Text Mining and Sentiment Analysis Project
"How do you feel, my dear (P8)"

Natalia Kartasheva: 943350

1 Introduction

The aim of this project is to create a model able to predict emotions in text. Recently, emotion detection in text has received attention in the literature on sentiment analysis. Detecting emotions is important for studying human communication in different domains, including fictional scripts for TV series and movies. The project aims at studying fictional scripts of several movies and TV series under the emotional profile.
In particular, this project consists of the following steps:
1. To built a model to predict emotions in text using the WASSA-2017 dataset as training set;
2. To exploit the model to study an emotional profile of the main characters in one of the movies included in the Cornell Movie-Dialogs Corpus;
3. To study how this emotional profile changes in time along the evolution of the movie story and how it is affected by the various relations among the different characters.

2 Research question and methodology

In order to detect emotions in movie dialogues, in this work it is proposed an approach based on Random Forest classifier model that is trained on tweets. Cross-validation through all the data helps to validate the best model and accuracy of results.
For each of the emotions (anger, fear, joy, sadness) a model has been trained. For each of the character speech, each emotion was predicted. One of the main things is that we used not the predicted label, but the prediction probability. This is better to interpret: closer to 0 - emotion not detected, closer to 1 - emotion has been detected. Using this method allow us to see much smoother and accurate result.

3 Experimental results

3.1 Datasets

The training of the model has been performed on the Wassa-2017 dataset. The dataset is provided for four emotions: joy, sadness, fear, and anger. For example, the anger training dataset has tweets along with a real-valued score between 0 and 1 indicating the degree of anger felt by the speaker.

Cornell Movie-Quotes Corpus. This corpus contains a large metadata-rich collection of fictional conversations extracted from raw movie scripts.

3.2 Pre-processing

The Wassa-2017 dataset has been subjected to a common part of pre-processing consisting in removing non-english words from the text, remove punctuation, remove double spaces and numbers, remove stopwords, and lemmatization. Moreover, the intensity of the emotion, initially expressed as a continuous number between 0 and 1, has been categorized. In particular, the "noisy tweets", containing ambiguous levels of emotions (that correspond to an intensity between 0.45 and 0.55), have been removed in order to obtain a more precise model. The intensity has become a binary variable: 1 if the tweet contains an emotion with a intensity greater than or equal to 0.55 and 0 if the tweet contain an emotion with intensity less than or equal to 0.45. Also tweets have been cleaned by Twitter's features (retweet, hashtag, mentions, etc.).

3.3 Training data (Wassa-2017)

Using the training data, a Random Forest classifier was created. We trained 4 random forest classifiers: for each of the emotions (anger, fear, joy, sadness) a model has been trained. By cross-validation, we understood the accuracy of the prediction of each of the models more accurately than if we had done it using a regular partition.

anger model

```
vectorizer = TfidfVectorizer(tokenizer = lambda x: tokenize_tweet(x))
features = vectorizer.fit_transform(df_anger['tweet_cleaned'])
cross_v = KFold(n_splits = 5, random_state = 1, shuffle = True)

model_1 = LGBMClassifier()
model_2 = RandomForestClassifier()

scores_lgbm = cross_val_score(model_1, features, df_anger['intensity'],
                              scoring = 'accuracy', cv = cross_v, n_jobs = -1)

scores_rf = cross_val_score(model_2, features, df_anger['intensity'],
                            scoring = 'accuracy', cv = cross_v, n_jobs = -1)
```

fear model

```
vectorizer = TfidfVectorizer(tokenizer = lambda x: tokenize_tweet(x))
features = vectorizer.fit_transform(df_fear['tweet_cleaned'])
cross_v = KFold(n_splits = 5, random_state = 1, shuffle = True)

model_1 = LGBMClassifier()
model_2 = RandomForestClassifier()

scores_lgbm = cross_val_score(model_1, features, df_fear['intensity'],
                              scoring = 'accuracy', cv = cross_v, n_jobs = -1)

scores_rf = cross_val_score(model_2, features, df_fear['intensity'],
                            scoring = 'accuracy', cv = cross_v, n_jobs = -1)
```

joy model

```
vectorizer = TfidfVectorizer(tokenizer = lambda x: tokenize_tweet(x))
features = vectorizer.fit_transform(df_joy['tweet_cleaned'])
cross_v = KFold(n_splits = 5, random_state = 1, shuffle = True)

model_1 = LGBMClassifier()
model_2 = RandomForestClassifier()

scores_lgbm = cross_val_score(model_1, features, df_joy['intensity'],
                              scoring = 'accuracy', cv = cross_v, n_jobs = -1)

scores_rf = cross_val_score(model_2, features, df_joy['intensity'],
                            scoring = 'accuracy', cv = cross_v, n_jobs = -1)
```

sadness model

```
vectorizer = TfidfVectorizer(tokenizer = lambda x: tokenize_tweet(x))
features = vectorizer.fit_transform(df_sadness['tweet_cleaned'])
cross_v = KFold(n_splits = 5, random_state = 1, shuffle = True)

model_1 = LGBMClassifier()
model_2 = RandomForestClassifier()

scores_lgbm = cross_val_score(model_1, features, df_sadness['intensity'],
                              scoring = 'accuracy', cv = cross_v, n_jobs = -1)

scores_rf = cross_val_score(model_2, features, df_sadness['intensity'],
                            scoring = 'accuracy', cv = cross_v, n_jobs = -1)
```

3.4 Prediction data

After pre-processing of Cornell Movie-Quotes Corpus dataset, we can see a table with characterID, movieID, character name and text that they said.

	lineID	characterID	movieID	character name	text
0	L1045	u0	m0	BIANCA	They do not!
1	L1044	u2	m0	CAMERON	They do to!
2	L985	u0	m0	BIANCA	I hope so.
3	L984	u2	m0	CAMERON	She okay?
4	L925	u0	m0	BIANCA	Let's go.

In the table below we can also see an example of how the dialog between two characters looks like.

	lineID	characterID	movieID	character name	text
68	L194	u0	m0	BIANCA	Can we make this quick? Roxanne Korrine and ...
67	L195	u2	m0	CAMERON	Well, I thought we'd start with pronunciation...
66	L196	u0	m0	BIANCA	Not the hacking and gagging and spitting part...
65	L197	u2	m0	CAMERON	Okay... then how 'bout we try out some French...

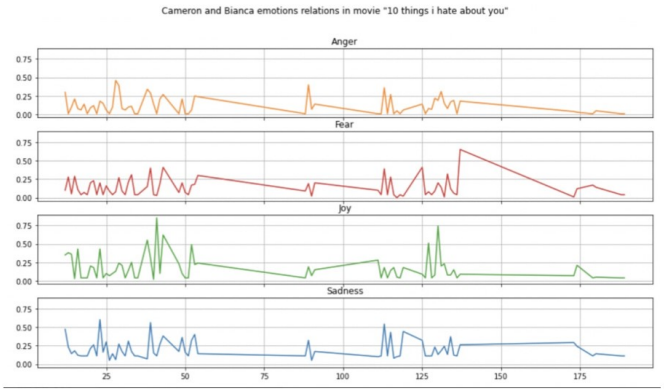
Moreover, in the table below we can see how we collected all the replicas in chronological order.

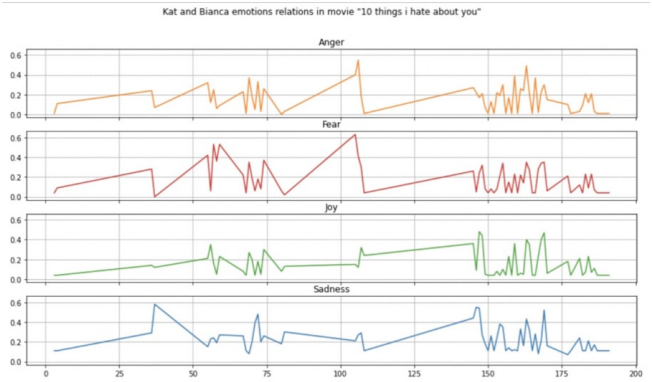
	lineID	characterID	movieID	character name	text
68	L194	u0	m0	BIANCA	Can we make this quick? Roxanne Korrine and ...
67	L195	u2	m0	CAMERON	Well, I thought we'd start with pronunciation...
66	L196	u0	m0	BIANCA	Not the hacking and gagging and spitting part...
65	L197	u2	m0	CAMERON	Okay... then how 'bout we try out some French...

Next, in the table below, we predicted the probability of every emotion (predict proba) for each replica of each character in each movie.

lineID	characterID	movieID_y	character name	text	anger_proba	fear_proba	joy_proba	sadness_proba
L194	u0	m0	BIANCA	Can we make this quick? Roxanne Korrine and ...	0.30	0.10	0.35	0.47
L195	u2	m0	CAMERON	Well, I thought we'd start with pronunciation...	0.01	0.28	0.38	0.23
L196	u0	m0	BIANCA	Not the hacking and gagging and spitting part...	0.10	0.05	0.36	0.14
L197	u2	m0	CAMERON	Okay... then how 'bout we try out some French...	0.21	0.29	0.03	0.18
L198	u0	m0	BIANCA	You're asking me out. That's so cute. What's...	0.08	0.11	0.43	0.12

At the end, we made a plot of change of emotions in the dialogues between the characters "Bianca" - "Cameron" and "Bianca" - "Kat" in the movie "10 things I hate about you". On the graphs, we can see that during the interaction with Cameron, Bianca experienced different emotions in the beginning and middle of the movie. By the end of the movie, they had almost no contact. Communication with Kat, on the contrary, increased by the end of the film and caused more and more sadness and joy, while fear and anger decreased. It should be noted that in this way, we can give thousands of different examples for any of the films in the list.




Kat and Bianca emotions relations in movie "10 things i hate about you"
Anger
Fear
Joy
Sadness
0.0
0.2
0.4
0.6
0
25
50
75
100
125
150
175
200