"Did you really mean what you said?" : Sarcasm Detection in Hindi-English Code-Mixed Data using Bilingual Word Embeddings

Akshita Aggarwal, Anshul Wadhawan, Anshima Chaudhary and Kavita Maurya

Department of Computer Engineering, Netaji Subhas University of Technology

Introduction

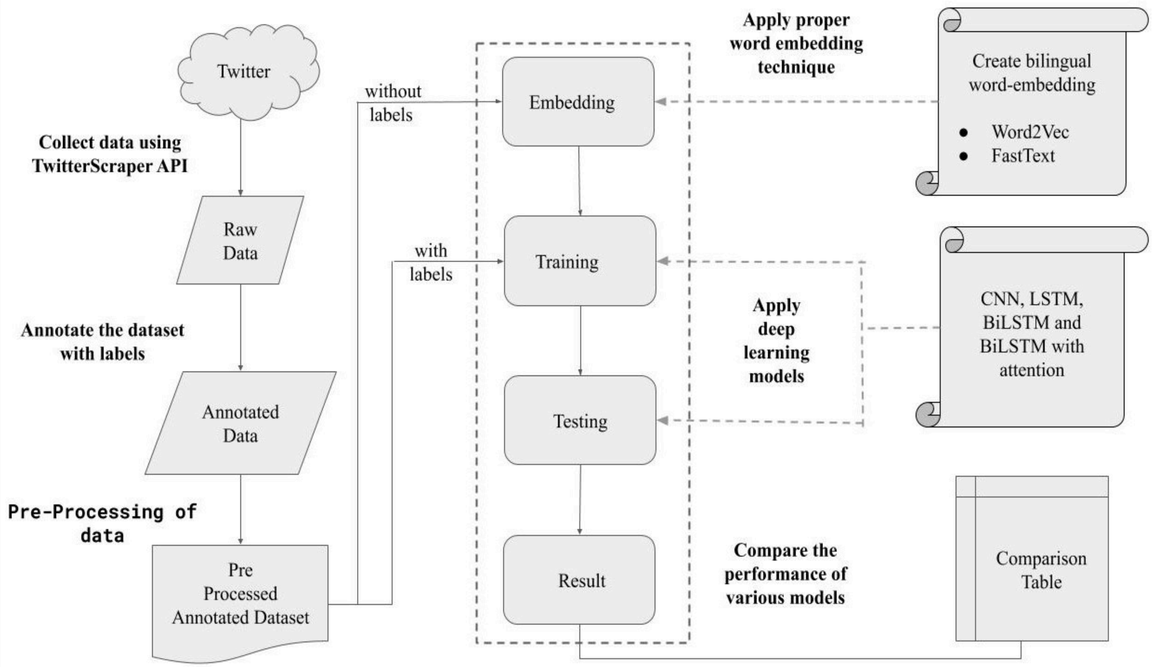
In this paper, we have experimented with deep-learning approaches to detect sarcasm in the Hindi-English code mixed dataset using bilingual word embeddings derived from FastText and Word2Vec approaches. A labelled Hinglish dataset for sarcasm detection is released as a part.

Results

•All the proposed deep learning models performed better than the traditional state-of-the-art models, where the attention based Bi-directional LSTM network produced the best accuracy of 78.49%

•Word2Vec embeddings produce better results than FastText embeddings, for all the models.

Proposed Methodology



Conclusions

•In this paper, we presented a class-balanced Hindi-English code mixed dataset for the problem of sarcasm detection, by scraping relevant tweets from twitter.

•We compared two representations, FastText and Word2Vec, both based on different word representation learning mechanisms and trained on custom scraped data from scratch.

•We analysed the performance of different deep learning models, which take as input the generated word embeddings, to solve the problem of sarcasm detection.

Literature cited

• Sahil Swami, Ankush Khandelwal, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018.A corpus of English Hindi code-mixed tweets for sarcasm detection..

Major Challenges :

•Lack of clean data and linguistic complexities associated with code-mixed data.

•Lack of large labelled datasets

Future work

As future scope, the problem can be solved to obtain even better results by carrying out a comparison of MUSE aligned vectors, pre-aligned FastText word embeddings and BERT word embeddings.