A Project Report on

Bot or Human? A Natural Language Processing Approach To Text Classification

Group Members:

Srivatsa Setu (202218021)
Zeel Gudhka (202218025)
Radhika Singhal (202218062)

Under the Guidance of:
Prof. Sourish Das Gupta

Text Preprocessing

- Lowercased, Lemmatized and removed all characters from the string text that are not alphabets or whitespace characters.
- Did not remove Stop words because it may help in classifying the text accurately.

Label Imbalance

SMOTE (Synthetic Minority Over-sampling Technique): It is an oversampling technique used in machine learning to address the problem of class imbalance by generating synthetic samples for the minority class.

Models

Applied word embedding techniques on the models mentioned below:

- Logistic Regression
- Random Forest Classifier
- SVM
- 1-Layer MLP
- Decision Tree Classifier

Word Embeddings

Word2Vec (SkipGram): Word2Vec Skip-gram is a variant of the Word2Vec algorithm that predicts the surrounding words given a target word to generate high-quality word embeddings that capture semantic relationships between words.

Word2Vec Continuous Bag of Words (CBOW): It is a variant of the Word2Vec algorithm that predicts a target word based on the surrounding words in a context window to generate high-quality word embeddings that capture semantic relationships between words.

Doc2Vec: Doc2Vec is an unsupervised algorithm for generating document embeddings that capture the overall semantic meaning of a document and its constituent words.

GloVe (Global Vectors for Word Representation): is an unsupervised algorithm that uses co-occurrence statistics to generate word embeddings that capture both semantic and syntactic relationships between words in a corpus of text.

BERT (Bidirectional Encoder Representations from Transformers): is a pre-training technique for natural language processing (NLP) that trains a deep bidirectional transformer model on a large corpus of text to generate contextualized word embeddings, which can be fine-tuned for a wide range of NLP tasks.

Accuracy Table

Word Embedding	Word2Vec(SkipGram)	Doc2Vec	Word2Vec(Continuous Bag of Words)
Logistic Regression	80.19%	78%	76.07%
Random Forest Classifier	80.75%	78.6%	77.01%
SVM	79.91%	81.31%	79.63%
1-Layer MLP	85.79%	80.93%	79.25%
Gradient Boosting Classifier	81.87%	78.50%	78.04%

Word2Vec(Skip-Gram) seems to perform slightly better than Doc2Vec and Continuous Bag of Words which contradicts our expectations that Doc2Vec performs better than Word2Vec
Word2Vec performs the best on 1-Layer MLP
Doc2Vec performs the best on SVM
Continuous Bag of Words performs the best on 1-Layer MLP

Accuracy Table for Over Sampled Text

Word Embedding	Word2Vec(SkipGram)	Doc2Vec	GloVe
Logistic Regression	75.2%	76.1%	66.6%
Random Forest Classifier	97.3%	90%	88.7%
1-Layer MLP	83.6%	91.5%	79.6%
SVM	80.7%	85.4%	76%
Decision Tree Classifier	95.5%	75.3%	87.9%

Word2Vec gives the best accuracy on Random Forest Classifier.
Doc2Vec gives the best accuracy on 1-Layer MLP.
GloVe gives the best accuracy on Random Forest Classifier.

RNN Model : We implemented a Recurrent Neural Network (RNN) model for our task. During training, the RNN model showed significant improvement in performance across epochs. The loss function decreased from an initial value of 0.5500 to a final value of 1.2475e-04, indicating that the model was effectively learning and making accurate predictions. The accuracy of the model increased from 0.7224 to 1.0000, demonstrating its ability to classify the data correctly. Additionally, we evaluated the precision and recall of the RNN model. The precision metric, which measures the ability of the model to correctly identify positive cases, improved from 0.7330 to 1.0000. The recall metric, which measures the model's ability to identify all positive cases, also showed improvement, increasing from 0.9625 to 1.0000.

Bert Model : We implemented Bert and our accuracy rose from 58% to 62% for 3 epochs(on undersampling).As we increase no of epochs we can improve our accuracy.

THANK YOU!