Final Assignment Report

Csenge Szabó, student ID: csz202, 2803781

October 28, 2023

Introduction

Analysis of the Test Data

The test data is derived from the annotated Eliza conversations, which is composed of 2129 test instances and the same number of test labels. Eliza's neutral utterances were excluded for this assignment, and only the speakers' turns will be analyzed. It is important to note that there was no code book defined for the annotation of the test instances; the annotators were the speakers themselves, and each instance was labeled with a single emotion by one annotator only. The units of the annotations were the individual turns of utterances, and the annotation procedure was not followed by adjudication.

The tag set is composed of seven emotion labels. The emotions in our test data are not equally distributed across the utterances. Most of the utterances are labeled as "neutral" (39%), while other emotions like "anger" (12%), "sadness" (14%), "joy" (16%), "fear" (6%), "surprise" (5%), and "disgust" (5%) are less frequent, which results in an imbalanced distribution. See Appendix 1.

1 Training a classifier on MELD and Tweets data (SVM)

1.1 Stop words

The instances of the data from MELD and WASSA Tweets were combined using the concat() function from the pandas library. The nature of the task is sentiment analysis, therefore all stop words were included when training the classifier, since each word can significantly modify the meaning of a sentence. For instance, "don't" in "I don't really like talking about my feelings" (test instance 114) changes the sentiment of a sentence. Stop words are also useful for maintaining sentence coherence and meaning, which explains the improved results with their inclusion in the evaluation stage. First I experimented with including stop words, then removed the words with negation from the stop word list, finally, not using a stop word list at all showed the most optimal results.

1.2 Word frequency threshold

The threshold(3) indicates the minimum number of occurrences a word must have in the combined training data from the two data sets to be included in the feature space. After some experimentation with higher and lower thresholds, this value was chosen in order to balance the ratio of informative words and reduce noise. Words that appear less than three times in the data set probably do not provide sufficient information for the classifier to make accurate predictions on the test set. Setting a higher threshold, however, may come with the danger of excluding potentially valuable information. This threshold yielded a good balance between precision and recall, leading to better classification performance during evaluation.

1.3 Hypotheses about the performance

Considering that the distribution of emotions is fairly similar in the training data and the test data, the following could be expected: (1) The classifier should have reasonable precision and recall score for most emotions, especially for "neutral" and "joy" considering their diversity and proportion in the training data. (2) Emotions with limited training instances, such as "disgust", might have low recall and precision scores due to the sparsity of data. See Appendix 2.

2 Report on the GO-Ekman Classification Results

2.1 Choice of threshold (0.03)

In order to choose an most optimal threshold for mapping GO-emotion labels to Ekman labels, I experimented with higher and lower ranges of thresholds. Higher ranges of 0.1 – 0.25 resulted in accuracy lower than 57% and in some cases no emotion was mapped ('None') to the utterances. Applying lower ranges 0.01 – 0.04 for the threshold resulted in accuracy higher than 62%, while keeping precision and recall generally balanced and relatively high for most emotions. It was noticeable that further decreasing the threshold resulted in higher recall at the cost of lower precision, since it enabled the classifier to assign an emotion label, but it also increased the chances of assigning the incorrect label. I attempted setting individual thresholds for emotions, which did not significantly change the evaluation outcomes. Finally, I chose the threshold of 0.03 for optimizing overall performance of the emotion classification model and minimizing the gap between precision and recall. The model's overall accuracy using this threshold is about 63.55%, which measures the proportion of the test data that was correctly classified into the Ekman emotion labels. This threshold supports a well-balanced trade-off between precision and recall, ensuring a relatively high number of true positives while minimizing false positives and false negatives.

2.2 Classification Report Analysis

Noticeably, "neutral" and "joy" have relatively high recall scores, which implies the model can capture these emotions well (Table 1). The emotion "disgust" however, shows a lower F1-score than other emotions due to its low recall, which means the model may benefit from further refinement for this particular emotion. This is the least frequent label in the test set with only 107 instances, the emotion "disgust" could also be highly context dependent and implicit, which might explain the low performance for this emotion.

2.3 Confusion Matrix Analysis

The confusion matrix gives an insight into which emotion labels were true positives (diagonal elements), false positives and false negatives. The model can effectively distinguish between different Ekman emotions with some instances of misclassification. The model's strongest capability is to classify the "neutral" emotion, which has the highest number of true positives. On the other hand, "neutral" is overly dominant, as there are several false positives for this class, for example "anger" was mislabelled 105 times as "neutral". Other noticeable confusion patterns are "anger" and "disgust", as well as "anger" and "sadness". This could be explained by the fact that these emotions might have overlapping expressions, which leads to confusion. See Appendix 3.

Emotion	Precision	Recall	F1-Score
Anger	0.6061	0.4598	0.5229
Disgust	0.6957	0.2991	0.4183
Fear	0.8454	0.5655	0.6777
Joy	0.5910	0.7289	0.6527
Neutral	0.6188	0.7599	0.6821
Sadness	0.7971	0.5238	0.6322
Surprise	0.5259	0.5680	0.5462
Accuracy			0.6355
Macro avg	0.6686	0.5578	0.5903
Weighted avg	0.6530	0.6355	0.6289

Table 1: GO Ekman Metrics with Threshold 0.03

Emotion	Precision	Recall	F1-Score
Anger	0.4332	0.4100	0.4213
Disgust	0.3800	0.1776	0.2420
Fear	0.5565	0.4759	0.5130
Joy	0.5987	0.5569	0.5770
Neutral	0.5717	0.7611	0.6529
Sadness	0.6055	0.4190	0.4953
Surprise	0.3387	0.1680	0.2246
Accuracy			0.5510
Macro avg	0.4978	0.4241	0.4466
Weighted avg	0.5397	0.5510	0.5337

Table 2: SVM classifier results with Threshold 3

3 Comparison of the SVM Classifier and GO-Ekman Classification

3.1 Differences in performance

Both models show reasonable performance in recognizing emotions, with F1-scores on average above 40%, the strongest capacity of both is identifying the "neutral" test instances. As expected in the hypotheses section, the SVM classifier has the lowest recall for "surprise" and "disgust", whereas "joy" and "neutral" have the highest F1-scores (Table 2). When comparing the two models, the BERT-based GO classifier performs better across the majority of the metrics. It demonstrates higher precision, especially for "disgust" and "fear", alongside higher recall scores, especially for "surprise" (Table 3). The F1-score is higher for each emotion label, the accuracy, macro and micro averages also indicate that the BERT-based classifier has superior performance.

The primary reason for these differences is the training method of the classifiers: the fine-tuned GO classifier is using contextual embeddings, which capture the nuanced semantic differences in the context. It was trained on 58k carefully selected and annotated English Reddit comments. On the other hand, the SVM classifier was created using BoW and TF-IDF feature extraction methods, thus it performs a lot weaker than the fine-tuned tranformer-based model since it cannot capture contextual meaning. The SVM classifier was trained on a lot fewer - only 13 602 instances -, combining utterances from the TV show *Friends* and WASSA Tweets.

Emotion	Precision Difference	Recall Difference	F1-Score Difference
Anger	0.1729	0.0498	0.1016
Disgust	0.3157	0.1215	0.1763
Fear	0.2889	0.0897	0.1647
Joy	-0.0077	0.1720	0.0757
Neutral	0.0471	-0.0012	0.0292
Sadness	0.1916	0.1048	0.1369
Surprise	0.1872	0.4000	0.3216
Accuracy			0.0845
Macro avg	0.1708	0.1338	0.1437
Weighted avg	0.1132	0.0845	0.0953

Table 3: Comparison of the performance of the SVM and GO-classification

3.2 Error analysis using test instances

Test Instance	GOLD Label	GO Label	SVM Label
I'm so disgusted by your behavior.	disgust	disgust	neutral
You're weird and it makes me uncomfortable.	disgust	disgust	joy
My new place is also very dirty and dark.	disgust	disgust	sadness
(...) it's disgusting that you are trying to hide.	disgust	disgust	anger
I'm surprised with the gift	surprise	surprise	neutral
That's astonishing!	surprise	surprise	joy
Surprisingly I feel like you're my friend	surprise	surprise	neutral
You're not scared you are alone?	surprise	surprise	fear
I feel disgusting	disgust	sadness	neutral
So please do not make me angry	disgust	anger	anger
How can people talk to this thing for 100 rounds?	disgust	surprise	neutral
How dare they call themselves a family?!	disgust	anger	surprise
I didn't expect you to say that at all!	surprise	neutral	anger
You bewildered me	surprise	neutral	anger
I am amazed by everything happened in my life.	surprise	joy	joy
I didn't expect you to ask such a challenging question.	surprise	neutral	neutral
I feel like we're both not showing our fun side to each other.	sadness	joy	joy
I can't be happy.	sadness	joy	joy
I just wish I was stronger	sadness	joy	joy

Table 4: Test Instances for Error Analysis

When selecting errors, I decided to focus on the emotions "disgust" and "surprise" since the precision, recall and F1-scores were the lowest for these emotions in case of both models. The first 8 selected test instances represent cases where the GO-model correctly assigned the label "disgust/surprise", whereas the SVM classifier misclassified the instance. These instances can highlight potential directions for improving our trained classifier. The second 8 selected test instances represent cases where neither of the classifiers managed to correctly label the test instance with the label "disgust/surprise". It was unexpected that both models would quite frequently confuse the complementary emotions "sadness" and "joy", the last 3 instances of Table 4 illustrate such test cases.

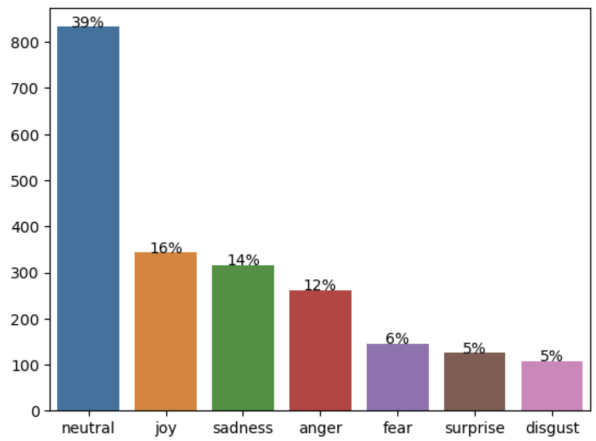
The examples demonstrate that the classification of emotions can be highly context-dependent and implicit. Both classifiers struggle with understanding nuances in sentiment differences, as shown by the misclassifications in the selected certain instances. The SVM classifier appears to be overfitting to "neutral" and "joy" due to an imbalance in the training data. This overfitting results in misclassifications, especially when other emotions are also expressed in the text. The SVM classifier is especially weak in identifying "disgust", which could be partially explained by the scarcity of this emotion in the training data (1%). The main reason for confusing contradicting emotions (sadness and joy) is that the models sometimes do not pick up on negation within the sentence, which changes the sentence meaning to its opposite. The gold labels themselves may also be impacted by subjectivity; some instances contain typos, which can impact the performance of both classifiers. Emotions are complex, and annotators might interpret them differently depending on the context.

4 Improvements of the SVM Classifier

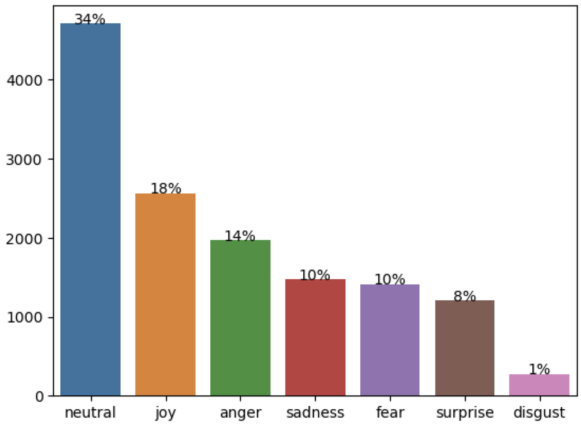
The error analysis highlights that lemmatizing the training and test utterances could potentially improve the SVM classifier in capturing key phrases, thereby resulting in higher accuracy. This step can help reduce the feature space and enabling the model to generalize across different word forms. I experimented with thorough text preprocessing, including removing special characters, emojis and profile tags from the tweets before training the classifier, surprisingly, this has not improved the performance. The imbalance of the training set should be addressed by reducing the number of neutral instances and increasing instances for underrepresented emotions like "disgust", this could help the model make more accurate predictions across all emotions and regulate the degree of overfitting. Collecting more diversified data could enhance the classifier's performance. Regarding feature engineering, experimenting with different techniques, such as using word embeddings instead of BoW text representations could better capture context and semantic relationships.

Appendices

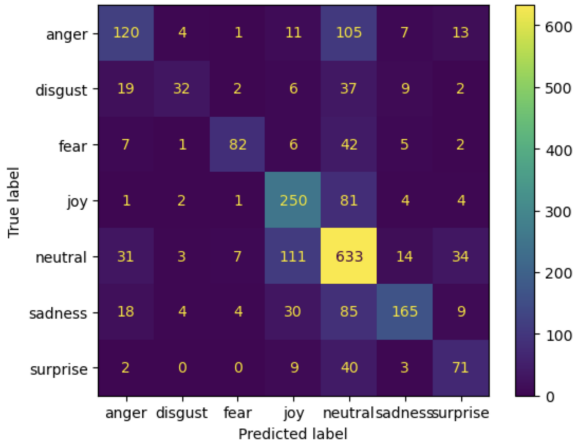
Appendix 1. Distribution of ELIZA test data



Appendix 2. Distribution of MELD and Tweets training data



Appendix 3. Confusion Matrix for GO-Ekman classification results



Appendix 4. Confusion Matrix for SVM classifier results