Project Specification

Sentiment Classification of Social Media v.1.0

The purpose of this project is to let you experience the whole process of big data analysis, from collecting data from social media sites till building models for sentiment classification. The topic is about sentiment classification, which is the process of analyzing data and classifying it based on its sentiment conveying properties and the process has a multitude of applications in different industries.

1. Introduction

Social media is a growing source of data and information spread. However, the information is convoluted with varying interests, opinions and emotions. Moreover, the form of communication lacks standardized grammar, spelling, use of slang, sarcasm and abbreviations, and more. These variables can make extracting critical points, facts, and the sentiment of the message difficult in situations where a number of these aspects are present. Through natural language processing (NLP) it is possible to study and analyze these messages and objectively classify sentiments presented in social media.

Sentiment classification is the task of labeling data with a polarity through analysis of the properties contained within the data. Classification can be binary, meaning either positive or negative, or describe a detailed range of polarity at the expense of increased implementation complexity. Social media increases the complexity of the problem, necessitating analysis of informal communication which does not necessarily adhere to any grammatical or contextual rules. An interesting aspect of this topic is the difference between spoken and written language and evaluating which variables are the most important in conveying sentiment in written form.

2. Research Question

Sentiment classification in social media is difficult due to the informal nature of the communication. The informal nature introduces additional variables and properties that have to be evaluated compared to formal texts, necessitating additional resources spent on annotating the data and training the classifiers.

The aim of this project is to investigate the performance of a number of sentiment classifiers on data from social media by the learning-based approaches of naive Bayes (NB), Support Vector Machine (SVM) and other methods.

The center research question is:

Which classification approach performs best when evaluating social media texts?

3. The Data

A data set is required for training, testing the classifiers and measuring their performance. Each group should first collect a data set from some social media sites for training and validating your classifiers. One example of such a dataset is the VADER data set, consisting of 4200 Tweets that have been manually annotated by trained individuals and represent the gold standard in sentiment annotation. The data set was chosen because of its relevance to the subject of sentiment classification in social media, while retaining the often omitted informal features that are important in conveying sentiment. Tweets are also characterized by their short length (maximum 140 characters), which imposes additional challenges for determining the sentiment of the Tweet. Moreover, the data set has already been manually annotated. This eliminates the need for distant supervision and the baseline to which the results are compared will be highly reliable. Of course, you are free to choose any social media datasets to do the training.

The second part of this project is to collect **at least** 1 million Tweets or other social media data to evaluate your classifiers. In the report you have to give a detailed explanation of why you chosen the datasets in your project.

4. Programming Frameworks

The programming language of choice is Python because of its wide adoption in the software development industry and the scientific world, as a result the language has well-supported libraries providing NLP tools. The NLTK and Scikit-learn libraries offer

implementations of machine-learning classifiers, including NB and SVM and techniques for feature extraction.

5. Measuring Performance

You will classify Tweets into three classes: *positive*, *negative* and *neutral*. The simplest measure of performance regarding text classification is accuracy. Accuracy is calculated as the ratio of correct classifications divided by total classifications. However, accuracy is not a good indicator of performance if data is unbalanced. Recall and Precision and combination of these two are alternative measures to accuracy for determining classification performance. In contrast to accuracy, they are defined in terms of predicted and actual classes.

6. Milestone and Deadlines

Project (due Feb. 28)

- One project: Group size <= 3 students
- Checkpoints
    - Proposal: Title and Goal (due Feb. 19)
    - Implementation and Demo (due Feb. 28)
    - Project Report (due Feb. 28)
    - Project Presentation (due March 2)
    - Final Project Report and Course Evaluation Report (due March 6)
- Each group should submit one-page project proposal to Haibo on Feb. 19 in the lab time.
- Each group should finish the project and show your demo and results to Haibo on Feb 28.
- Each group should provide a ten-page document on the project; the responsibility and work of each student shall be described precisely, on Feb. 28.

- Each group should give a ppt presentation on your project in seminar 3 on March 2 (15-20 minutes).
- Each group will receive some feedback from teachers and should do revision accordingly on March 3.
- Each group will submit final report and one half page course evaluation report to Haibo on March 6, 18:00.