CM3203 – ONE SEMESTER PROJECT

Initial Plan

Building a taxonomy of tweet types and automatically classifying tweets into these types

Author
David HARRISON

Supervisor
Irena SPASIĆ

Moderator
Andrew JONES

Initial Plan

1. **Project Description**
2. **Project Aims & Objectives**
3. **Work Plan**

1. Project Description

People use Twitter for different reasons, e.g. business, personal, sharing information or emotion, etc, and broadcast tweets of different nature. The goal of this project is to analyse text data on Twitter to develop a taxonomy of the basic types of tweets. A corpus of tweets will then be collected and manually mapped to the classes in the taxonomy. The corpus be initially analysed manually in order to investigate the language usage across different types (e.g. personal messages probably start with pronouns such as 'I' or 'my'). After collecting an initial set of lexical (words) and syntactic (phrases) clues, a classifier will be implemented that will automatically map tweets to the most appropriate class in the taxonomy. The classification performance will be evaluated in terms of precision, recall and F-measure.

Twitter supplies a proportion of tweets as part of their *Streaming API*[1]. These are gathered in real time. Using this stream, this project intends to collect a relatively large sample of tweets which can then be categorised manually in order to create two datasets: training and testing.

Once annotated, the corpus of tweets can be analysed in order to build a set of classes to sort tweets into, as well as a set of rules with which the tweets can be sorted. These rules can then be implemented to automatically sort or classify a tweet in real time.

Hopefully, the project will also be able to learn from erroneous classifications that can be manually reclassified. These will feed back into the rules and decision making process to increase precision and accuracy of future classifications.

[1] Twitter. (2012). *Getting Started.* Available: https://dev.twitter.com/start. Last accessed 27th January 2014

2. Project Aims & Objectives

As mentioned in the Project Description, it is anticipated that this project will fall into two phases, in addition to an initial phase for research.

Initial Research	
Phase 1: Learning	Gather & Store Tweets in database
	Create web interface to display and allow tagging of tweets.
	Collect tags for each tweet.
Phase 2: Implementation	Identify classification & identifying features (e.g. phrases, words, etc.).
	Implement a system to automatically classify tweets.
Evaluation	

Naturally, this basic structure will be reflected in the Work Plan.

Phase 1: Learning

Within the learning phase, it will be necessary to design a method of storing tweets, and their accompanying data, in a way that allows for the data to be manipulated and analysed. Most importantly, it will be necessary to identify trends and patterns within the types of tweets stored.

Once stored, tweets will need to be manually tagged and classified. The simplest way to do this will be to create a web interface which allows users to manually describe the tweet. This information will then be saved in the same database.

Phase 2: Implementation

Building on the information gathered in the first phase, the next step will be to identify common traits of tweets that are similarly classified. As mentioned in the project description, it may be that personal tweets use first person pronouns more often, or that promotional tweets use phrases such as “win” or

“buy”. These trends and traits will then need to be codified in the form of features and/or rules that can be utilised by some automated system.

At this early stage in the project, there is no clear advantage to using any given programming language or technology. Most likely, the first phase will be undertaken using PHP and MySQL. However, for the second phase, this will be something that will need to be identified within in the project, possibly as part of the **Initial Research**.

Evaluation

Here it will be important to assess the performance of the implemented system. This will be done using qualitative evaluation as well as a quantitative assessment where the success rate of the system is compared to the similarity of the initial tweet sets using mathematical means.

3.Work Plan

January 2014

Mon	Tue	Wed	Thu	Fri	Sat	Sun
		1	*2*	*3*	*4*	*5*
6	*7*	*8*	*9*	*10*	*11*	*12*
13	*14*	*15*	*16*	*17*	*18*	*19*
20	*21*	*22*	*23*	*24*	*25*	*26*
27	28	29	30	31		

February 2014

Mon	Tue	Wed	Thu	Fri	Sat	Sun
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28		

March 2014

Mon	Tue	Wed	Thu	Fri	Sat	Sun
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

April 2014

Mon	Tue	Wed	Thu	Fri	Sat	Sun
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	*15*	*16*	*17*	*18*	*19*	*20*
21	*22*	*23*	*24*	*25*	*26*	*27*
28	*29*	*30*				

May 2014

Mon	Tue	Wed	Thu	Fri	Sat	Sun
			1	*2*	*3*	*4*
5	**6**	*7*	*8*	*9*	*10*	*11*
12	*13*	*14*	*15*	*16*	*17*	*18*
19	*20*	*21*	*22*	*23*	*24*	*25*
26	*27*	*28*	*29*	*30*	*31*	

Time outside project

Project Plan

Initial Research

Collect & Store Tweets

Build website

Collect manual classification data

Identify classification trends and key words

Implement classification system

Evaluation

Deliverable deadline

Dates shown in red are those outside of the project.

Those marked in italics fall outside of Cardiff University term dates[2]. Most notably, this includes Easter Recess from the 12th of April to the 4th of May.

[2] Cardiff University. *Semester Dates for Undergraduate Modular Programmes 2013/14.* Available: http://www.cf.ac.uk/regis/sfs/dates/1314/semester-dates-201314.html . Last accessed 3rd Jan 2014