Project Proposal

Authors: Daniyar Nariman, Borgalinov Timur
Topic: Tweet classifier

Description

Nowaday, Twitter have become an important part of the daily life of many of users. This microblogging services are used as communication media, recommendation services, real-time news sources and information sharing sites. The large amount of new data created as result makes automatic analysis essential for processing this data. Thus, Twitter has become an attractive area for many studies such as text classification. Social media has become a common place where people manifest their opinions. Through sentiment analysis, companies can automatically process what their consumers write in natural language and get valuable insights in order to take decisions. Text classification aims at labeling natural language texts into a fixed number of predetermined categories.

Preliminary plan

We plan to analyze the tweets, with different approaches(see below), to explore which method fits best on different kind of information that we extract. Also we plan to try different kind of preprocessing methods to analyze what data is best suited for a particular method.

Dataset : https://www.kaggle.com/c/twitter-sentiment-analysis2/data

Pre-processing

- **Text cleaning**

Tweets contain different kind of noise that can harm machine learning algorithms performance. We need to carefully get rid of them. To this particular task we will take advantage of regular expressions.

- **Spelling correction**

People make typos (‘cudtomers’), use abbreviations (‘ppl’), acronyms (‘asap’) and different words have the same meaning (‘iphone’ and ‘phone’). These are just four examples of words that need to be fixed. This examples were found by inspecting the frequency vocabulary. For bigger datasets or automated processing spell checker can be used

- **NLTK features**

Baseline preprocessing methods. For example **wordnet** corpus to find out the senses of different words if needed.

Methods

- Logistic Regression
- SVM
- Gaussian Naive Bayes classifier
- KNN
- Ensemble classifier
- Convolutional Neural Network

Roles

Nariman Daniyar: CNN, Ensemble, GaussianNB, Preprocessing
Borgalinov Timur: LR, SVM, KNN, Preprocessing

Relevance to the course

With this project, we will focus on extraction of relevant information from tweets, based on different methods from machine learning and different kind of preprocessing and spelling correction methods.