Unsupervised Sentiment Analysis on Tamil Texts

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

① Introduction

Sentiments are central to almost all human activities and are key influencers of individuals or organizations. It is in the core of understanding sentiments expressed in the text. It groups opinions of written text into *positive*, *negative* or *neutral*.
Several algorithms are there in grouping opinions into positive or negative or neutral. This study uses clustering and lexicon based approaches to determine a suitable approach to performing Sentiment Analysis in Tamil text using a corpus and lexicon.

Clustering is the task of grouping a set of objects in such a way that objects in the same group are similar to each other with respect to certain features. There may be several clusters in a set of objects.

Bag of Words is the representation of the words by their counts appeared in a document.

② Corpus and Lexicon

UJ_Corpus_Opinions corpus consists of reviews and comments with tags of positive/negative [Positive- 1518 and negative- 1173].
For example,

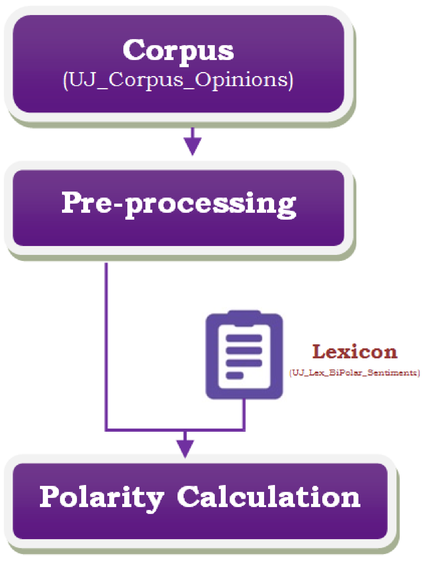
- பயிற்சி ஆட்டத்தில் நியூசிலாந்து அணி வெற்றி 😀 ஆரம்பமே ரொம்ப அமர்களமாயிருக்கு 😁😁😁
- படம் முன் பாதியைக் காட்டிலும் ,பின் பாதி காட்சிகள் வேக வேகமாக ஜம்ப் ஆவது சற்றே பலவீனம்

UJ_Lex_BiPolar_Sentiments lexicon consists of words and emojis expressing sentiments with tags positive/negative [Positive- 820 and negative- 1237].
For example,

- சூப்பர், பெரிய, சிறந்த, கேவலம், குறைவான, எதிரான

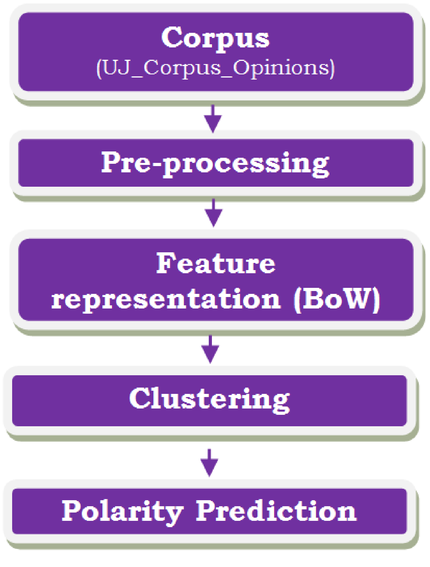
③ Methodology

In this work three approaches are experimented: Lexicon based, K-means with BoW and K-modes with BoW. Approaches are trained and tested using *UJ_Corpus_Opinions* corpus taking 70% and the remaining 30%.



Algorithm 1 Lexicon based approach

Require: UJ_Lex_BiPolar_Sentiments lexicon(UJLexicon)
UJ_Corpus_Opinions (UJCorpus)
Ensure: Accuracy (Acc)
Step1: for each comment ∈ UJCorpus do
words ← tokenised comment
Step2: Initialise variables: pos, neg for positive and negative polarity count with zero.
pos ← 0
neg ← 0
Step3: for each word ∈ words do
if word in UJLexicon then
if Polarity(word) is Positive:
pos← pos+1
else: neg← neg+1
Step4: polarity ← positive if pos>neg else negative
Step5: $\text{Acc} \leftarrow \frac{\text{No. of correctly classified comments}}{\text{Total no. of comments in the corpus}} \times 100$



Algorithm 2 K-means with BoW approach

Require: UJ_Corpus_Opinions (UJCorpus)
Ensure: Accuracy (Acc)
Step1: Split UJCorpus into training and testing sets
Step2: for each comment ∈ UJCorpus do
words ← tokenised comment
Step3: feature vector (BoW)
Step4: centroids ← K-means cluster (K=2) and feature vector (train)
Step5: for each vector ∈ feature vector (test) do
distances ← euclidean(centroids, vector)
polarity ← Label of the centroid with minimum euclidean distance
Step6: $\text{Acc} \leftarrow \frac{\text{No. of correctly classified comments}}{\text{Total no. of comments in the corpus}} \times 100$

K-modes with BoW approach: This approach is same as K-means with BoW approach but here instead of means mode is used.

④ Results

Tests results of the three approaches:

Approach	Accuracy
Lexicon based approach	57
K-means with BoW approach	61
K-modes with BoW approach	62

⑤ Discussion and Conclusion

Lexicon based Sentiment Analysis approach gives low accuracy (57%) compared with other two models due to the limited size of the lexicon. We are working on this to increase the accuracy of this approach by enhancing lexicon. K-modes with BoW based approach achieved highest accuracy of 62%.

In K-means and K-modes approaches one centroid is used to represent positive class and the other one is used to represent negative class. Each class contains the texts with different patterns thus, these two centroids in both models failed to capture the patterns of the two classes. Increased number of cluster centers K can be used to capture different patterns of text.

⑥ References

[1] E. Nivedhitha, S. P. Sanjay, M. Anandkumar, 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] B. G. Patra, D. Das, A. Das, and R. 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.

[3] S. Phani, S. Lahiri, and A. 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.