Twitter Sentiment Analysis

Christian Hotz-Behofsits, Dominik Pichler, Matthias Reisinger,
Thomas Schmidleithner, Florian Taus

Advanced Internet Computing

5th December 2014

Introduction

Aim

Aggregate sentiment values of different tweets by a value specific keyword via the Twitter API.

- Load Tweets from Twitter API
- Preprocessing
- Classification
- Aggregation of overall sentiment

Twitter API

Twitter offers several APIs for different purposes. All require authorization and a valid Twitter API-account.

- REST API
  - limited to 180 search queries/15 minutes
  - limited to 15 fetches/15 minutes (100 tweets each)
  - fetch all tweets at once
- Streaming API
  - no limitation regarding tweets per minute
  - continuous tweet-stream

Preprocessing

Preprocess a tweet message and extract important information (feature vector) for classification.

Tokenizer splits input into related parts (tokens). e.g. Stanford NLP PTBTokenizer, Carnegie Mellon Twokenizer

Tagger adds information to each token (e.g. word-class). e.g Stanford NLP MaxentTagger

token list filtering based on tagged-information.

Classification

- Aims at dividing inputs into at least two classes
- In this course: assign sentiments to textual inputs
- Supervised and unsupervised approaches
- Common supervised algorithms: Support vector machines (SVM), naive bayes, ...
    - Textual input in form of a *feature vector*
    - The feature vector contains relevant words such as verbs, adjectives and hashtags

Classification (Implementation)

- Based on WEKA
- SVM as classifier (LibSVM)
- Train model with stanford twitter test corpus[1] (manually reduced to 1000 positive and 1000 negative instances)
- Evaluation with testset (359 instances):
  64.35 % classified correctly

[1] http://help.sentiment140.com/for-students

Aggregate Classified Sentiments

Calcuate the weighted sentiment for each tweet whereby more recent dates are weighted heavier

- The weight of the most recent positive tweet is 1
- The weight of the oldest is 0.8
- The weights in between decrease linear
- $\sum_{i=0}^{\#tweets-1} 1 - i * \frac{decreaseFactor}{\#tweets-1}$
- Just positive tweets are weighted

Prototype

Some libraries and attributes of our prototype.

- Spring Boot
- REST Backend
- AngularJS Frontend
- PTBTokenizer and Twokenizer
- Maxent Tagger
- Weka

→ live demo!

Code available at: https://github.com/inkrement/aic