The best title in the world

Øyvind Selmer & Mikael Brevik

November 12, 2012

Contents

List of Figures

Chapter 1

Introduction

1.1 Background and Motivation

1.1.1 Task Description

Sentiment Analysis using the Twitter Corpus

In recent years, micro-blogging has become prevalent, and the Twitter API allows users to collect a corpus from their micro-blogosphere. The posts, named tweets are limited to 140 characters, and are often used to express positive or negative emotions to a person or product.

In this project, the goal is to use the Twitter corpus to do sentiment analysis. Pak and Paroubek (2010) have shown how to do this using frameworks like Support Vector Machines (SVMs) and Conditional Random Fields (CRFs), benchmarked with a Naive Bayes Classifier baseline. They were unable to beat the baseline, and the goal of this project will be to experiment with these and other machine learning frameworks as Maximum Entropy learners to try to beat the baseline.

1.2 Project goals

In this section we will describe our goals for this project. As this assignment was intended for a master thesis, we had to scope down the goals to fit the time schedule for a specialization project. We've scoped the assignment down to a set of goals that will make a good foundation for a master thesis.

1.2.1 Research on state-of-the-art sentiment analysis

A lot of work has already been done in the field of sentiment analysis, also when using the Twitter corpus. To be able to make a contribution to this work, we have to do research to gather knowledge about existing solutions and their performance. - Mention SLR here -

1.2.2 Python CLI for sentiment classification

We will design and implement a highly modular python CLI with with a basic form of classification. This system will work as a foundation for implementing the complete classification system in the proceeding projec. For that reason we want it to be as modular as possible, to make it easy to swap out different parts of the system when necessary.

1.2.3 API layer with twitter API integration

Twitter offers a well documented REST API to obtain data from their corpus. To make our system easy to use for developers already using the Twitter platform, we will implement a web service that is compatible with the Twitter API. And as an extension to the already existing API we'll add a *sentiment* attribute to the returned tweet. This attribute should hold the result of the sentiment classification.

Chapter 2

Background Theory and Motivation

2.1 Background Theory

2.1.1 Sentiment analysis

Naive Bayes Classifier

The naive Bayes classifier(NBC) is a practical Bayesian learning model that are easy to understand and implement. For some classification tasks it has proven to be equally performing to more complex classifiers like artificial neural networks(ANN) and decision trees(DT)[ref]. NBC is used for learning cases where an instance x consists of a number of attribute-value pairs, and the target function $f(x)$ consists of a finite number of values from a set V.

The NBC is based on an assumption that all the attribute values are conditionally independent given the target value of the instance.

$$v_{NB} = P(v_j) \amalg p(a_i|c_j) \tag{2.1}$$

To classify an instance, the classifier's using the Maximum Likelihood Estimation(MLE) method to find the ratio of an attribute value and a given target value in the same instance in the training corpus. This means that it has to calculate the probability estimate P for each attribute a_i, given the target value. It then assigns the target value as the one that gives the highest product from multiplying all the probabilities P from the training data.

2.1.2 Twitter API

2.1.3 Nodejs

2.1.4 Natural Language Toolkit for Python

Appendices

2.2 Systematic Literature Review Protocol

2.2.1 Introduction

This SLR protocol is developed during the specialization project of the fall semester 2012. This protocol will be used in both the specialization project and the master thesis. For the fall project this protocol and the SLR in general, will be used for the authors to gain sufficient knowledge about sentiment analysis using the Twitter corpus.

Twitter is a microblogging platform used by millions of people all over the world. In contrast to other social media platforms, the Twitter messages, called tweets, is limited to a maximum length of 140 characters.

The goal of this fall project is to implement a bare-bone, modular and highly customizable application for doing sentiment analysis on tweets. In addition an extension of the existing Twitter API (Application Programming Interface) will be developed. This will be achieved by mimicking the API interface and passing on the query to the Twitter API. By simply extending the API, the sentiment analysis data will be easy to use for existing Twitter developers, and already be heavily document by the Twitter API team.

The focus for this SLR is to search for papers with existing solutions for sentiment analysis on the Twitter corpus. To uncover the different performances and how the problem has been solved by other researchers. This information will be used to implement a basic sentiment analysis application for the specialization project and to implement a more sophisticated application for the master thesis project.

2.2.2 Research Questions

RQ1 What are some of the existing solutions for SA (sentiment analysis) in the Twitter Corpus.

RQ2 How does the different solutions found by addressing RQ1 compare to each other with respect to micro-blogs like Twitter.

RQ3 What is the strength of the evidence in support of the different solutions?

RQ4 What implications will these findings have when creating the application/system?

2.2.3 Search Strategy

ACM digital library	Mikael
IEEE Xplore	Mikael
ScienceDirect	Mikael
CiteSeer	Both
SpringerLink	Øyvind
ISI web of knowledge	Øyvind
Wiley Inter Science	Øyvind

Table 2.1: Domains used for search

The different search domains used for the SLR, can be found in Table 2.2.3. The source list is compiled based on the most common digital computer science archives.

A set of terms is defined closely based on the first research questions (**RQ1**). The terms are split into groups where one group consists of words that are synonyms or have similar semantic meaning.

All search terms is placed in a table with the groups as columns and search term as a row. The entire search string will be constructed by using Boolean notation. All terms in a group are concatenated by the keyword OR, and the groups themselves are concatenated by the keyword AND. This search string is represented by the following formula:

```
([G1, T1]) AND ([G2, T1] OR [G2, T2]) AND ([G3, T1] OR [G3, T2])
```

	Group 1	Group 2	Group 3
Term 1	Sentiment 1	Analysis	Twitter
Term 2		Classifier	Microblog

Table 2.2: Search terms and groupings

All results from found by using the search string, will be collected in a document, and reduced by removing duplicated papers, the same studies published from different sources and studies published before the year 2008.

2.3 Selection of primary studies

To reduce the studies even more, they are assessed using three different screenings; primary, secondary and by quality. The primary and secondary inclusion criteria is used to filter out the non-thematically relevant studies. The primary criteria is used on meta data such as title and abstract and the secondary is used on the full text paper. The quality screening is also used on the full text as the last step of selection.

Primary inclusion criteria

IC1 The study's main concern is Sentiment Analysis.

IC2 The study is a primary study presenting empirical results.

IC3 The study focuses on sentiment analysis on the english language.

Secondary inclusion criteria

IC4 The study focuses on the Twitter corpus.

IC5 The study describes an implementation of an application.

All studies that make it passed the primary and secondary selection criteria, will be passed on to the quality assessments.

2.3.1 Study quality assessment

To further filter the papers and assess the quality of the different papers, a set 10 of quality criteria is defined. The first two criteria is used in quality screening, to assess whether the papers includes a basic research data.

Each study should be classified according to all 10 quality criteria. They can either be classified as "Yes" (1 point), "Partly" (1/2 point) or "No" (0 points).

QC1 Is there is a clear statement of the aim of the research?

QC2 Is the study is put into context of other studies and research?

QC3 Are system or algorithmic design decisions justified?

QC4 Is the test data set reproducible?

QC5 Is the study algorithm reproducible?

QC6 Is the experimental procedure thoroughly explained and reproducible?

QC7 Is it clearly stated in the study which other algorithms the study's algorithm(s) have been compared with?

QC8 Are the performance metrics used in the study explained and justified?

QC9 Are the test results thoroughly analysed?

QC10 Does the test evidence support the findings presented?

2.3.2 Data Extraction

From each paper, the following data will be extracted for the SLR:

- Name of author(s)
- Title
- Study identifier
- Year of publication
- Type of article
- Aims, objectives and contributions of study
- Name of system
- Type of machine learning algorithm
- User modelling technique
- Experimental design
- Test set source
- Training set
- Findings and conclusions

The data will be presented in table format. Whereas the data type is divided into columns, and each paper is on its own row.