# Computer Science Tripos Part II Project Proposal

# Sentiment Analysis for Multilingual Tweets

M. Drozdzynski, King's College

Originator: Dr Daniele Quercia

22 October 2010

## Special Resources Required

File space on PWF – 2 GB
Developer Account at Twitter
Dataset of English tweets
Use of my own machine (MacBook Pro)

**Project Supervisor:** Dr Daniele Quercia

**Director of Studies:** Dr Simone Teufel

**Project Overseers:** Dr A. Copestake & Dr R. Harle

# 1 Introduction

Approaches for automatically classifying the sentiment (either positive or negative) of Twitter messages have been recently introduced. They are used by online shoppers to check the sentiment associated with products, or by companies to monitor the popularity of their brands. One limitation of those approaches is that they work for Twitter messages in English. This project considers the three most promising existing approaches (Naïve Bayes, Maximum Entropy, and SVM) and evaluates them for Twitter messages in a number of continental European languages. The three approaches will be made available through an API (similar to http://twittersentiment.appspot.com/).

# 2 Starting Point

This project builds upon two previously published papers on sentiment analysis of Twitter messages:

- Alec Go et al. – *Twitter Sentiment Classification using Distant Supervision*: [1]

- Hyung-il Ahn et al. – *"How Incredibly Awesome!" – Click Here to Read More*[2]

Significant research will need to be done around sentiment classification using Naïve Bayes, Maximum Entropy and Support Vector Machines. My prior experience in creating web applications and interacting with Twitter API will also be useful to this project.

The data set for English containing over 1,000,000 classified tweets has already been collected and will be provided by the Project Supervisor.

# 3 Substance and Structure of the Project

This project spans the domains of information retrieval, natural language processing and web development and can subsequently be divided into sections representing each of these three modules.

Work will commence by implementing a scraper which collect necessary training and test data from Twitter. The API provides an interface to query

[1] `http://www.stanford.edu/~alecmgo/papers/TwitterDistantSupervision09.pdf`

[2] `http://web.media.mit.edu/~hiahn/publications/ahn-etal-icwsm2010.pdf`

for Twitter messages in 19 different languages (Arabic, Danish, Dutch, English, Farsi / Persian, Finnish, French, German, Hungarian, Icelandic, Italian, Japanese, Norwegian, Polish, Portuguese, Russian, Spanish, Swedish, Thai) and allows for crawling tweets without violating any Terms of Service. The scope of this project will be limited to the following 5 languages, chosen to be spoken by either me or the Project Supervisor should a fallback to manual sentiment classification be required:

1. English
2. Spanish
3. German
4. Italian
5. Polish

The approach I describe will however be easily replicable to all remaining languages available via the API provided enough time to collect necessary training data.

In order to easily separate tweets into training sets of positive and negative sentiment, the scraper will run two individual queries for ':)' and ':(', with results from the former being qualified as positive and the latter as negative. Twitter API uses the following equivalence classes for emoticons: ':)', ':-)', ': )', ':D' and '=)' are mapped to ':)' and will subsequently be marked as positive; ':(', ':-(' and ': (' are mapped to ':(' and thus marked as negative. These emoticons will then be stripped off to avoid biasing the classifiers towards them.

As determined in previously cited paper on *Twitter Sentiment Classification using Distant Supervision*, this approach has been proven successful and will provide a working base for the classifiers. To further validate this approach, a random manual check will be carried out on 100 positive and 100 negative emoticons in each language to estimate accuracy of our training data set.

Further time will be spent studying the following classification algorithms: Naïve Bayes, Maximum Entropy and Support Vector Machines. I am hoping to be able to utilise existing solutions, hence research will be done into freely available implementations (including, but not limited to, *Weka*, *libsvm* and *SVM Light*) to verify if they meet the needs of this project.

Accuracy of classification algorithms will be measured by determining the number of correctly classified tweets. Two different metrics will be provided:

first determined using leave-one-out and second 5-fold cross-validation, provided the latter could be determined.

Later in the project significant work will be done on creating a web application to utilise classification tools implemented in previous sections. The main goal of the application is to aid potential users in monitoring sentiment for particular Twitter queries and plot it against time or location. An API will also be provided which will enable running search queries against Twitter and returning results supplemented by sentiment analysis.

# 4 Success Criteria

## Essential aims

For the successful completion of the project, the following criteria shall be met:

1. Collection of a limited number (1,000) of tweets in the following four languages: Spanish, German, Italian and Polish and implementation of feature reduction & duplicate filtering algorithms.

2. Evaluating the classifiers against the existing dataset of English tweets.

3. Preliminary results of classification using the limited dataset of Spanish, German, Italian and Polish tweets using the leave-one-out cross-validation technique.

4. API: Querying Twitter API for a given keyword or user and classifying returned tweets.

5. Application: Plotting sentiment for a given keyword or user over time.

## Desirable aims

1. Collection of a larger dataset for Spanish, German, Italian & Polish

2. Classification of the extended dataset and evaluation using 5-fold cross-validation technique.

3. Application: Plotting sentiment for a given keyword or user over location.

# 5 Plan of Work

## Aim

My aim is to consider and analyse three classification algorithms (Naïve Bayes, Maximum Entropy and Support Vector Machines) with respect to their accuracy in determining sentiments of multilingual tweets, as well as create an web application which can be used to plot sentiment of tweets over time or location.

## Preparation

Research into relevant topics will be carried out in the first 4 weeks of the project. During this time a scraper will be written and preliminary data will be collected for all four languages.

## Implementation

Approximately 10 weeks are allocated to the implementation of the project. An iterative methodology will be followed and upon successful completion of all iterations of the project all desired features will be included.

## Evaluation

My project will be evaluated against the criteria set out in section 4 of this document. A successful project would meet the essential goals, and if possible the desirable ones.

## Timeline and Milestones

The following timetable divides the time starting November 1st until the due date into 2-week iterations with individual goals to be achieved by the end of each block. Upon completion of these tasks, the final working project together with the dissertation will have been completed.

### Michaelmas Term

Weeks 1 to 2: WB 01/11/10 and 08/11/10

- *Research:* Twitter API
- *Research:* Sentiment classification

- Set up development environment
- *Implementation:* Twitter scraper with a working duplicate filter. Successfully mine samples of 20 tweets per language

Weeks 3 to 4: WB 15/11/10 and 22/11/10

- *Implementation:* Feature reduction algorithms – parsing out Twitter specific properties (usernames, links), normalisation of repeated letters
- Collect preliminary data sets for all four languages
- *Research:* Naïve Bayes, Maximum Entropy, Support Vector Machines

Weeks 5 to 6: WB 29/11/10 and 6/12/10

- *Research:* Data mining and classification software
- Outline the overall structure of the dissertation
- Complete the Introduction and Preparation chapters of the dissertation

**Christmas Vacation**

Weeks 7 to 8: WB 13/12/10 and 17/01/11

- *Implementation:* Naïve Bayes
- Running the classifier against tweets
- Evaluation of accuracy metrics
- Section on Naïve Bayes in Implementation and Evaluation chapters

**Lent Term**

Weeks 9 to 10: WB 24/01/11 and 31/01/11

- *Implementation:* Maximum Entropy
- Running the classifier against tweets
- Evaluation of accuracy metrics
- Section on Maximum Entropy in Implementation and Evaluation chapters

- *Milestone:* 04/02/11 Progress Report due

Weeks 11 to 12: WB 07/02/11 and 14/02/11

- *Implementation:* Support Vector Machines
- Running the classifier against tweets
- Evaluation of accuracy metrics
- Section on SVM in Implementation and Evaluation chapters

Weeks 13 to 14: WB 21/02/10 and 28/02/11

- *Implementation:* API for querying Twitter API and supplementing the results with sentiment analysis
- Section on API in the Implementation and Evaluation chapters

Weeks 15 to 16: WB 07/03/11 and 14/03/11

- *Implementation:* Application for plotting sentiment of tweets over time and, if time permits, location.
- Section on the Application in the Implementation and Evaluation chapters

**Easter Vacation**

Weeks 17 to 18: WB 21/03/11 and 28/03/11

- Refinements and finalisation

**Easter Term**

Week 19: WB 25/04/11

- Evaluation chapter complete

Week 20: WB 02/05/11

- Conclusion chapter complete

Week 21: WB 9/05/11

- Formatting and finalisation of dissertation. Printed and bound by end of the week.

Week 22: WB 16/05/11

- *Milestone.* 20/05/11 Dissertation Due