Fake News Detection using Machine Learning and Natural Language Processing

Group Number : 5
Palak Tiwari (MT20103), Deepankar Kansal (MT20007), Vineet Maheshwari (MT20020)

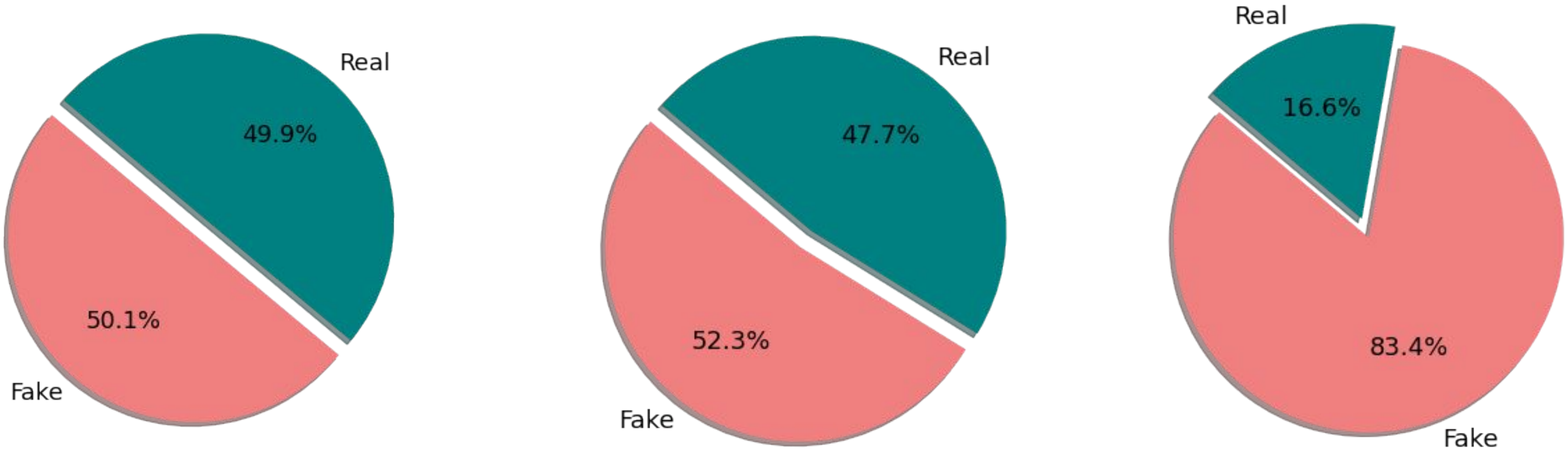
Previous Work

- Worked on one dataset
- Text preprocessing : Tokenization, Stop-words, Stemming
- Feature Extraction using Bag of Words
- Models like:

  Logistic Regression (92%), SVM (92%) and

  Decision Tree (88%)

Datasets Description



We used three different datasets for evaluation

Main Libraries Used:

For text preprocessing and feature extraction:

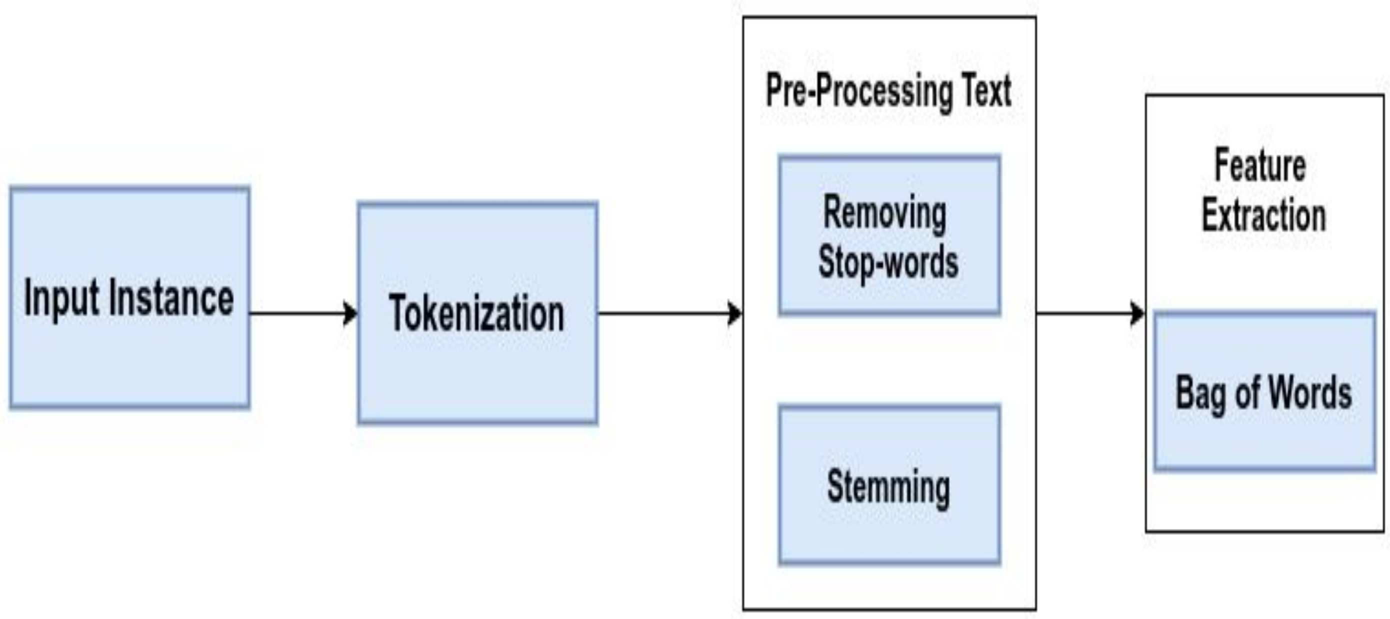
1. nltk (Natural Language Toolkit):- Like Corpus, PorterStemmer
2. re (Regular Expression)
3. CountVectorizer class from SciKit-learn

Classification Models Applied: From SciKit-learn

Logistic Regression (LR), Support Vector Machines (SVM), Gradient-Boosting , Multinomial Naive Bayes

Bi-LSTM from keras module using Glove as feature extraction

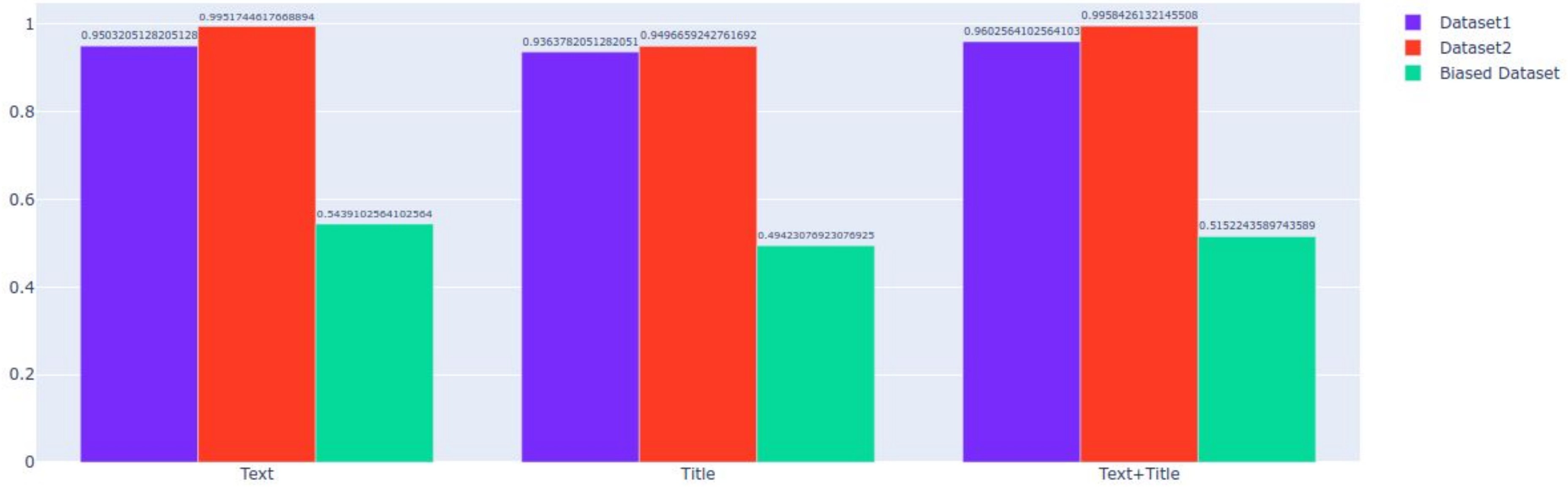
Text Preprocessing



Bag of words made using Unigram and Bi-gram combinations

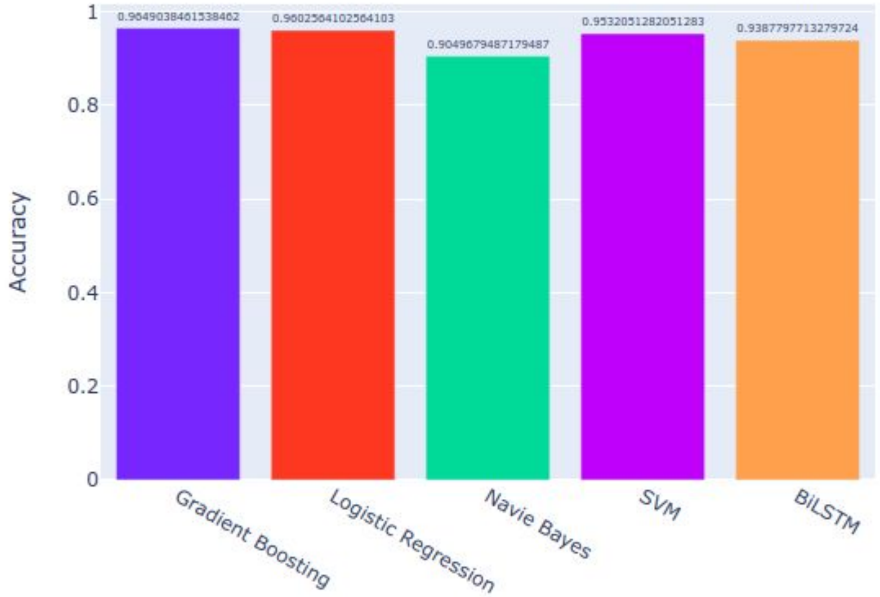
Using libraries nltk, re, and CountVectorizer (class)

Evaluation on Variations

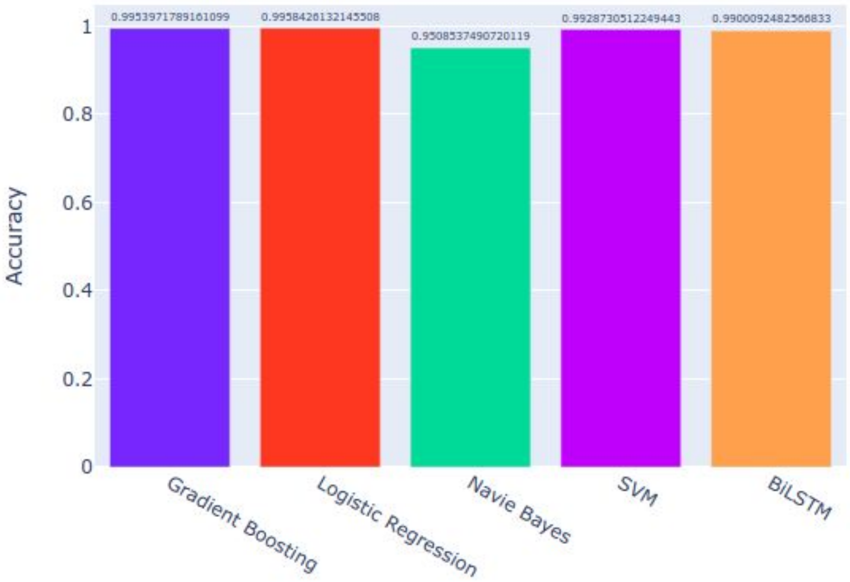


We can see that Text+Title combination used for building the corpus and training the model, gives us the best accuracy

Models Evaluation and Comparison



Accuracy plot for Dataset1



Accuracy plot for Dataset2

We have tried various Machine Learning algorithms and found that, these 5 algorithms are giving encouraging accuracies.

Conclusion

Best evaluation models found are Gradient Boosting and Logistic Regression(LR), this is happening because features that we are extracting from the corpus are well separated that LR and Gradient Boost are able to capture the context.

Future Aspects

Various emerging algorithms and neural nets using feature extraction techniques like BERT (are able to capture the semantics of the text) can be used and checked for accuracies.

Thank You