MAP THE DEBATE

Micro Blog Sentiment Analysis

Joe Root

1. Introduction

1.1 Motivation

From women's rights to civil rights, the influence of public opinion on government policy has been pivotal. Within a healthy democracy the voice of the electorate should be heard and recognised by those chosen to represent them. Throughout history platforms have often be provided for public opinion to be made known, from the early public forums of Greece and Rome, to speaker's corner and the house of commons today. Providing a means for people to express their opinion enables them to both challenge and shape the direction their elected governments take. Finding ways of gathering and understanding this opinion has increasingly proven fundamental if a government wishes to be successful.

Current methods of measuring opinion are largely statistical, with methods such as polling looking at the opinion of a sample group, before using their results to make further predictions. These can be very accurate, however their small sampling rates mean that figures can often be askew. Furthermore polling is both costly and time consuming to conduct, and thus can neither be used to find opinion on breaking news or on a variety of topics. None the less, as methods for measuring public opinion have increased both in accuracy and detail, politicians and policy makers are starting to look to them not only for affirmation of their policies, but for guidance and new initiative.

As the web has become more prevalent throughout society, it is increasingly becoming a platform for discussion and opinion. The initial growth of blogging demonstrated the web's ability to serve as a forum for debate and opinion. None the less, the technical knowledge required to start a blog, alongside the time required to write a post meant that adoption was limited. In the past two years however, we have seen the rise of micro-blogging (essentially 140 character blog posts) through services such as Twitter. These have seen unprecedented levels of adoption, with Twitter's 200 million users posting 25 billion micro-blog posts in 2010. Due to the simple nature of writing short posts, micro-blog discussion tends to break quickly around news topics, and offers genuine insight into public opinion surrounding news topics.

This project hopes to utilise the growth of publicly available opinion on the web, using it as a source upon which new methods for analysing and measuring public opinion can be built. In particular the project will focus on understanding sentiment on micro-blogging services such as Twitter.

1.2 Contributions

Ruby implementation of a sentiment analysis engine based upon current research. The engine should be able to correctly identify sentences containing sentiment, and classify the sentiment as positive or negative. Different implementations should be tested and compared to determine which algorithms and techniques work best.

An algorithm for classifying sentiment as a range of emotion and feeling, rather than just a score along a scale of positive to negative. The algorithm should be implemented in Ruby and included in the sentiment analysis engine.

Research optimisations for current algorithms, in order to tailor the sentiment analysis engine for micro-blog posts from services such as Twitter. These optimisation should be implemented within the engine.

Research optimisations for current algorithms, in order to better facilitate the understanding of Politically focussed micro-blog data. These optimisation should be implemented within the engine.

A Ruby based Twitter module to store and classify live data from Twitter. Visualisations should be designed and implemented to help better understand the data and classification results.

2. BACKGROUND

Prior to the advent of the web little research had been conducted into sentiment analysis, largely due to the lack of digitalised opinion pieces. However, as the web has flourished as a platform for debate and opinion, there have been many significant studies into how we can better understand the sentiment of digitalised textual information. Much of the research conducted thus far has focussed on understanding the sentiment of online articles, often with a focus on gaining insight into product-based online reviews. These studies often combine elements of machine learning, natural language processing and linguistics in order to calculate measures of opinion along a scale of positive to negative.

Little detailed research has been published with regards to sentiment analysis on micro-blogs, and although the problems posed are not entirely disparate from traditional analysis, many differences do arise. The character constraint enforced by micro-blogs along with the volume and variety of their content, raises many problems which have not traditionally been considered within sentiment analysis. We will discuss these issues in more detail in *2.2 Sentiment analysis and micro-blogging*.

Current sentiment analysis focuses on measuring how positive or negative opinions directed at *objects* and their *attributes* are. Although informative when trying to understand the sentiment of reviews, the current measure is fairly limited as a measure of human emotion and response. Within *2.3 Expanding upon sentiment*, we will look at how we can develop current measures of sentiment, so as to express a broader spectrum of emotion and feeling.

2.1 State of sentiment analysis

In the following section, we will explore the general problems which have arisen within the field of sentiment analysis, before looking at current approaches to solving these problems and the theory behind their solutions.

2.1.1 Overview

In recent years, research into sentiment analysis has posed many challenging problems for researchers. Textual information can be broadly categorised as either opinion or fact. When trying to find the sentiment of an article, often the first problem posed is trying to identify sentences or phrases which are subjective, and thus have opinion, and those which are not. This has led researchers to develop methods for determining subjectivity, which we will look deeper into in *2.1.2 Subjectivity classification*.

Once we have identified those sentences which are subjective and from whom an overall understanding of sentiment can be gained, we are then left with two primary challenges. Firstly we must ascertain as to whom or what the opinion is being directed at. Broadly speaking it hopes to identify the target of discussion and in particular, for any given subjective sentence or phrase it should also identify the attribute of the target being discussed. This topic of feature extraction will be elaborated upon further in *2.1.3 Feature extraction.*

For any given subjective sentence or phrase, once the target object and attribute of discussion have been identified, the final problem posed is as to how sentiment can both be classified and measured. This is largely done by classifying the sentiment of individual phrases within a subjective sentence, and we will discuss this in *2.1.4 Sentiment* classification.

2.1.2 Subjectivity classification

Subjectivity classification is used within sentiment analysis to determine whether a sentence contains any form of opinion. As noted by Minahlcea et al. [1], and later re-iterated by Pang et al. [2], the problem of subjectivity classification is often more challenging than the eventual classification of sentiment.

The problem is generally approached through supervised machine learning techniques, often based upon naive Bayesian classifiers or support vector machines. Essentially these classification algorithms are provided with a set of objective and subjective sentences as training data. The properties of each sentence are analysed and used to determine how and which properties affect the subjectivity or objectivity of a sentence. Properties such as the number of opinion words and positioning of adjectives are used as measures of subjectivity by many classifiers.

The classification algorithms used in determining subjectivity are very much dependant upon both the quality and size of the training data set provided. Building a set of training data large enough to ensure a high quality of classification is difficult to approach manually due to the amount of time required to assemble it. As a result of the difficulty in manually compiling the training set, research has been conducted into developing high precision un-supervised algorithms which can accurately classify objectivity and subjectivity. These are then run on sentences, and the results are added to the training set, ensuring that a large enough set of examples can bet gathered.

When trained with a large and accurate data set, facilitated by un-supervised algorithms, supervised classifiers such as the naive Bayes classifier and support vector machines perform well.

2.1.3 Feature extraction

2.1.4 Sentiment classification

Sentiment classification aims to determine the polarity of sentiment in a subjective sentence or phrase. The majority of work conducted thus far into sentiment classification, aims to classify sentences in terms of the positivity or negativity of the data. Sentiment classification, like subjectivity classification, is largely achieved through a mixture of supervised machine learning techniques and un-supervised algorithms.

Discussion of un-supervised algorithms.

Discussion of supervised algorithms.

Like subjectivity classification, it has been found that with strong training data, traditional supervised classifiers perform well when classifying sentiment as positive or negative. Pang et al. [3] found that support vector machines in particular have strong accuracy rates, correctly classifying over 80% of sentences.

2.2 Sentiment analysis and micro-blogging

2.2.1 Micro-blogging and Twitter

2.2.2 Current approach

2.3 Expanding upon sentiment

3. References

[1] R Mihalcea, C Banea, J Wiebe - Learning multilingual subjective language via cross-lingual projections

[2] B Pang, L Lee - Opinion mining and sentiment analysis

[3] B Pang, L Lee - Thumbs up?: sentiment classification using machine learning techniques