Data Science and Economics
Universit‘a degli Studi di Milano

How do you feel, my dear?

Student: Vojimir Ranitovic 963780
Email: vojimir.ranitovic@studenti.unimi.it
Module: Text Mining and Sentiment Analysis

Introduction

Emotion detection, as part of sentiment analysis, involves automatically identifying and classifying the emotional content of text into discrete categories, such as anger, fear, joy, enthusiasm, worry, love, sadness, etc. Besides the discrete representation of emotions, there is also representation as a continuous space such as Valence-Arousal-Dominance. Emotion detection can be used in various applications in many fields such as marketing, customer service, political analysis, education, mental health, and healthcare in general.

One of the main difficulties of emotion detection in the text is the intrinsic sophistication of human emotions. Emotions are often subtle and unclear, and their expression in a text can vary considerably based on the context and the individual. In addition, the same emotion can be expressed in different ways, making it difficult to specify clear and consistent criteria for classification. However, the ability to detect emotions in text with high accuracy has the potential to provide valuable insights of human behavior, communication, and latent feedback.

In this paper, the focus is on emotion detection in text using different machine-learning algorithms. Particularly, performance of different classification algorithms was explored, such as Naive Bayes, Support Vector Machines (SVM), Logistic Regression, and Linear SVM, on a labeled dataset of tweets. After finding the best algorithm, it was applied to movie dialogues to find emotions in texts, and also to find similar scenes based on emotions predicted by the best model.

Research question and methodology

The main aim of the project is to find emotions in movie scripts but also to describe interactions between characters and overall emotions throughout the movie. Besides this, the goal is also to find similar characters and scenes between movies.

Dataset and preprocessing

The dataset[1] used for training and testing supervised algorithms is WASAA-2017. It consists of 7102 tweets, all labeled with one of four categorical labels: fear, anger, joy, and sadness. The dataset was slightly unbalanced so minor classes were randomly oversampled and new instances were slightly changed with the function *replace_with_synonyms* in code. That function randomly changed words with their synonyms, so there could be more diversity in newly generated data. The dataset was split, before oversampling, into train and test sets, where 80% of tweets were in the train set.

After the split, the preprocessing part was the usual one for tweets. All mentions, links, digits, double-blank spaces, punctuation, and stopwords were removed. Also, words with only two or fewer letters were cleaned. In this case, hashtags remained because they

[1]Dataset can be downloaded at: http://saifmohammad.com/WebPages/EmotionIntensity-SharedTask.html

contain a lot of information about particular emotions as we can see in Figure 1. Emojis were also transformed into words since they could also contain useful details.

Figure 1: The most common hashtags in each class.

Below is an example of one tweet before and after preprocessing.

Original tweet:
"Goodfellas was authentic and kind of provided the anchor for the genre. Besides, Pesci, DeNiro, and Liotta were perfect three . But, popular impression is that Casino has way well ending !@DeNiro that acting was 🔥🔥😊 *. I want sequel #Casino!* 😭*"*

Preprocessed and lemmatized tweet: *"goodfellas authentic kind provide anchor genre pesci deniro liotta perfect popular impression casino way end acting fire fire smile face smile eye want sequel casino loudly cry face"*

After getting the best model, it was used on the new dataset Cornell Movie-Dialogs Corpus[2], that has 304,713 utterances and 83,097 conversations of 617 movies. From that corpus, only two movies were chosen, *Casino* and *Goodfellas*. In addition, to understand emotions throughout the whole movie, two extra datasets were downloaded. Those were subtitle files with timestamps, downloaded from the *podnapisi*[3] website.

Classification algorithms
In this project algorithms like SVM, Random forest, Logistic regression, and Naive Bayes were used. Two additional promising algorithms were the Passive aggressive (PA) and optimization SGD classifier with different losses. The Passive-Aggressive algorithm is a learning algorithm designed to make fast updates to the model as new train examples

[2]Dataset can be downloaded at: https://www.cs.cornell.edu/ cristian/Cornell_Movie-Dialogs_Corpus.html
[3]Subtitles can be downloaded at: www.podnapisi.net

become accessible while maintaining high accuracy. The idea behind the PA algorithm is to adjust the model parameters in the direction of the new training example, while still maintaining a margin between the positive and negative examples. The amount of adjustment is determined by a C variable in code, that allows the algorithm to tolerate some degree of misclassification. A lower value of C results in a more conservative model, which might help in reducing overfitting to the training data. On the other hand, a higher value of C can result in a model with high accuracy on the training data but may lead to overfitting if the model becomes too complex. SGDClassifier (Stochastic Gradient Descent) with hinge loss is often used to train linear Support Vector Machines (SVMs) for classification problems. On the other hand, SGD with log loss is usually used to train logistic regression models. The SGDClassifier works by iteratively updating the weights of a model using gradient descent on randomly selected subsets of the training data (mini-batches). This makes the algorithm well-suited for large datasets, as it can process the data in small chunks rather than loading the entire dataset into memory. Hyperparameters tuning was done for each of the mentioned algorithms, where the one with the best F1 score on both training and test sets was chosen, also 5-fold cross-validation was used for model assessment.

Results

In Table 1. are the results of each algorithm after tuning, where the SGD classifier with hinge loss was the best one and logistic regression was the worst one for this specific task.

Algorithm Score	F1 - train	F1 - test
SVC (SVM)	0.90	0.85
Passive Aggressive	0.91	0.86
Random Forest	0.88	0.84
SGD (hinge Loss)	**0.91**	**0.87**
Naive Bayes	0.87	0.82
Logistic Regression	0.82	0.79

Table 1: Performance comparison of different algorithms on the train and test sets.

Interestingly, both movies have a similar distribution of those four emotions, as it can be seen in Figure 2.

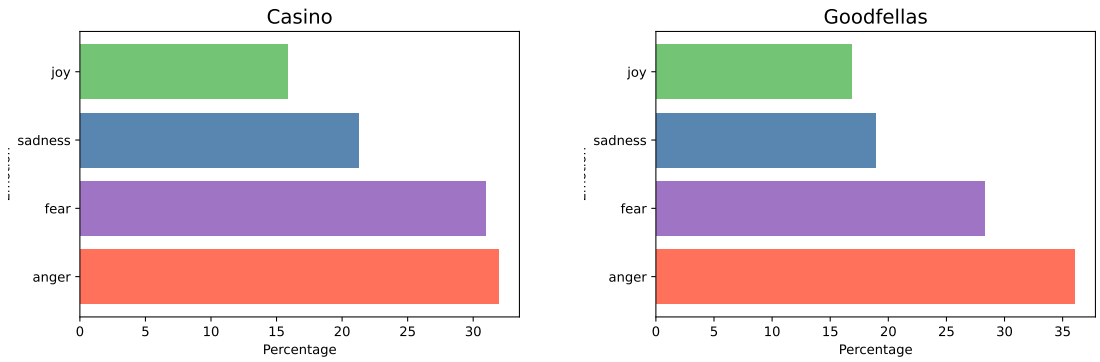

Figure 2: The percentage of emotions in both movies.

In Figure 3. we can see emotions between characters when they are having conversation. It is noticeable that emotions like fear and anger are dominant in both movies between all conversations. Casino's characters are on the left, and Goodfellas's are on the right graph. Figures showing emotions for each character in both movies are in appendix in Figure 7. and Figure 8.

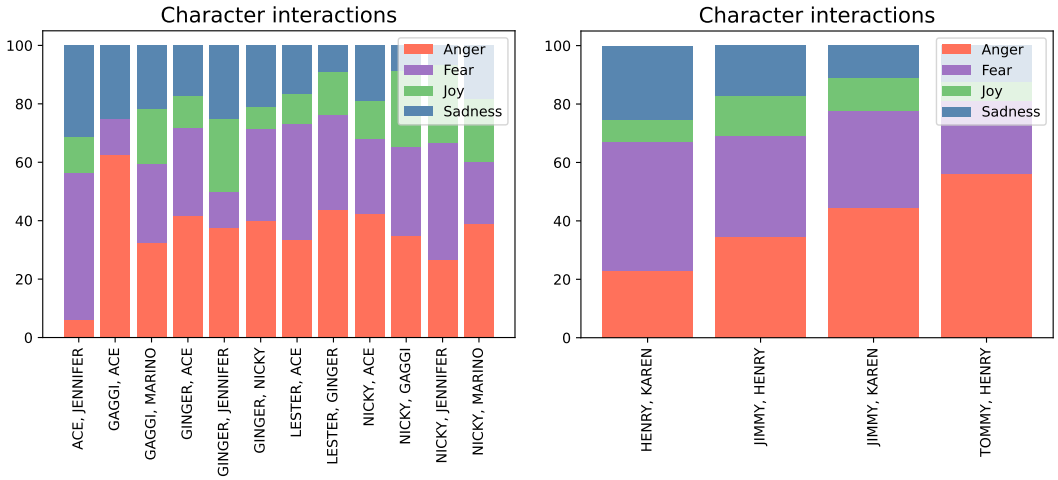

Figure 3: The percentage of emotions in both movies.

Figure 4. and Figure 5. show the emotions throughout the whole movies. It is a great way to see the peaks and analyze them better. All the peaks suit the scenes. For example, the peak of joy in Casino is when Ace is speaking about Ginger and how he loves her. In the scene where the anger is at its peak, Nicky, Ace, and Ginger were involved, and Ginger furiously smashed Ace's car.

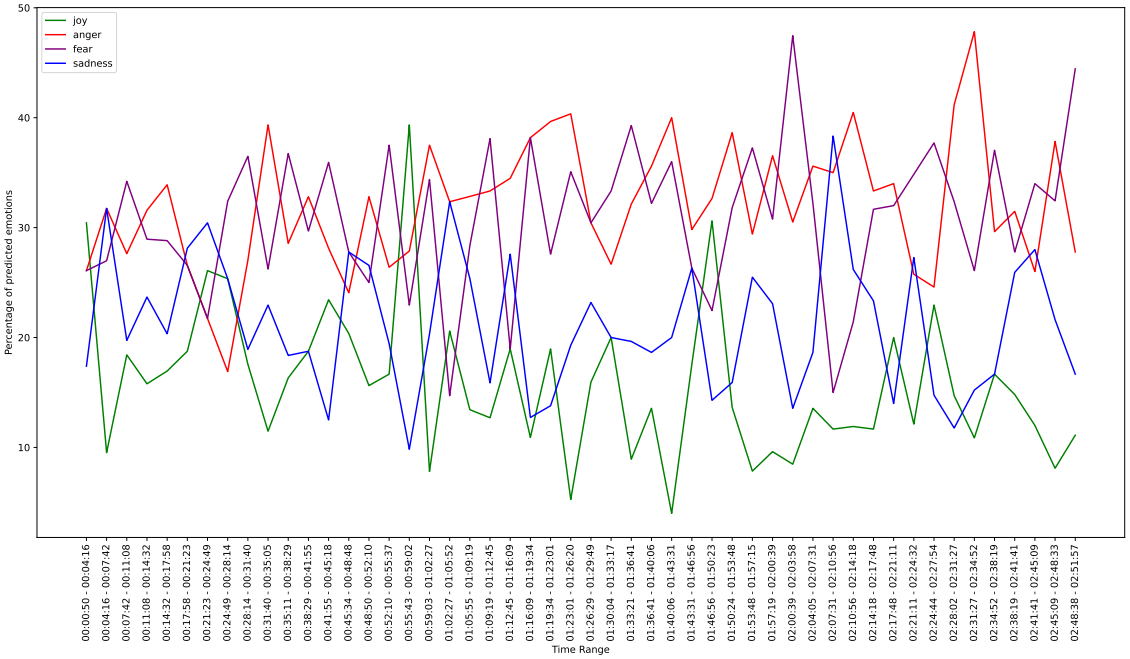

Figure 4: Casino - Emotion status throughout whole movie.

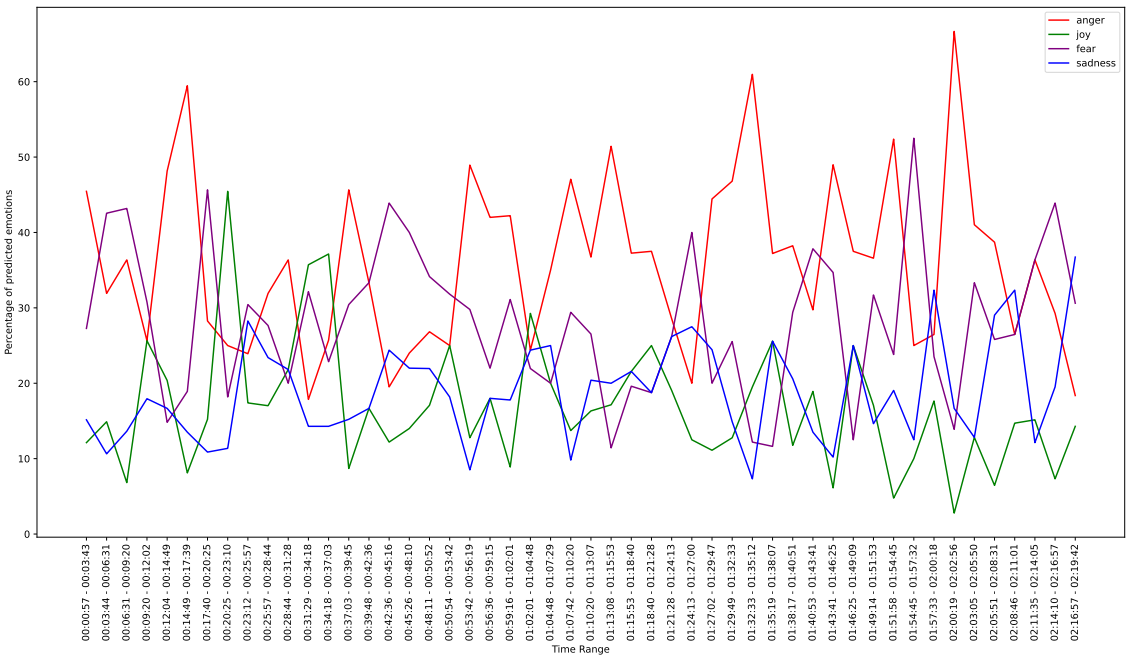

Figure 5: Goodfellas - Emotion status throughout whole movie.

In the end, we inspected similarities between characters and scenes across the movies. Figure 6. shows that Nicky and Tommy are the most similar characters. It is somehow expected, since those two characters had similar stories, and lines, they were similarly aggressive in both movies, and both are played by Joe Pesci (who acts in both movies in the same style). The table of similar scenes between movies is in the appendix in Figure 9.

Casino character	Goodfellas Character	Similarity
NICKY	TOMMY	0.995234
GAGGI	JIMMY	0.994413
GINGER	TOMMY	0.991113
ACE	HENRY	0.987514
ACE	JIMMY	0.986548
LESTER	HENRY	0.982325
MARINO	JIMMY	0.979159
NICKY	JIMMY	0.960231
LESTER	KAREN	0.959776
ACE	TOMMY	0.956709

Figure 6: Cosine similarity between characters across the movies.

Conclusion and future work

The algorithms used in this project gave satisfactory results on the WASAA-2017 dataset, but there could be improvements. For example, the dataset was slightly unbalanced, and there could be used more precise function to change random words with synonyms, instead of my custom-made one. Also, even though those algorithms were adequate, the deep neural network could be tried and used, but for it, we should probably increase the number of instances in the dataset. Even with this fast and "simpler" algorithm such as SGD classifier we could correctly catch and analyze some of the peaks in those two movies and also find similar characters and scenes. As future improvement, analysis could also be done on whole movie scripts, where we could find all emotions for each character and each conversation they made with others, since Cornell Movie-Dialogs Corpus does not have them all.

Appendix

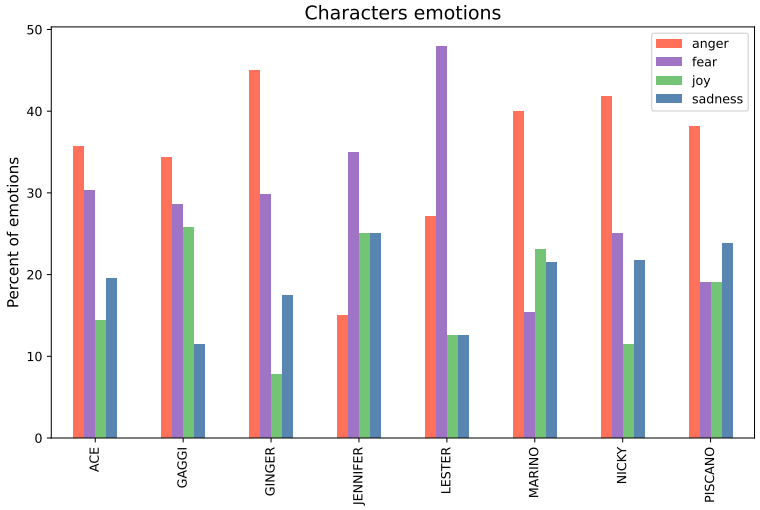

Figure 7: Casino - Percentage of emotions per each character in.

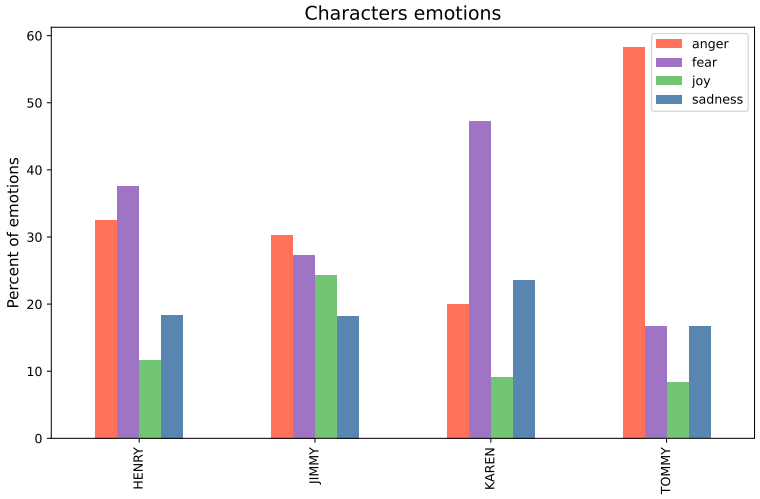

Figure 8: Goodfellas - Percentage of emotions per each character in.

Chunk	Time Range	fear	joy	anger	sadness	Chunk	Time Range	fear	joy	anger	sadness
26	01:26:29 - 01:29:49	34.782609	14.492754	30.434783	20.289855	10	00:25:57 - 00:28:44	34.042553	14.893617	29.787234	21.276596
45	02:31:27 - 02:34:52	26.086957	10.869565	50.000000	13.043478	25	01:07:42 - 01:10:20	25.490196	11.764706	50.980392	11.764706
16	00:52:10 - 00:55:37	41.666667	15.277778	22.222222	20.833333	17	00:45:26 - 00:48:10	42.000000	14.000000	22.000000	22.000000
50	02:48:38 - 02:51:57	41.666667	13.888889	27.777778	16.666667	7	00:17:40 - 00:20:25	41.304348	15.217391	28.260870	15.217391
15	00:48:50 - 00:52:10	21.875000	15.625000	32.812500	29.687500	36	01:38:17 - 01:40:51	23.529412	14.705882	32.352941	29.411765
25	01:23:01 - 01:26:20	33.333333	5.263158	42.105263	19.298246	22	00:59:16 - 01:02:01	33.333333	6.666667	40.000000	20.000000
44	02:28:02 - 02:31:27	32.352941	16.176471	39.705882	11.764706	15	00:39:48 - 00:42:36	33.333333	16.666667	37.500000	12.500000
42	02:21:11 - 02:24:32	39.393939	9.090909	27.272727	24.242424	31	01:24:13 - 01:27:00	40.000000	10.000000	25.000000	25.000000
30	01:40:06 - 01:43:31	34.000000	6.000000	42.000000	18.000000	22	00:59:16 - 01:02:01	33.333333	6.666667	40.000000	20.000000
15	00:48:50 - 00:52:10	21.875000	15.625000	32.812500	29.687500	30	01:21:28 - 01:24:13	23.809524	14.285714	33.333333	28.571429

Figure 9: Similar scenes between movies.