LTAT.01.001 Natural language processing

Detecting emotion intensity in tweets

Olha Kaminska and Viacheslav Komisarenko

UNIVERSITY OF TARTU
Institute of Computer Science

Introduction

Nowadays, detecting **emotions intensity** is interesting not only for scientists, but also actively using in business.

Social networks are perfect sources for gathering people thoughts, text is good way for sharing feeling and ideas, but these days actively developing hashtags, emoticons, links etc.

Twitter is one of the most popular social network, millions of tweets are written every day, which is a huge field for analysis.

In this work we propose model that predicts for four basic emotions - **anger, fear, joy or sadness** - their intensity in tweets. As a result, we obtained working model and interesting visualization.

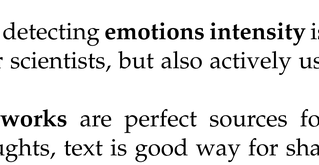
Materials and Methods

Preprocessing input data

As input we have file with **tweets and emotion intensity** for every emotion: anger, fear, joy, sadness.

(a) Tweet with text. **(b)** Score.

Figure 1: Example of anger input tweet.



(a) Tweet with text. **(b)** Score.

Figure 2: Example of joy input tweet.

Source text with emotional tweet was separated by emotions, then **preprocessed and tokenized**.

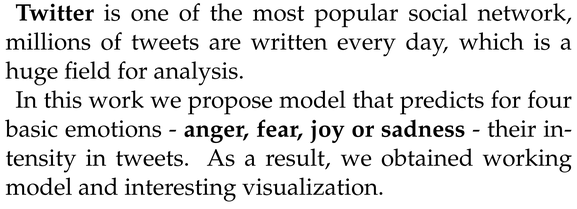


Figure 3: Tweet tokanization.

During preprocessing we deleted all tokens without sense or, potentially, emotions. We left meaningful words, emojis and hashtags without '#'-symbol.

Deleted tokens:

- Standard English stop words;
- Punctuation marks;
- HTML-tags / @-mentions of other users / URLs / numbers.

Features extracting

From obtained tokens we extracted useful features, which then were vectorized. From every tweet from each training set were extracted **1803 features**. The most features were based on **Word Embedding Features**, that was implemented using Google News 300-dimensional pre-trained word embeddings. Input emotions' intensity we used as regression labels.

Extracted features:

- Arithmetic mean and median of WEF vectors,
- Minimum and maximum for each coordinate among all WEF vectors,
- 25% and 75% percentiles for each coordinate,
- Total number of tokens in the tweet,
- Total number of characters in tweet,
- The average number of symbols in the token.

Machine learning models

All vectors and labels were used in several machine learning **regression** models. We used models from 'sklearn' library in Python with special set of parameters for each.

Quality of models was evaluated by **RMSE** (Root-mean-square error) function.

Used models:

1. Support Vector Regression with RBF, linear and polynomial kernels
2. Decision Tree Regressor
3. K-Neighbors Regressor
4. Random Forest Regressor
5. Gradient Boosting Regressor

Results

The best models

In the Table 1 represented **the best results** for each model. For every emotion was chosen the best.

Model	Parameter	Emotion	RMSE
SVR, rbf	c = 0.2	anger	0.0097
	0.1	fear	0.0088
	0.5	joy	0.0107
	0.1	sadness	0.0092
SVR, linear	0.01	anger	0.0088
	0.01	fear	0.0079
	0.01	joy	0.0103
	0.01	sadness	0.0084
SVR, poly	1.0	anger	0.0089
	1.0	fear	0.0079
	1.0	joy	0.0097
	1.0	sadness	0.0084
DT R	feats = 110	anger	0.0116
	10	fear	0.0118
	110	joy	0.0150
	210	sadness	0.0124
K-N R	k = 18	anger	0.0097
	16	fear	0.0089
	10	joy	0.0108
	16	sadness	0.0095
RF R	trees = 60	anger	0.0092
	110	fear	0.0082
	160	joy	0.0097
	110	sadness	0.0086
GBR	steps = 160	anger	0.0089
	210	fear	0.0080
	260	joy	0.0098
	60	sadness	0.0086

Table 1: The best results for all models

References

1. D.Davidov, O.Tsur and A.Rappoport. 2010. *Enhanced Sentiment Learning Using Twitter Hashtags and Smileys*
2. S.M.Mohammad, F.Bravo-Marquez. 2017. *Emotion Intensities in Tweets*
3. V.Duppada, S. Hiray. 2017. *Tweet Emotion Intensity Estimator*

Testing on real tweets

We tested the best models for every emotion on **new data**.

Figure 4: Tweet example.

As output we got **intensity** of all four emotions in these data.

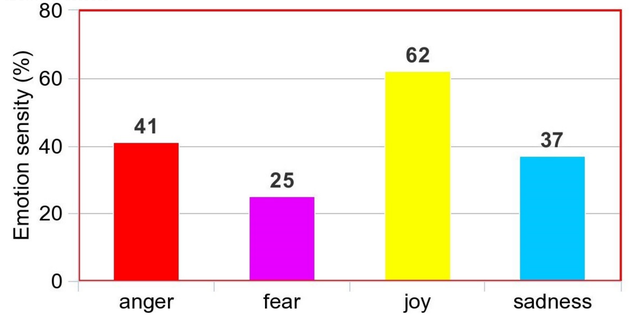


Figure 5: Emotions intensity in tweet.

Conclusions

- **Main idea** of this project was to build model that extracts emotion intensity from tweet for basic emotions: anger, fear, joy and sadness. As **result** was chosen the best regression model for each emotion by RMSE.
- Results of this project shows, that it is **possible** to detect emotion intensity from tweet - short text with a lot of noise, like links, emojis, hastags etc., with quite high accuracy.
- However, during work occurred some **difficulties**. The main one was feature extracting. Mostly features were chosen on a base of word embedding, also were attempts to use word-based features, but it is not improving results. Other important problem was parameters choices for different models.
- Results could be **improved** in several ways, one way is to perform more proper features selection, second is to try more complicated ensembles of models or neural networks.

Advisor: Kairit Sirts