Master Thesis Proposal: Sentiment Analysis of Social Media Content in the Context of Learning Environments

Seyed Ehsan Badakhshan

ehsan.badakhshan@rwth-aachen.de

**ABSTRACT**

In recent years, a noticeable attention has been directed to web documents as a new source of individuals’ opinions and experiences. This situation leads to an increasing interest in methods for automatically extracting and analyzing such opinions that are included in customer reviews, weblogs and comments on news. This information can be used in the context of learning environments, by considering the user’s emotional state over social networks. There has been a large amount of researches in the field of sentiment analysis on social media such as on twitter and Facebook. The purpose of this master thesis project is analyzing and comparing currently available methods in sentiment analysis focusing on using them in the context of education and learning environment environments of RWTH after tracking data over social networks.

**INTRODUCTION**

Nowadays, social media platforms such as Twitter and Facebook are popular microblogging services. They allow countless number of users to create and exchange unlimited number of content. In many cases, This content (called tweets in twitter and status updates in Facebook) express opinions about different topics. This includes statements that are related to universities’ topics and events. Such opinion rich data resources can be used to for extracting and analyzing opinions in terms of specific topics. Along with the help of data mining and natural language processing techniques it is possible to detect and analyze opinions related to learning context from large amount of data.

Sentiment analysis, also called opinion mining, is the field of study that analyzes people’s opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes [1].

The research field for this project will be natural language processing and data mining specifically it will explore existing sentiment analysis technologies in learning context in order to extract useful information to evaluate universities.

**RELATED WORKS**

Since the sentiment analysis - especially in social networks - has recently become an important topic, many researchers have focused and published their work in this area. There has been a large amount of prior research in sentiment analysis, especially in the domain of product reviews, movie reviews, and blogs [2].

There has been done researches in the field of sentiment analysis on social media specifically on twitter. Feature selection approaches such as using n-grams [3] [4] or Part of Speech (POS) tags [10] and lexicons [3] [5] has been examined in the context of sentiment analysis.

Researchers have also analyzed and compared data mining classification techniques such as Naïve Bayes classifier, Support Vector Machine (SVM) and Maximum Entropy for classifying tweets [6] [7]. Apart from using different features and classifiers, there are variety of used methods such as using emoticons [6], opinion reversal words etc., for identifying sentiments.

From the social media networks, Facebook is the more popular around the world [Resourse]. On October 2012, it reached 1 billion monthly active users (that is, 1 billion users accessed the network within a month) and more than 550 million daily active users [8]. One of the recent researches with the purpose of extracting information about users’ sentiments from the messages they write in Facebook [9]. They have developed a new and non-intrusive method for sentiment analysis in this social network. It consists on a hybrid approach, combining lexical-based and machine learning techniques to perform sentiment analysis in Facebook with high accuracy (83.27%).

**SOCIAL MEDIA SENTIMENT ANALYSIS**

The focus of this master thesis is in the context of education, just the tweets related to university learning environment is important in first phase. According to the pilot research that has been done in the IMA/ZLW & IFU at RWTH Aachen University, the process of sentiment analysis consists of 3 general phases.

1. Data collection

2. Data Processing

3. Test and evaluation

According to the study, from October 1st, 2014 till March 31st, 2015 there were 16488 tweets related to selected universities in Germany (TU9¹) in both English and German language have been posted on Twitter. If we subtract retweets from it, just 10189 original tweets in entire winter semester 2014/2015 has been collected. The biggest limitation associated with supervised learning is that it is sensitive to the quantity and quality of the training data and may fail when training data are insufficient. For solving this problem one solution would be collecting tweets from more universities. Another solution would be to consider another resources such as Facebook.

Data processing phase consists of preprocessing, feature selection and classification steps. In feature selection step, they adopted a combination of uni- and bigrams and they considered emoticons a part of n-gram features. In the classification step, they used naïve Bayes technique. Their classifier accuracy performance is 73.6%, while Go et. al. [mention the name][7] achieved around 80% accuracy rate. An idea to increase sentiment analysis accuracy rate would be to considering individual or mixture of different sentence features such as emoticons, parts of speech (POS) tags and lexicons, then testing other machine supervised learning classification techniques such as Support Vector Machine (SVM) and Maximum Entropy for classifying tweets.

Besides establishing a comparison between the TU9 based on the tweets related to each university, they investigated the tweets sentiment on daily basis for each university to obtain feedback on different events and activities. Comparison based on daily events is giving us a general result of sentiment analysis for each university. Another idea would be to classifying tweets based on topics such as Advertisement/Announcement, City News, Course/Class/Teaching/Professor, Exam/Homework, University Event/Sport Day, Party/Fun/Drinking and Conference, then applying a sentiment analysis methods on each of the topics for each university. On one hand, educational environments can make use of this information to come up with specific indicator. On the other hand, topic based sentiment analysis information can act as feedback for the university.

Add a section explaining the final intended output, is what form and what will it add to the previous study.

**PROPOSED TIMETABLE**

The estimated needed time to accomplish this project would be 6 months.

**4.1. Literature Review: 3 weeks**

We have a review to the current works and knowledge in the area of sentiment analysis on twitter. Sentiment analysis on Facebook would also be considered. We will compare our planed work with existing solutions.

**4.2. Preparation/Initialization Phase: 2 weeks**

In this step, we prepare some prerequisite of our work, including required tools and environment, etc. In addition, review some Python programming materials.

**4.3. Data collection: 1 week**

Sentiments collection from Twitter and Facebook APIs

**4.4. Text filtering: 2 week**

The process of cleaning tweets texts removing all irrelevant text for the sentiment classifier learning step.

**4.5. Features selection: 4 weeks**

One of the main parts of the project. Features are the sentence properties that we analyze in an attempt to correlate it to the tweet sentiment.

**4.6. Classification: 4 weeks**

Another important part of the sentiment analysis is sentiment classification. We are considering supervised classifiers which requires training and testing sets.

**4.7 Evaluation and discussion: 4 weeks**

The results section evaluates three main aspects. Measuring the classifier efficiency. Establishing a comparison between universities. Investigation the tweets on new aspects for each university to obtain feedback on different topics.

**4.8 Documentation 4 weeks**

We finally document our findings in the thesis. Nevertheless, there would be a continuous process of writing notes during the whole project.

**REFERENCES**

1. Alexander Pak, Patrick Paroubek. “Twitter as a Corpus for Sentiment Analysis and Opinion Mining” 2010.
2. B. Pang and L. Lee. “Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval”, 2008.
3. Kush Shah, Naser Munshi, and Pavan Reddy. “Sentiment analysis and opinion mining of microblogs”, 2013.
4. Apoorv Agarwal, et al. “Sentiment Analysis of twitter data”, 2011.
5. Efthymios Kouloumpis, et al. “Twitter Sentiment Analysis: The good the bad and the OMG”, 2011.
6. Go et al. “Twitter sentiment analysis”, 2009.
7. Go et al. “Twitter sentiment classification using distant supervision”, 2005.
8. Kiss, J. (10.04.2012). Facebook hits 1 billion users a month (Retrieved February 2013). The Guardian. <http://www.guardian.co.uk/technology/2012/oct/04/facebookhits-billion-users-a-month>.
9. Alvaro Ortigosa et al. “Sentiment analysis in Facebook and its application to e-learning”, 2014.
10. Kevin Gimpel et al. “Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments”, 2011.