OFFENSIVE SPEECH CLASSIFICATION USING SVM AND CNN MODELS

CE807 – ASSIGNMENT 2 - 2213409

OVERVIEW

- Objective: Classify offensive speech using the OLID dataset
- Models: Support Vector Machine (SVM) and Convolutional Neural Network (CNN)
- Experiments on different data sizes (25%, 50%, 75%, and 100% of training data)
- Evaluation metrics: accuracy, precision, recall, and F1-score

Links to be used:

Code: https://colab.research.google.com/drive/1an9JnhrIM8-ykm-6mBVMTtjygub49n0G?authuser=2#scrollTo=UWRetx31Sumx

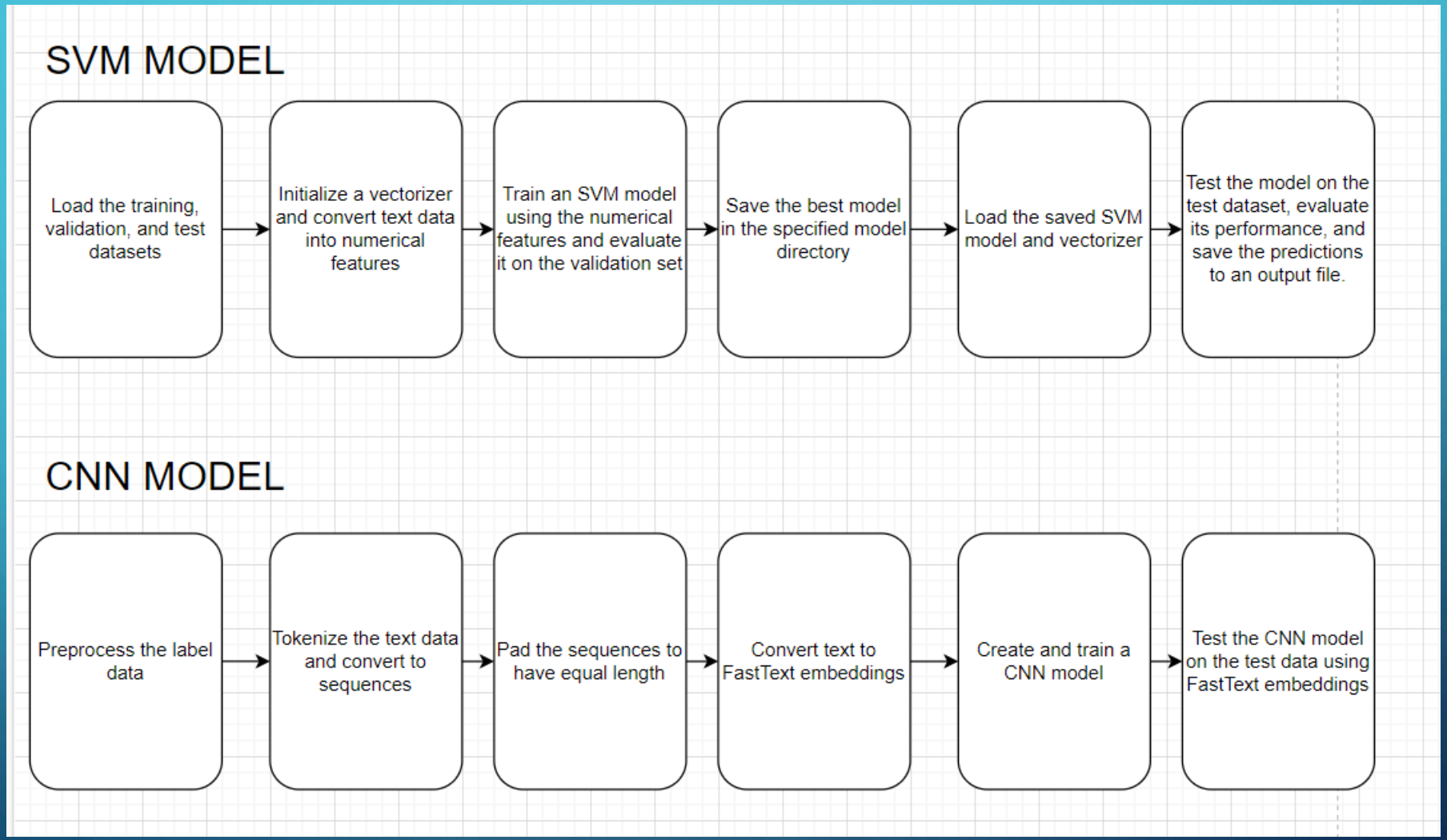
Google drive folder:
https://drive.google.com/drive/folders/1PpTsYYDvVtx4ZUm24Qk0XD8IkKw1gp15?usp=share_link

SVM MODEL

- Preprocessing: lowercasing, removing stopwords and punctuations
- Tokenization and TF-IDF weighting
- Trained on various data sizes (25%, 50%, 75%, 100%)
- Performance improves with larger data sizes

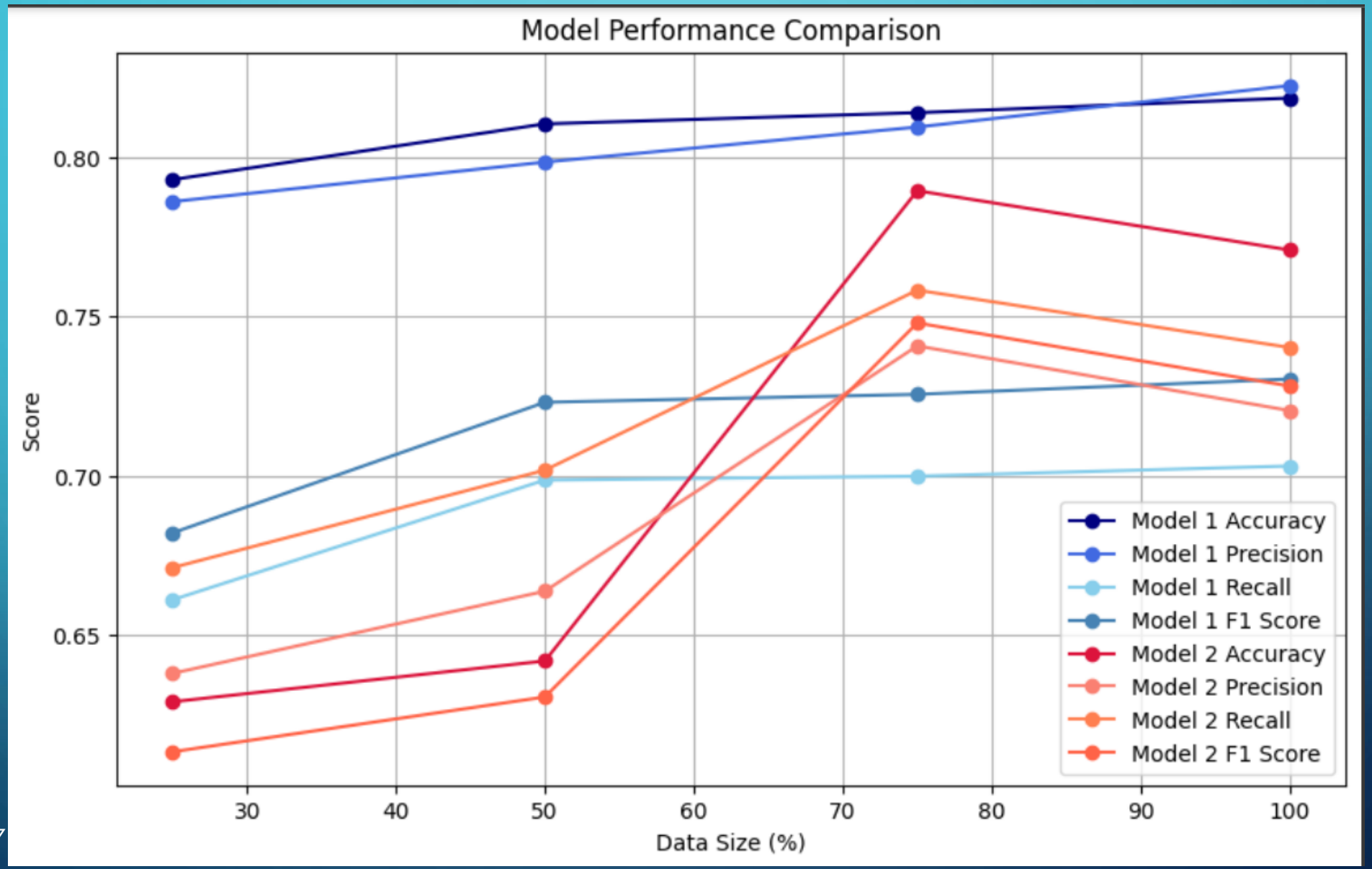
CNN MODEL

- Model architecture: Dense layer, Conv1D layer, GlobalMaxPooling1D layer, Dropout layer
- Tokenization, sequence padding, and FastText embeddings
- Class weights to handle class imbalance
- Trained on various data sizes (25%, 50%, 75%, 100%)
- Mixed performance trend with increasing data sizes


SVM MODEL
Load the training, validation, and test datasets
Initialize a vectorizer and convert text data into numerical features
Train an SVM model using the numerical features and evaluate it on the validation set
Save the best model in the specified model directory
Load the saved SVM model and vectorizer
Test the model on the test dataset, evaluate its performance, and save the predictions to an output file.
CNN MODEL
Preprocess the label data
Tokenize the text data and convert to sequences
Pad the sequences to have equal length
Convert text to FastText embeddings
Create and train a CNN model
Test the CNN model on the test data using FastText embeddings

PERFORMANCE COMPARISON

- SVM consistently outperforms CNN in terms of accuracy, precision, recall, and F1-score
- CNN shows potential for improvement with larger training data, but performance plateaus or decreases for certain metrics
- Importance of understanding the strengths and weaknesses of different algorithms in varying data conditions


Model Performance Comparison
Score
Data Size (%)
0.80
0.75
0.70
0.65
30
40
50
60
70
80
90
100
Model 1 Accuracy
Model 1 Precision
Model 1 Recall
Model 1 F1 Score
Model 2 Accuracy
Model 2 Precision
Model 2 Recall
Model 2 F1 Score

INSIGHTS & IMPLICATIONS

- SVM is better suited for offensive speech classification on the provided dataset, especially when limited training data is available
- CNN could potentially perform better with larger training data, but may require further fine-tuning or architectural adjustments
- Performance metrics should be carefully considered when evaluating models, as the optimal trade-off may vary depending on the problem domain
- Insights can guide model selection, tuning, and experimentation for future applications in offensive speech classification and similar text classification tasks

CONCLUSION

- SVM is recommended for offensive speech classification using the OLID dataset
- Future work could explore alternative deep learning architectures or techniques to improve CNN performance
- Model selection should consider the specific problem domain, available data, and performance metrics to achieve optimal results