ITMO University

Lab#1&Lab#2
Sentiment Analysis of Microblog Data Streams

Golovin Pavel, M4139c

23 октября 2020 г.

Содержание

1 Outline

- Preprocessing
  - List of methods that was used
- Data Analysis
  - Organization guessing
  - 3-way sentiment guessing
  - Using temporal data
  - Ideas for 5-way sentiment guessing
- Demo

2 Preprocessing

List of preprocessing steps:

- Normalize unicode

  First of all we need unify char encoding, because it influence on comparing symbols and seems like very generic and independent step.

- Expand contraction

  This step have no big effect, because produce mostly stop word.

- Replace emoticons (`:)`) with keywords like `<smile>`

- Unify latter case

- Replace mention and hashtag of companies (`@Apple`) with its name (`apple`)

  That's needed for save context from mention and hashtag removing

- Remove irrelevant structure like URL, HASHTAG, mentions, emoji, date, time, number

- Reduce repeated latter (`loooool -> lol`)

- Remove punctuation

- Normalize language

  Currently only english normalizing vocabulary is using.

- Remove stopwords

- Remove redundant whitespaces

3 Data Analysis

3.1 Organization guessing

- Features: words and character Tf-Idf vectorization.

- Classificator: LinearSVC

3.1.1 Results:

	precision	recall	f1-score	support
apple	0.95	0.96	0.95	98
google	0.84	0.77	0.80	79
microsoft	0.90	0.73	0.81	78
twitter	0.72	0.89	0.79	87
accuracy			0.85	342
macro avg	0.85	0.84	0.84	342
weighted avg	0.85	0.85	0.85	342

3.2 3-way sentiment prediction

- Features: word vectorization from organization prediction and result of that prediction (label of organization)
- Classificator: LinearSVC

3.2.1 Result:

	precision	recall	f1-score	support
irrelevant	0.74	0.93	0.82	105
negative	0.70	0.47	0.56	49
neutral	0.79	0.74	0.76	156
positive	0.53	0.50	0.52	32
accuracy			0.74	342
macro avg	0.69	0.66	0.67	342
weighted avg	0.74	0.74	0.73	342

3.3 3-way sentiment + temporal data

- Tries to take into account time of tweets by adding new categorial features: day and month.

3.3.1 Results: A bit better

	precision	recall	f1-score	support
irrelevant	0.73	0.93	0.82	105
negative	0.70	0.53	0.60	49
neutral	0.81	0.72	0.76	156
positive	0.52	0.50	0.51	32
accuracy			0.74	342
macro avg	0.69	0.67	0.67	342
weighted avg	0.74	0.74	0.73	342

3.4 5-way sentiment

We haven't 5 level labeling data, so we can try to extrapolate/interpolate 3 level sentiment marks. In my work classificator (SVC) was replaced with linear ridge regression in range from -1 (negative) to 1 (positive). And that range was split into 5 pieces:

- -2 – $(-\infty; -0.75)$
- -1 – $(-0.75, -0.25)$
- 0 – $(-0.25, 0.25)$
- 1 – $(0.25, 0.75)$
- 2 – $(0.75, +\infty)$

4 Demo for 3-way sentiment predictor

```
msg = "Google is very good"
time = "Tue Oct 18 21:53:25 +0000 2011"
test_df['weekday'] = test_df['TweetDate'].apply(lambda s: s.split()[0])
test_df['month'] = test_df['TweetDate'].apply(lambda s: s.split()[1])

a = sentiment_feature.transform(test_df['cleaned'])
b = time_feature.transform(test_df)

print(guess_org(msg, time)) # output: google
print(guess_sentiment(msg, time)) # output: positive
```