Analysis

4.1 Analysis of the Dataset

An analysis of the dataset given on Moodle will be provided in the following parts

4.1.1 Description

It is predominant to commence the analysis of the dataset with a brief description of the GoEmotions set itself. Emotions are a key aspect of social interactions, having a profound impact on the manner people behave and construct relationships. Due to the previously-mentioned significance of sentiments, it has been a long-term goal for machines to understand various emotions embedded in phrases. Consecutively, the latter led to a creation of a vast “GoEmotions: A Dataset of Fine-Grained Emotions: dataset, which consists of 58 thousand of Reddit posts extracted from popular English subreddits.

4.1.2 Sentiments

As the dataset was briefly introduced, it is time to dive deep into the categories of emotions it encapsulates as well as their distributions. That is, there are four major categories of sentiments: neutral, negative, positive, and ambiguous (entries, 11% of the dataset). Please refer to the table below in order to observe the exact distribution of categories of sentiments presented in the dataset.

Emotion Category	Number of Entries	Percentage
Positive	58968	34%
Neutral	55298	32%
Negative	38545	22%
Ambiguous	19009	11%
TOTAL	171820	100%

It is clear from the table that the category of ambiguous emotions is underrepresented, which will affect the choice of the score function later on.

4.1.3 Positive Emotions

Now, it is also significant to consider the emotions themselves that are represented in GoEmotions dataset, which are the following: admiration, amusement, approval, caring, desire, excitement, gratitude, joy, love, optimism, pride, relief, anger, annoyance, disappointment, disgust, embarrassment, fear, grief, nervousness, remorse, sadness, confusion, curiosity, realization, surprise. As there are four categories of emotions, they will be displayed in a table manner one by one, commencing with the category of positive emotions. It is clear that the positive emotions are not equally represented, which will affect the choice of the score function later on.

Emotion	Number of Entries	Percentage
Approval	11259	19.1%
Admiration	10531	17.86%
Gratitude	7075	12%
Amusement	6130	10.4%
Love	4957	8.41%
Optimism	4519	7.66%
Joy	4329	7.34%
Caring	3523	5.97%
Excitement	3020	5.12%
Desire	2147	3.64%
Relief	788	1.34%
Pride	690	1.17%

4.1.4 Negative Emotions

Now, let us consider the negative emotions, which are the following: anger, annoyance, disappointment, disapproval, disgust, embarrassment, fear, grief, nervousness, remorse, and sadness. Please refer to the table below in order to observe the exact distribution of negative emotions presented in the dataset.

Emotion	Number of Entries	Percentage
Annoyance	8342	21.64%
Disapproval	7686	19.94%
Anger	5202	13.05%
Disappointment	4706	12.21%
Sadness	3827	9.93%
Disgust	2914	7.56%
Fear	1778	4.61%
Remorse	1510	3.92%
Embarrassment	1433	3.72%
Nervousness	796	2.07%
Grief	351	0.91%

It is clear that the negative emotions are not equally represented, which will affect the choice of the score function later on.

4.1.5 Ambiguous Emotions

Now, let us consider ambiguous emotions, which are the following: confusion, curiosity, realization, and surprise. Please refer to the table below in order to observe the exact distribution of ambiguous emotions presented in the dataset.

Emotion	Number of Entries	Percentage
Curiosity	5885	30.96%
Confusion	4938	25.98%
Realization	4714	24.8%
Surprise	3472	18.27%

4.1.6 Analysis

Now, as the dataset was described and analyzed in great detail, it is possible to argue for the best metric that will be used to assess the score of the models. First and foremost, it is significant to mention that the choice will vary between recall and precision, since accuracy is taken into account only if the classes are equally-represented (which is obviously not the case according to the tables presented previously). Consecutively, from my perspective, I believe that the metric that will be best suited to assess the score heavily depends on the task that the model will have to perform. For instance, if we care how many times the model is correct when predicting that the phrase has a positive emotion, we would go for precision. On the other hand, if we care how many times the model predicted correctly out of all data of positive emotions, we would go for recall. In my opinion, the metric that would suit better in general for tasks of text classification is recall. This is due to the fact that it is

of supreme importance for the model to indeed understand what emotion is expressed in the phrase, we want the classifier to be “sensitive” and “complete”, rather than having a high chance that the model correctly predicts a specific emotion that it rarely chooses. Therefore, it is recall that was selected for the purposes of assessment.

4.2 Analysis of the Results

An analysis of the results of all the models for both classification tasks will be provided in this part.

4.2.1 Words as Features

In order to provide an extensive analysis of the performances of all the models trained, it is first predominant to provide the following tables with all scores of all models.

Words As Features				
	Sentiments		**Emotions**	
Base-MNB	*Accuracy*	54%	*Accuracy*	39%
	Macro Average F1	50%	*Macro Average F1*	17%
	Weighted Average F1	54%	*Weighted Average F1*	31%
	Recall	54%	*Recall*	39%
Top-MNB	*Accuracy*	54%	*Accuracy*	40%
	Macro Average F1	50%	*Macro Average F1*	25%
	Weighted Average F1	54%	*Weighted Average F1*	35%
	Recall	54%	*Recall*	39%
Base-DT	*Accuracy*	54%	*Accuracy*	36%
	Macro Average F1	53%	*Macro Average F1*	28%
	Weighted Average F1	55%	*Weighted Average F1*	36%
	Recall	55%	*Recall*	36%
Top-DT	*Accuracy*	44%	*Accuracy*	40%
	Macro Average F1	28%	*Macro Average F1*	14%
	Weighted Average F1	36%	*Weighted Average F1*	28%
	Recall	44%	*Recall*	40%
Base-MLP	*Accuracy*	57%	*Accuracy*	44%
	Macro Average F1	51%	*Macro Average F1*	24%
	Weighted Average F1	56%	*Weighted Average F1*	37%
	Recall	57%	*Recall*	44%
Top-MLP	*Accuracy*	56%	*Accuracy*	44%
	Macro Average F1	50%	*Macro Average F1*	23%
	Weighted Average F1	55%	*Weighted Average F1*	36%
	Recall	57%	*Recall*	44%

First and foremost, it is predominant to underline which metric has to be put first in order to draw a conclusion on the best performing model out of all trained. As it was underlined in the previous part, it is of more importance to consider recall, since the previously-mentioned metric is more descriptive with regards to how complete our classifiers are, and whether they are really able to distinguish one emotion from another. Therefore, taking into account recall metric, it is possible to draw a conclusion that the Multi-Layered Perceptron Classifier performed the best, having a decent result of 57% of recall for sentiments and 44% for emotions classifications. Consecutively, the Decision Tree and Multinomial Naïve Bayes classifiers achieved smaller recall, contributing to 55% and 54% for sentiments and 36% and 39% for emotions for base models.

Apart from that, it is essential to mention that the base models and the top models for the MLP classifier achieved almost the same results, albeit the fact that the top models were supposed to use the best set of hyperparameters found using GridSearchCV. However, the top MNB model showed an improvement in comparison with the base model in absolutely all metrics for emotions, while achieving the same result for sentiments. I would assume that this is due to the fact that the sentiments classification achieved its limit of performance, while the emotion classification is a harder task with more categories, therefore the alpha float found using GridSearchCV increased the performance for the model. Lastly, an interesting trend is found for the DT, since the metrics for the top DT classifier decreased a maximum of 8%. The latter may be explained by the hyperparameter of max depth, which was ought to be found between seven and thirteen, which is relatively small for the number of categories presented in the dataset.

4.2.2 Words as Features (Stop Words Removed)

Words As Features [Stop Words Removed]		
	Sentiments *(Recall)*	**Emotions** *(Recall)*
Base-MNB	54%	39%
Top-MNB	54%	39%
Base-DT	55%	36%
Top-DT	44%	40%
Base-MLP	57%	44%
Top-MLP	57%	44%

Taking into account recall metric (it was the only metric recorded for the purposes of training the models without stop words), it is possible to draw a conclusion that the Multi-Layered Perceptron Classifier performed the best, having a decent result of 57% of recall for sentiments and 44% for emotions classifications. Consecutively, the Decision Tree and Multinomial Naïve Bayes classifiers achieved smaller recall, contributing to 55% and 54% for sentiments and 36% and 39% for emotions for base models. Therefore, summarizing the previously-mentioned, it is possible to draw a conclusion that the values of recall did not change at all for absolutely all the classifiers.

I assume one possible reason for the aforementioned trend: stop words do not affect the emotion classification, since they do not embed in themselves any emotion whatsoever. That is, one might say “I hate my life, what a waste” with “a” stop word, or one might also pronounce “I hate my life, what waste”. The emotion encapsulated within the phrase is not changed at all, since it is not the stop words that “enrich” the phrase with the emotional

meaning. Apart from that, it is essential to mention that base models and top models for all classifiers achieved exactly the same results as the ones trained with stop words.

4.2.3 Embeddings as Features (Word2Vec)

Embeddings As Features [Word2Vec]				
	Sentiments		**Emotions**	
Base-MLP	*Accuracy*	52%	*Accuracy*	38%
	Macro Average F1	43%	*Macro Average F1*	11%
	Weighted Average F1	50%	*Weighted Average F1*	26%
	Recall	52%	*Recall*	38%
Top-MLP	*Accuracy*	52%	*Accuracy*	38%
	Macro Average F1	42%	*Macro Average F1*	11%
	Weighted Average F1	50%	*Weighted Average F1*	26%
	Recall	52%	*Recall*	38%

Again, considering the recall metric, the base and the top models of the Multi-Layered Perceptron Classifier performed exactly the same, without any change at all in the metric. Furthermore, even if we consider all other metrics, they vary maximum by 1%, meaning that the hyperparameters did not affect the result at all.

Apart from that, as we analyze a new approach to text classification, it is possible to compare results from embeddings with the results from word frequencies. After a detailed observation, it is possible to conclude that absolutely all metrics decreased. That is, if we consider solely recall as our main metric of assessment, sentiments decreased from 57% to 52%, and emotions from 44% to 38%. The latter-mentioned trend may be explained by the extremely low hit rate of the Word2Vec model, which is 37% for the training set and 42% for the test set. Therefore, it is obvious that the classification results will be smaller due to the fact that the pre-trained model does not “understand” more than half the words that are utilized in the Reddit posts.

4.2.4 Embeddings as Features (Glove-Twitter-25)

Embeddings As Features [Glove-Twitter-25]		
	Sentiments *(Recall)*	**Emotions** *(Recall)*
Base-MLP	45%	35%

As we analyze a new approach to text classification, but now with a different model called Glove-Twitter-25, it is possible to compare results from embeddings with the results from word frequencies. After a detailed observation, it is possible to conclude that absolutely all metrics decreased. That is, if we consider solely recall as our main metric of assessment, sentiments decreased from 57% to 45%, and emotions from 44% to 35%. The latter-mentioned trend may be explained by the extremely low hit rate of the Glove-Twitter-25 model, which is 39% for the training set and 44% for the test set. Therefore, it is obvious that the classification results will be smaller due to the fact that the pre-trained model does not "understand" more than half the words that are utilized in the Reddit posts.

Apart from that, it is still possible to make a conclusion that the hit rates of Glove-Twitter-25 are higher by 2% in comparison with the ones of Word2Vec. However, albeit the previously-mentioned fact, the results of the recall are still not as good as with Word2Vec, since they are smaller by 7% for sentiments classification, and by 3% for emotions classification. One possible explanation that I will provide is the size of the array of embeddings, where the Word2Vec arrays are of size 300 and Glove-Twitter-25 arrays are of size 25. The previously-mentioned aspect might have affected the performance of the classifications since the model had fewer values of embeddings to work with.

4.2.5 Embeddings as Features (Glove-Wiki-Gigaword-100)

Embeddings As Features [Glove-Wiki-Gigaword-100]		
	Sentiments *(Recall)*	**Emotions** *(Recall)*
Base-MLP	48%	36%

As we analyze a new approach to text classification, but now with a different model called Glove-Wiki-Gigaword-100, it is possible to compare results from embeddings with the results from word frequencies. After a detailed observation, it is possible to conclude that absolutely all metrics decreased. That is, if we consider solely recall as our main metric of assessment, sentiments decreased from 57% to 48%, and emotions from 44% to 36%. The latter-mentioned trend may be explained by the extremely low hit rate of the Glove-Wiki-Gigaword-100 model, which is 39% for the training set and 44% for the test set. Therefore, it is obvious that the classification results will be smaller due to the fact that the pre-trained model does not "understand" more than half the words that are utilized in the Reddit posts. It is predominant to mention that it is still higher than the hit rates of Word2Vec, but almost exactly the same as the hit rates of Glove-Twitter-25.

Apart from that, the results of the recall are still not as good as with Word2Vec, since they are smaller by 4% for sentiments classification, and by 2% for emotions classification. However, they are higher than the results of Glove-Twitter-25, since the recall metric increased by 3% for sentiments classification and by 1% for emotions classification. One possible explanation that I will provide is the size of the array of embeddings, where the Glove-Twitter-25 arrays are of size 25 and Glove-Wiki-Gigaword-100 arrays are of size

100. The previously-mentioned aspect might have affected the performance of the classifications since the model had more values of embeddings to work with.

4.3 Description of Responsibilities

The project was done individually by Artem Chernigel, Student ID 40115241.