Hate Speech Classifier

Sirash Phuyal, Sam Macy,
Sushen Kolakaleti

Introduction

- Hate speech become common social platforms triggers the need for more effective detection mechanisms.

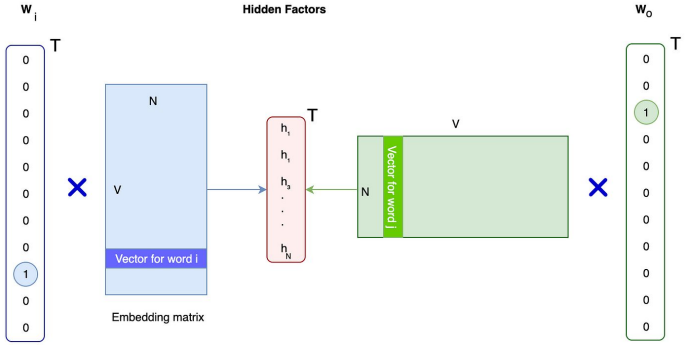
- Project attempts classification of hate speech in tweets, leveraging three distinct embedding approach: BERT-based and GloVe/word2vec

- This threefold embedding exploration aims to offer a comparative analysis on the efficacy of each approach for hate speech detection.

Methodology

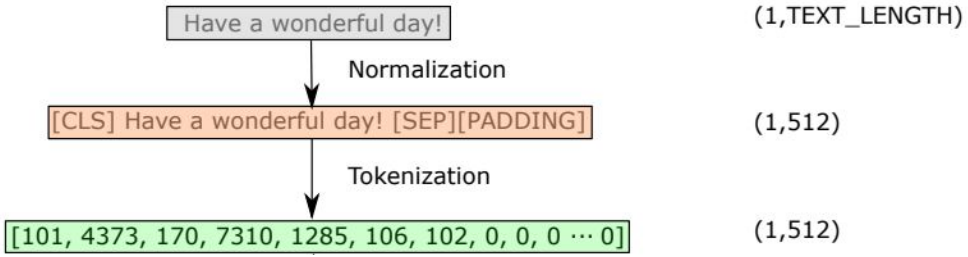
- Dataset processing / cleaning:

  - Removing stop-words for GloVe/word2vec
  - Majority voting of labels
    (Dataset contained multiple labels for many tweets)

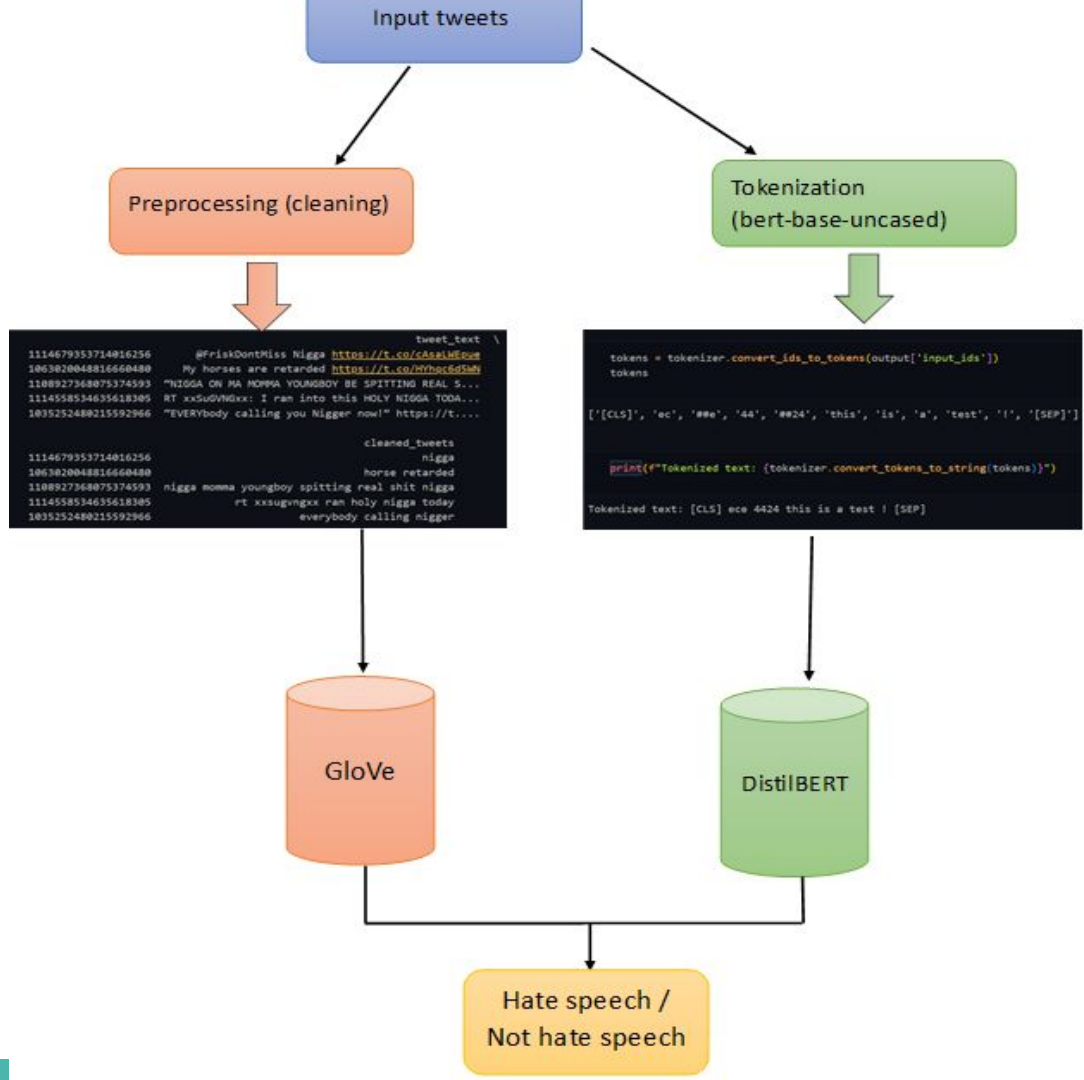


- BERT (Bidirectional Encoders Representation from Transformers)

  - Context-rich pretrained model
  - Fine-tuned to our dataset, resource-intensive, used condensed version called DistilBERT

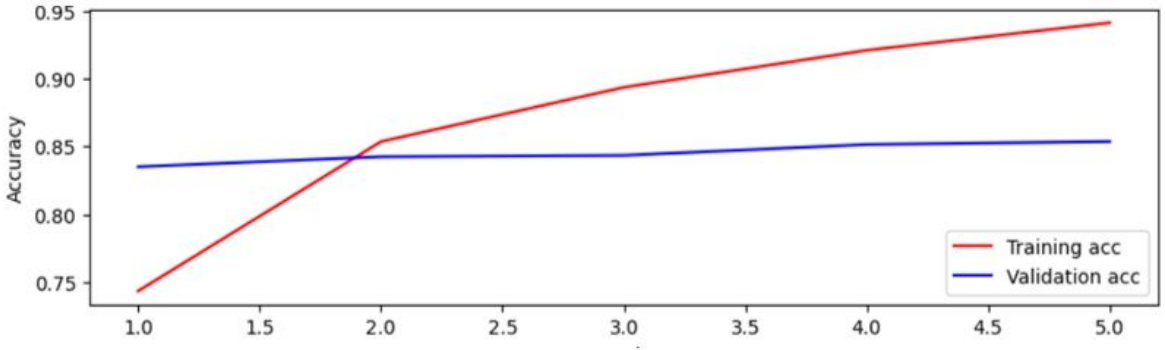


Process


Input tweets
Preprocessing (cleaning)
Tokenization
(bert-base-uncased)
tweet_text
1114679353714016256 @FriskDontMiss Nigga https://t.co/cAAeiWEpwe
1063020048816660480 My horses are retarded https://t.co/HYhqc6dSWN
1108927368075374593 "NIGGA ON MA MOMMA YOUNGBOY BE SPITTING REAL S...
1114558534635618305 RT xxSuGVNGxx: I ran into this HOLY NIGGA TODA...
1035252480215592966 "EVERYbody calling you Nigger now!" https://t....
cleaned_tweets
1114679353714016256 nigga
1063020048816660480 horse retarded
1108927368075374593 nigga momma youngboy spitting real shit nigga
1114558534635618305 rt xxsugvngxx ran holy nigga today
1035252480215592966 everybody calling nigger
tokens = tokenizer.convert_ids_to_tokens(output['input_ids'])
tokens
['[CLS]', 'ec', '##e', '44', '#024', 'this', 'is', 'a', 'test', '!', '[SEP]']
print(f"Tokenized text: {tokenizer.convert_tokens_to_string(tokens)}")
Tokenized text: [CLS] ece 4424 this is a test ! [SEP]
GloVe
DistilBERT
Hate speech /
Not hate speech

Results & Analysis

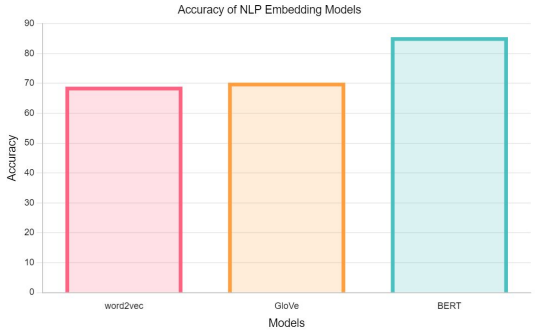
- Word2vec: 73.57% training accuracy, 68.92% testing accuracy
- GloVe: 69.71% training accuracy, 70.24% testing accuracy
- BERT: 45.34% loss, 85.48% testing accuracy



Training vs validation accuracy: Validation consistent around 85%

Conclusion

- Performance is highly dependent on the quality of the dataset.
- BERT performance and accuracy better than GloVe/word2vec
- BERT initially applied stop-word removal and lemmatization before tokenization, it is context-aware
- Potentially even better performance with a larger scale Deep NN model



Any Questions