Journal of Multidisciplinary Science and Technology, 8(2):172–178, 2017.