Project 2 - Emotion

Henrik Daniel Christensen[hench13@student.sdu.dk]
Frode Engtoft Johansen[fjoha21@student.sdu.dk]

DM873: Deep Learning
University of Southern Denmark, SDU
Department of Mathematics and Computer Science

1 Introduction

The objective of this project is to develop a deep learning model capable of classifying the sentiment of a given text. The model will be trained on Emotion Dataset, a dataset of 16k tweets labeled with one of 6 emotions: sadness (0), joy (1), love (2), anger (3), fear (4) and surprise (5).

2 Label Distribution

We started by analyzing the distribution of the labels in the dataset. The distribution of the labels is shown in Figure 1.

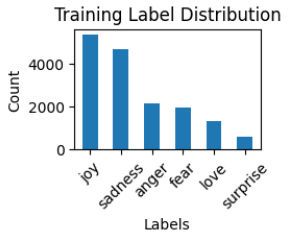

Fig. 1. Distribution of the labels in the dataset.

As can be seen, the dataset is very unbalanced, with the majority of the tweets being labeled as joy or sadness. This could potentially lead to the model being biased towards these labels.

3 Tokenizing and Vocabulary

Next, we tokenize the text data for that we created a custom RegexpTokenizer which tokenizes the text by only allowing words, numbers and some special characters. Even though the dataset is rather cleaned. We choose to keep exclamation marks and question marks as they potentially could add value for sentiment analysis.

Next, we check the top 10 most common words in the dataset. Many stopwords are present in the dataset also some words that are not relevant for sentiment analysis. Also, many words have similar meanings, e.g. 'feel' and 'feeling', therefore we also stem the words.

To see which words are most common in the dataset, we plot a WordCloud, which can be seen in Figure 2.

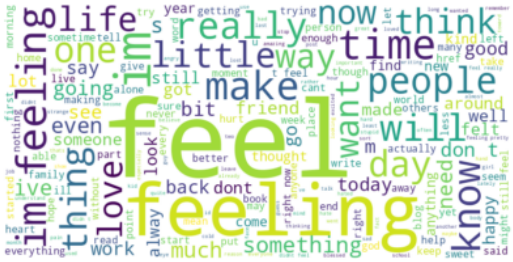

Fig. 2. WordCloud of the vocabulary.

We see a clear tendency to words accoiated with...

To be sure also all the text was converted to lowercase. We then create a vocabulary of all the unique tokens in the dataset. The vocabulary is then used to convert the text data into sequences of integers. The vocabulary for the training dataset ended up being 15,212 words.

Hereafter, we need to determine the length of the sequences. We do this by plotting the distribution of the lengths of the sequences. The distribution given as a boxplot can be seen in Figure 3.

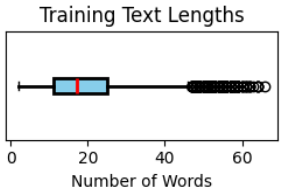

Fig. 3. Distribution of the lengths of the sequences.

From the boxplot, we see that the majority of the sequences have a length of around 20 words. We choose to set the maximum sequence length to 23 words. This means that all sequences longer than 23 words are truncated, and all sequences shorter than 23 words are padded with a special token, '<PAD>', to make them all the same length.

4 Model 1

5 Model 2

6 Analysis and Final Prediction

7 Pretrained Model (RoBERTa)

8 Conclusion