Sentiment Analysis

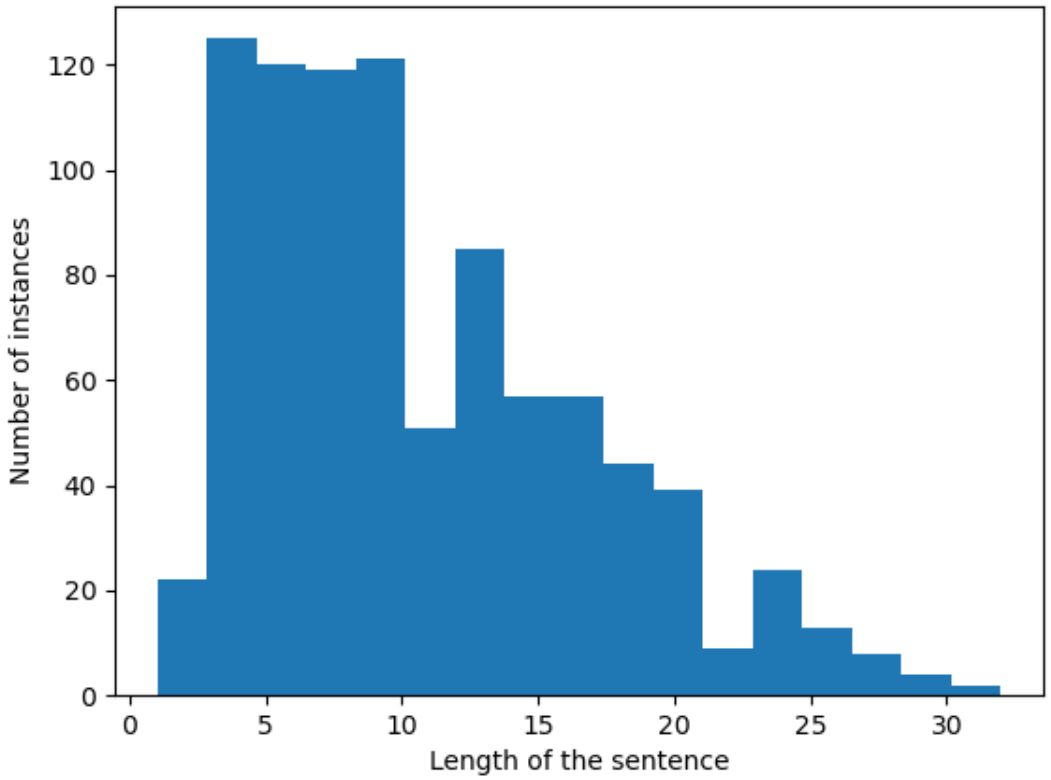
CHANTROUX G. **NUZZO M.** **PIERRE J.**

Plan

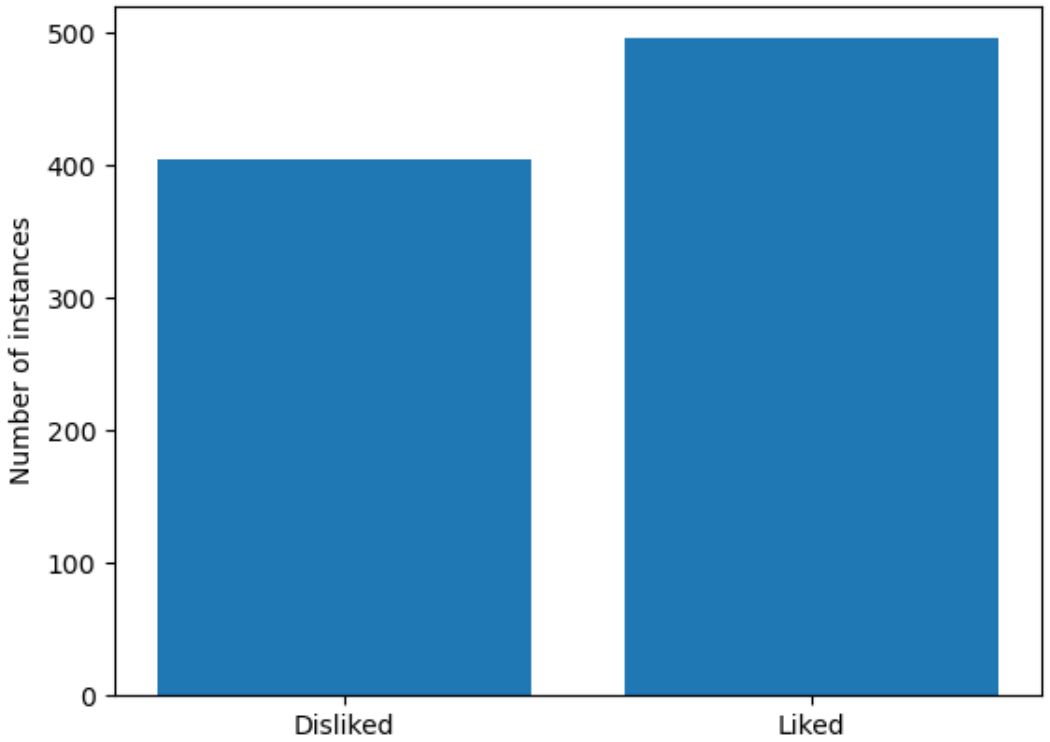
I. **Text Preprocessing**
II. **Models**
III. **Results**
    - **Simple Model**
    - **Attention Model**

EDA

Histogram of the length of the sentences



Data Imbalance



Data preprocessing

1. Taking care of specific issues (misspelled words, name of places, …)
   (Food was so gooodd . ---> Food was good .)
2. Dealing with contractions (I'd love to go back. ---> I would love to go back.)
3. Lower case letters (I would love to go back. --> i would love to go back)
4. Stop words (exception for not, pronouns, …)
5. Getting rid of non-alphabetical characters
6. Taking care of spaces

Word embeddings

Word2Vec

- Based on a neural network
- Predicts a word based on the context word or viceversa
- Each word is represented as a vector which contains semantic links between words

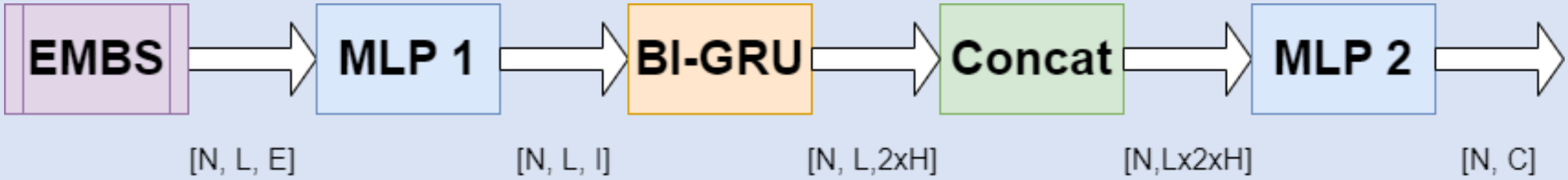
GloVe

- Based on the co-occurrence matrix which takes into account the context
- Considers the entire corpus's word global co-occurrence
- Captures both semantic and syntatic relationships betwenn words

FastText

- Introduces sub-word representation (wrt word2vec)
- Higher computational efficiency in words prediciton
- Suitable for out of vocabulary words

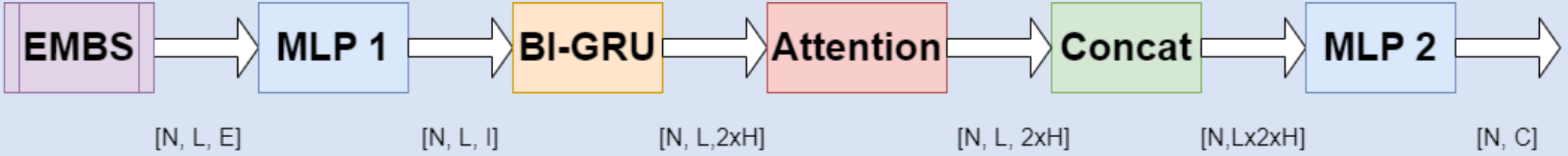
Models – Simple model



N: Number of batches
L: Length of the sequence
I: Input shape of the GRU
2xH: Output shape of GRU
C: Number of classes

Attention model

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

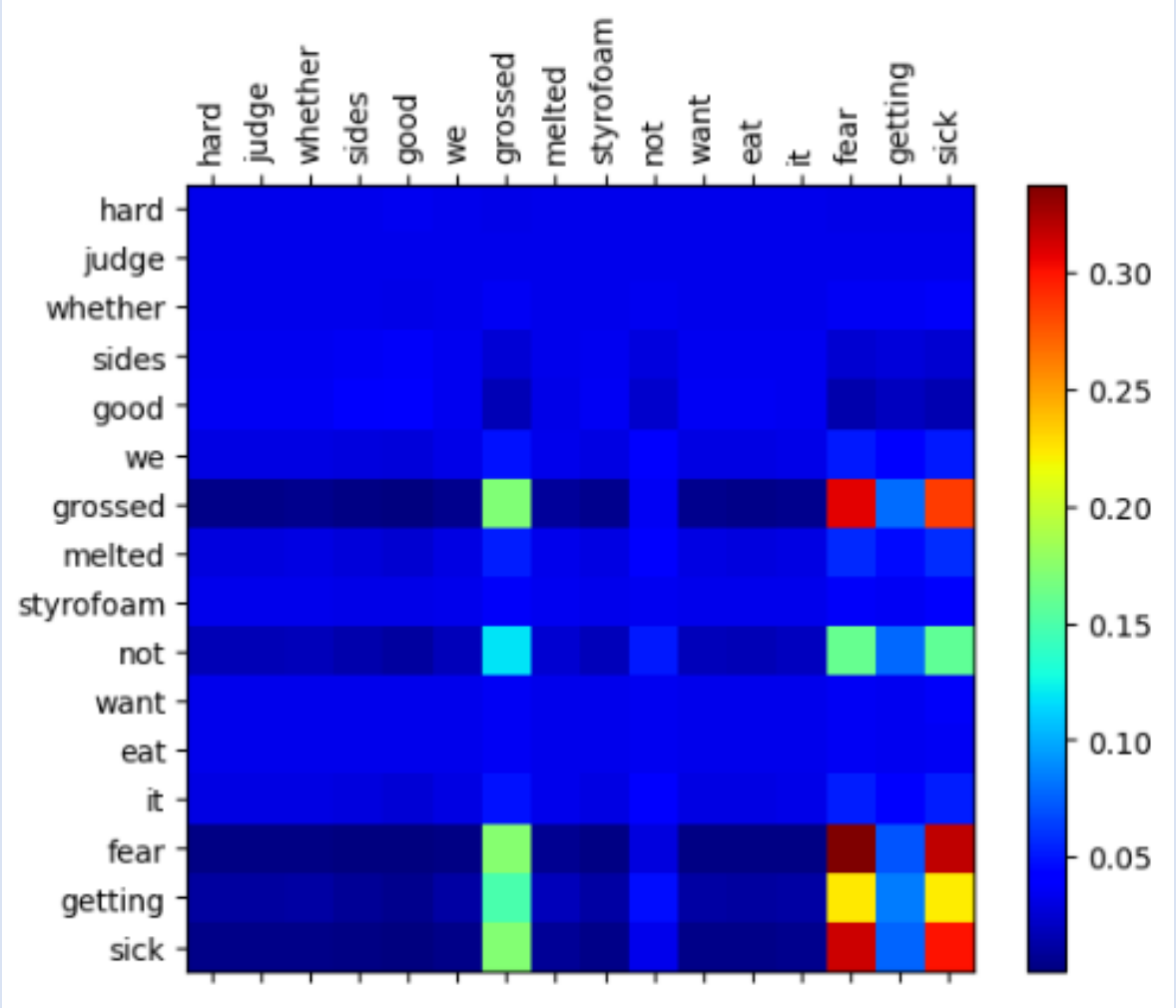


N: Number of batches
L: Length of the sequence
I: Input shape of the GRU
2xH: Output shape of GRU
C: Number of classes

Attention model

Hard to judge whether these sides were good because we were grossed out by the melted styrofoam and didn't want to eat it for fear of getting sick.

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

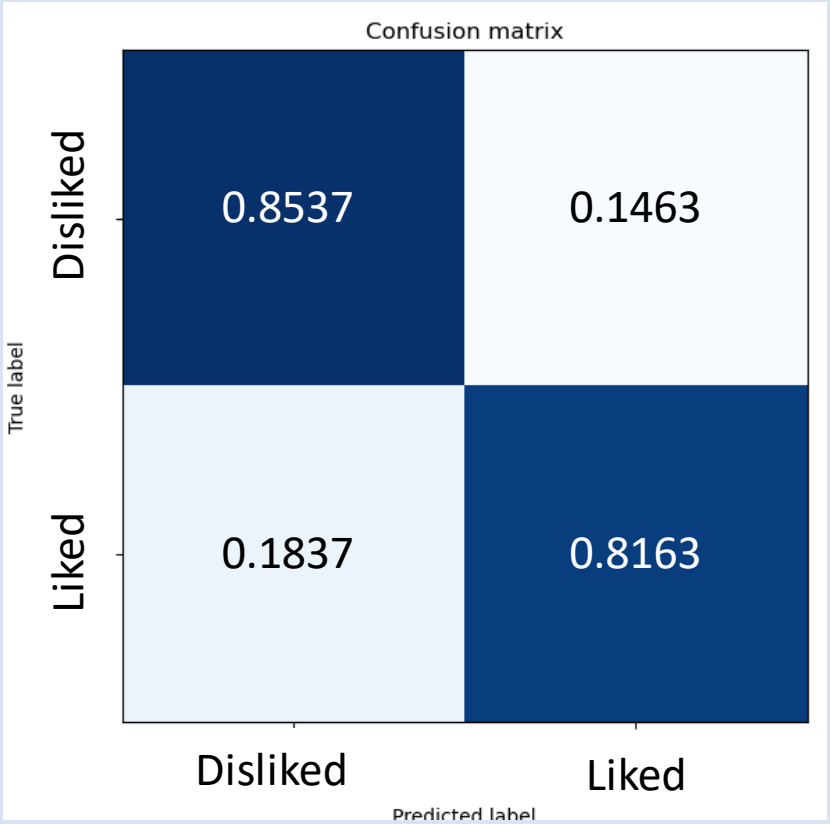


Training

- ADAM optimizer (lr=0.005) with weight decay
- Cross-entropy loss
- 100 epochs

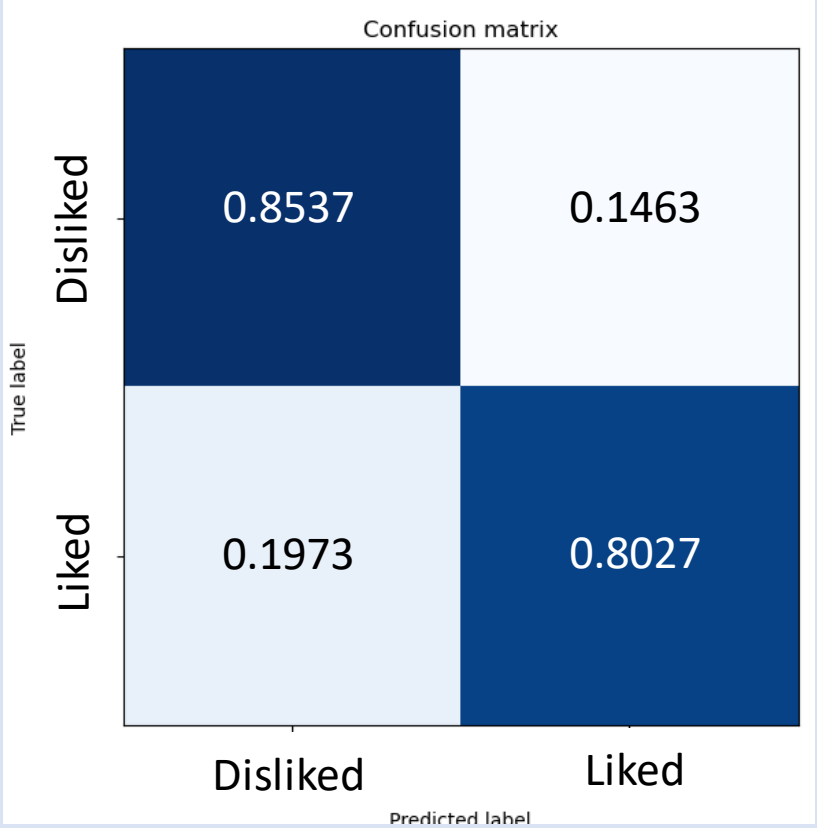
Results – Simple model

Word2Vec



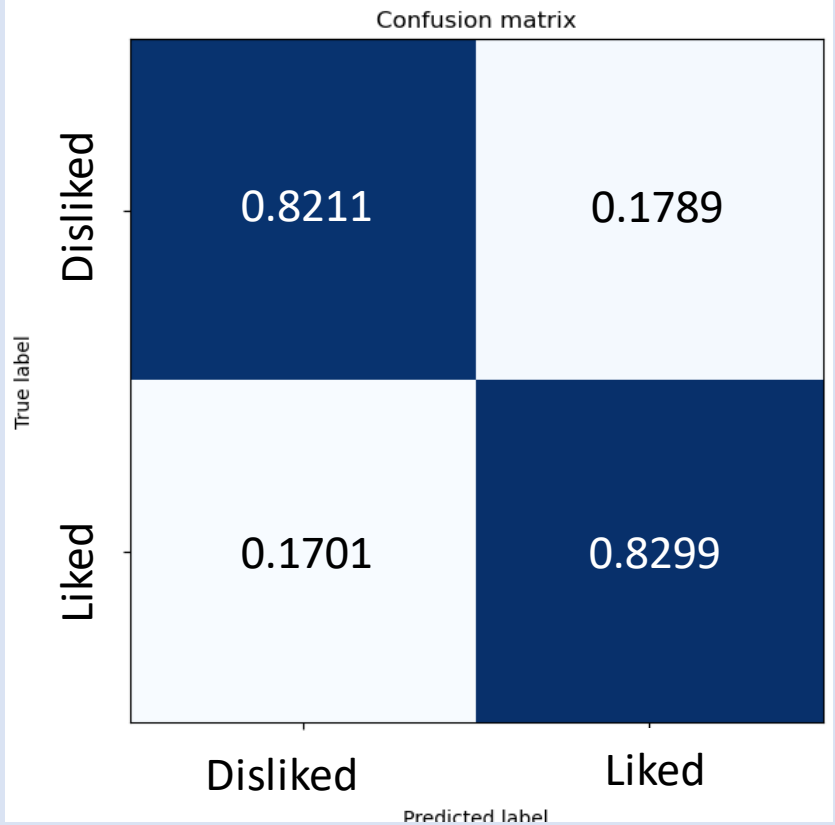
Accuracy: 83.3 %

GloVe



Accuracy: 82.5 %

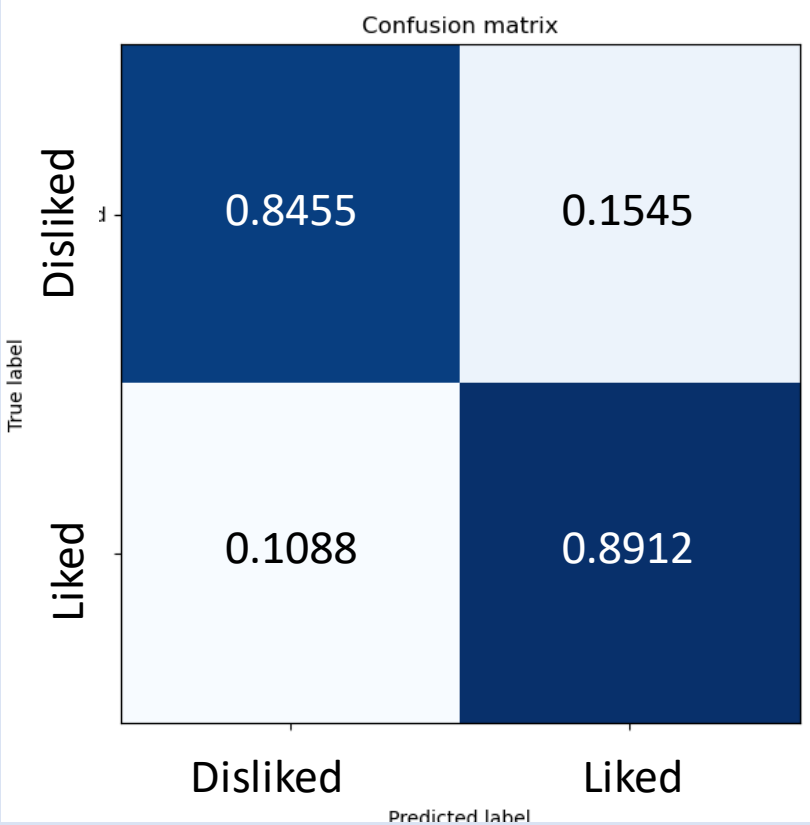
FastText



Accuracy: 82.6%

Results – Attention model

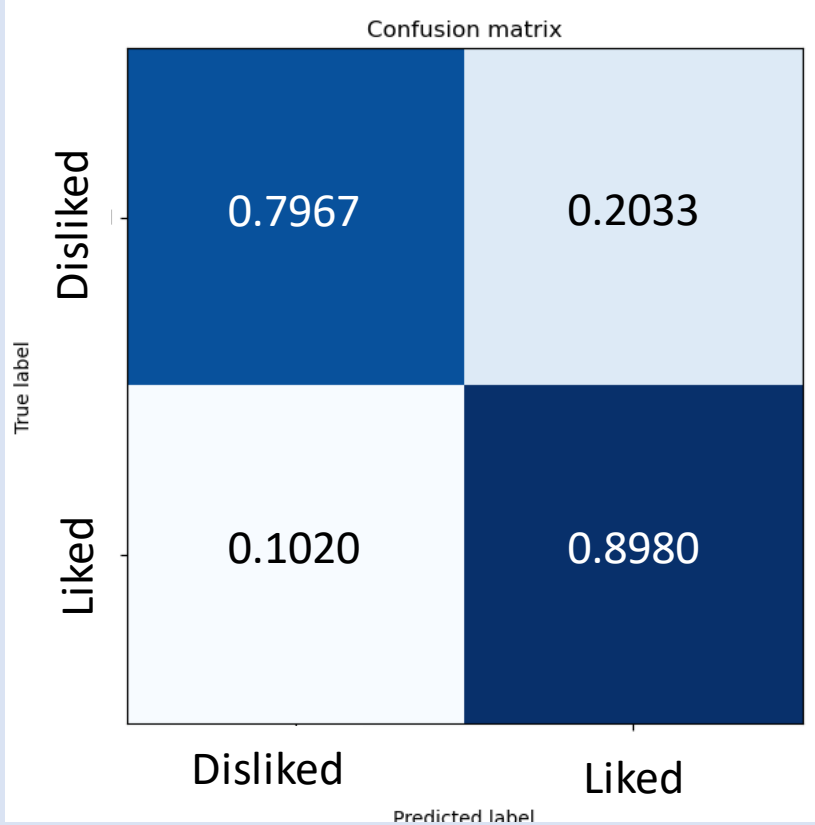
Word2Vec



Accuracy: 87.05%

Before: 83.32 %

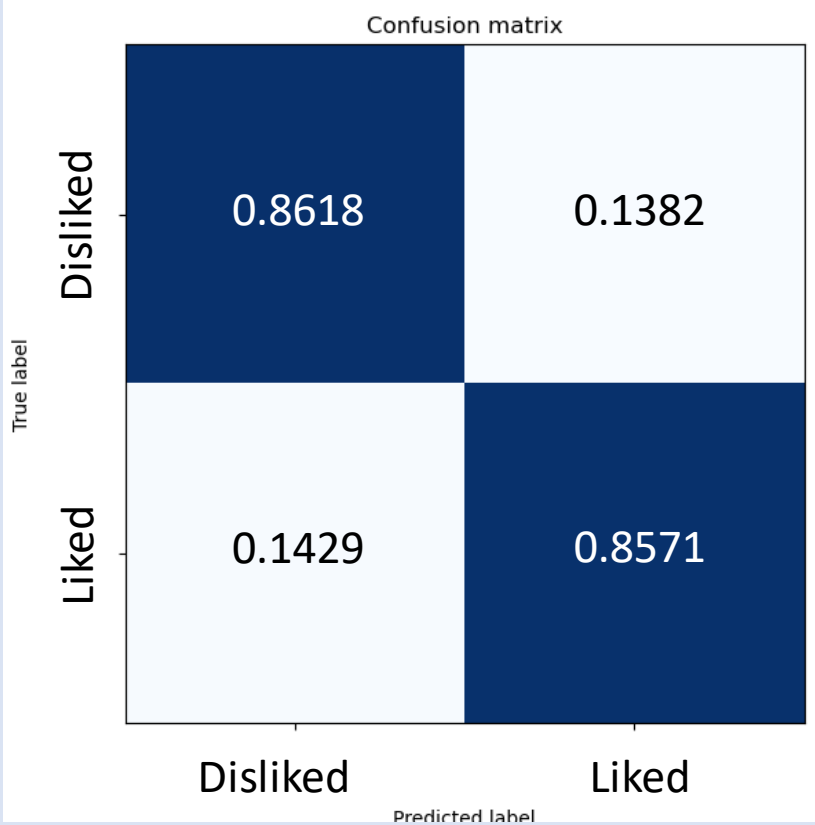
GloVe



Accuracy: 85.19%

Before: 82.5 %

FastText



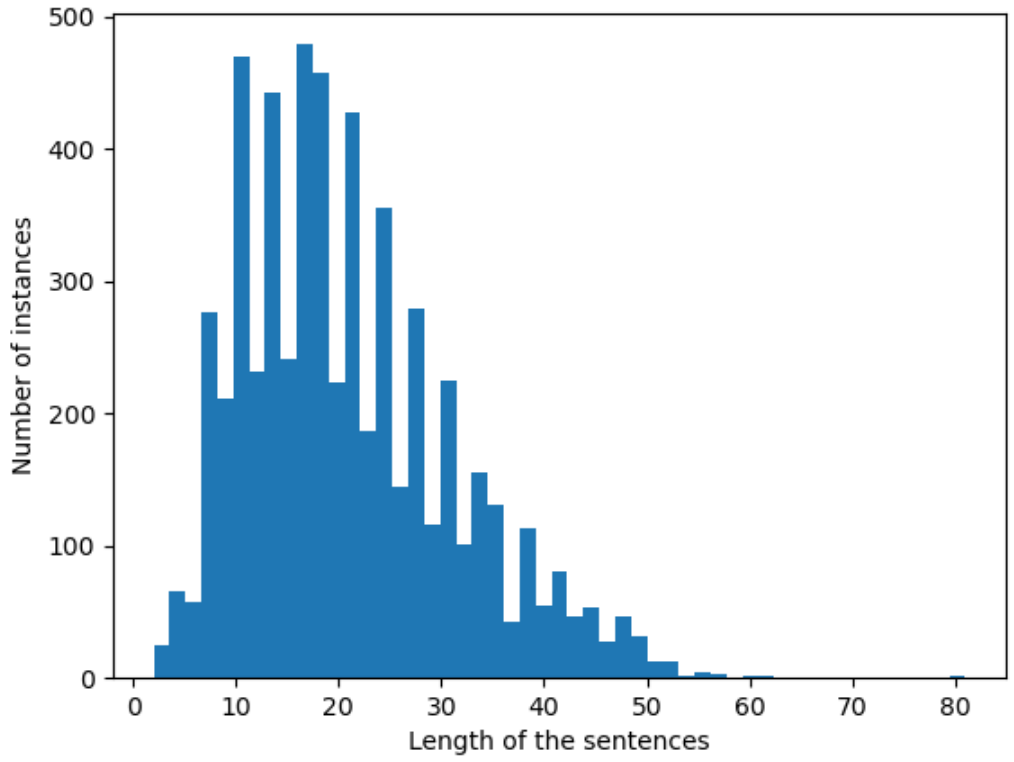
Accuracy: 85.93%

Before: 82.6 %

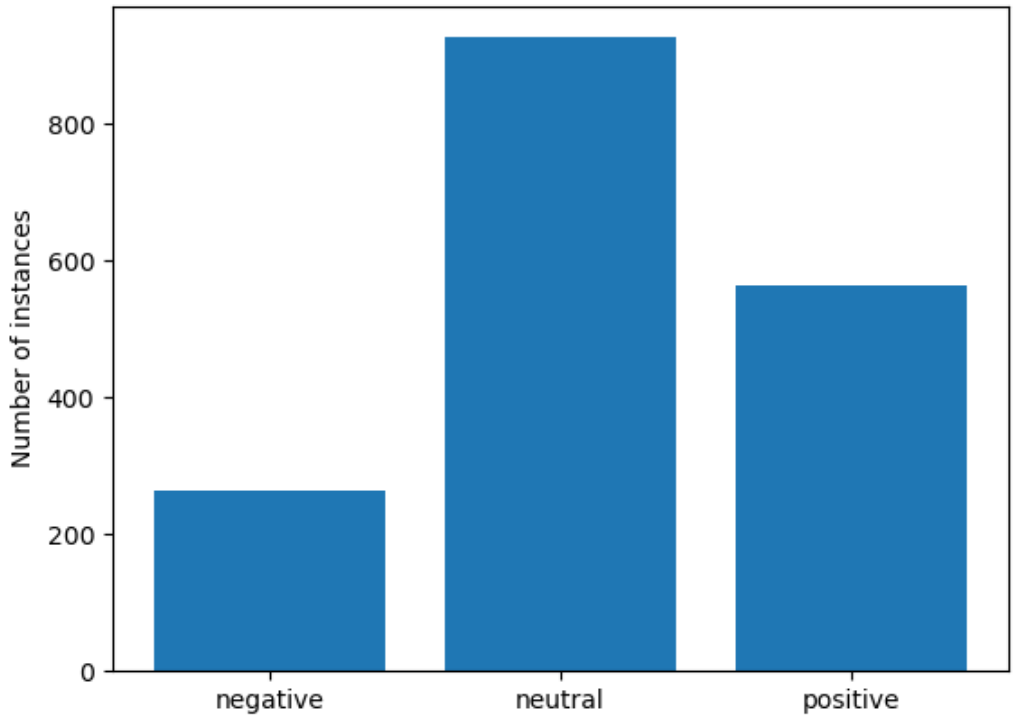
Testing with more classes

Financial data statements dataset

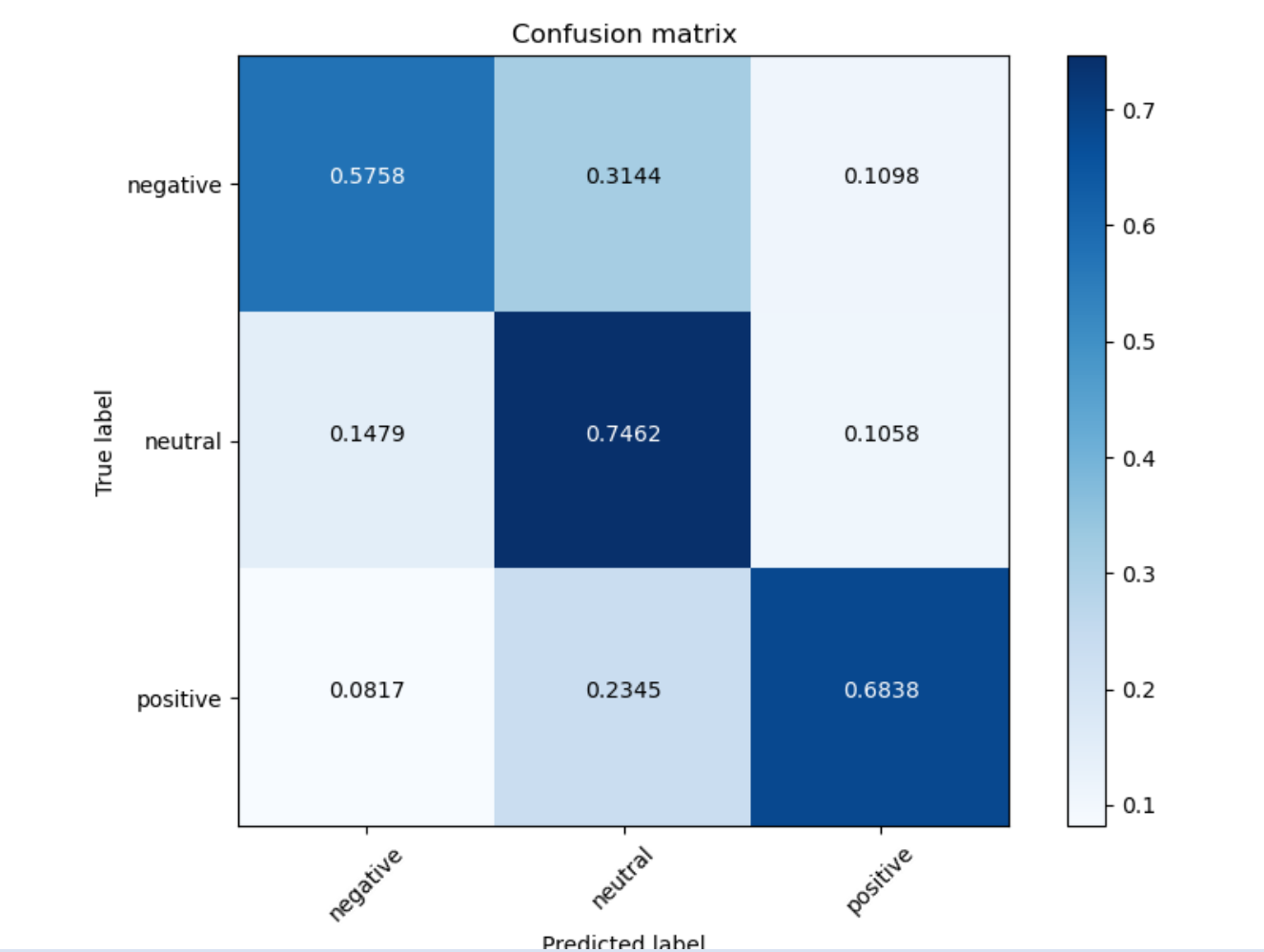
Histogram of the length of the sentences



Data imbalance



Results – Simple model with FastText

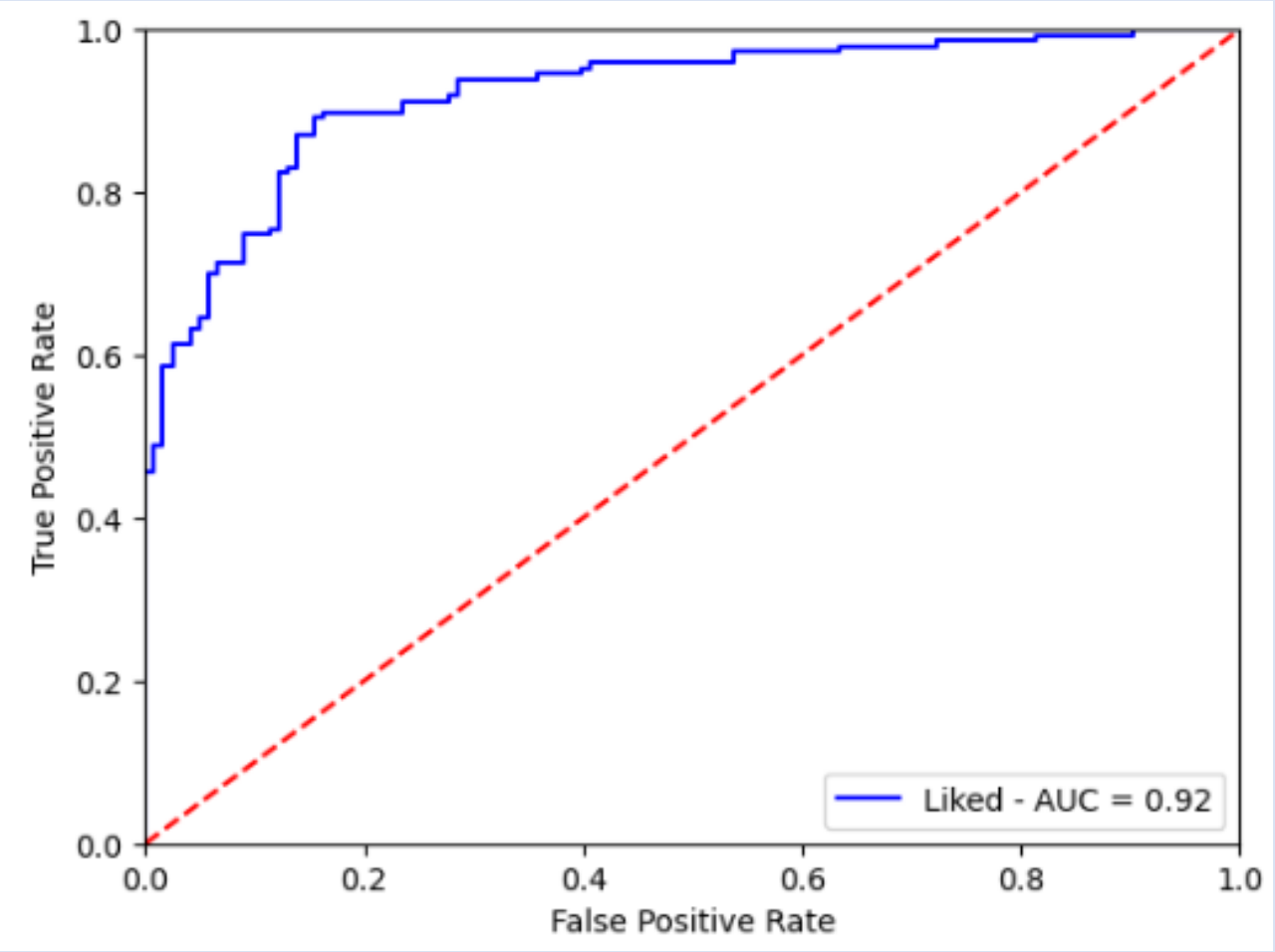


Accuracy: 70.05%

References & Links

- Our github: https://github.com/Jepi1202/Web-and-Text-Analytics
- Attention is all you need: https://arxiv.org/pdf/1706.03762.pdf
- Datasets:
  - Restaurant Reviews https://www.kaggle.com/datasets/hj5992/restaurantreviews
  - Financial Sentiment Analysis https://www.kaggle.com/datasets/sbhatti/financial-sentiment-analysis

W2v ROC CURVE



GLOVE ROC CURVE