Master Thesis Proposal

Sentiment Analysis of Social Media Content in the Context of Learning Environments

Seyed Ehsan Badakhshan
Student of M.Sc. Media Informatics RWTH Aachen University
Ehsan.badakhshan@rwth-aachen.de

Professors:

Prof. Dr. Sören Auer
department of Enterprise Information Systems (EIS)
University of Bonn

Univ.-Prof. Dr. rer. nat. Sabina Jeschke
Anas Abdelrazeq M.Sc

Institutscluster IMA/ZLW & IfU

November 2015

IMA ZLW IfU

universitätbonn

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 social media 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 which are included in customer reviews, weblogs and comments on news. This information can be useful in the context of learning environments as well, 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 Twitter and Facebook. The purpose of this master thesis project is to analyze and compare the available methods in sentiment analysis focusing on using them in the context of education and learning environment environments of universities 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 update 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

Sentiment analysis has been defined as the computational study of opinions, sentiments and emotions expressed in texts [2]. An example of a sentence transmitting a positive sentiment would be ''I love it!'', whereas ''It is a terrible movie'' transmits a negative one. A neutral sentiment does not express any feeling (e.g. ''I am commuting to work''). Most of works in this research area focus on classifying texts according to their sentiment polarity, which can be positive, negative or neutral [3].

With respect to the techniques that are used for sentiment analysis, two main approaches are considered: Machine-learning methods and lexicon-based approach. The survey written by Pang and Lee [3] covers the most popular techniques and approaches.

On the one hand, machine-learning methods such as Vector Machine (SVM), Maximum Entropy and Naïve Bayes are used to classify texts. An example of using machine-learning techniques in order to classify movie reviews is presented in Pang and Lee [4]. It compares different techniques to classify movie reviews, obtaining 82.9% of accuracy when applying Support Vector Machines (SVM).

Generally, it is difficult to obtain better results, due to characteristics of natural language. However, in specific domains, the use of machine learning algorithms for classifying texts according to their sentiment orientation performs well [13].

On the other hand, the lexicon-based approach consists of analyzing the text grammar and executing

a function to give a sentiment score to the text, considering a predefined sentiment lexicon [5] [6]. The advantage of the lexicon-based approach is that it is not necessary to have a labeled training set to start classifying texts. This approach tends to get worse results than machine learning approaches in specific domains, but when the domain is less bounded the results are better. This is because the lexicon approach analyzes the text grammar, whereas the machine-learning methods fit the algorithms to the training dataset particularities [13].

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 [3]. There has been done researches in the field of sentiment analysis on social media specifically on twitter.

An important part of sentiment analysis is feature selection. Features are the sentence properties that are getting analyzed in an attempt to correlate it to the tweet sentiment. Feature selection approaches such as using n-grams [7] [8] or Part of Speech (POS) tags [14] and lexicons [7] [9] has been examined in the context of sentiment analysis.

Researchers have also analyzed and compared machine-learning methods such as Naïve Bayes classifier, Support Vector Machine (SVM) and Maximum Entropy for classifying tweets [10] [11]. Apart from using different features and classifiers, there are variety of used methods such as using emoticons [10], opinion reversal words etc., for identifying sentiments.

From the social media networks, Facebook is the more popular around the world [13]. 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 [12]. One of the recent researches with the purpose of extracting information about users' sentiments from the messages they write in Facebook has developed a new method for sentiment analysis in this social network [13]. 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 three 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 (TU9[1]) 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 [15]. 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. [11] 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 supervised machine 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.

The final intended out of the master thesis is establishing a comparison between different machine-learning techniques to find the most accurate method for sentiment analysis in the context of learning environment.
Based on the results which are coming from testing and implementation of above mentioned ideas, an easy-to-use web-based platform for online sentiment analysis on education related data and report the result in different formats will be implemented.

CONCLUSION

The purpose of presented master thesis proposal is comparing sentiment analysis methods and using them in the context of education and learning environment of universities by tracking data over social networks such as Twitter and Facebook.
It will explore and compare the natural language processing and data mining techniques and applies them to existing sentiment analysis methods to extract useful information from social networks and implements a web-based sentiment analysis platform to report the results in different formats.

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. , (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *In Proceedings of the Seventh Conference on International Language Resources and Evaluation*. 3.1, 3.2, 6, 4.2, 4.3, 5.1

[2] Liu, B. (2010). Sentiment analysis and subjectivity. In N. Indurkhya & F. J. Damerau (Eds.), Handbook of natural language processing (pp. 627–666). Chapman and Hall: CRC Press.

[3] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1–2), 1–135.

[4] Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of Empirical Methods in Natural Language Processing (EMNLP) (pp. 79–86).

[5] Turney, P. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proceedings of the Association for Computational Linguistics (ACL) (pp. 417–424).

[6] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. Computational Linguistics, 37(2), 267–307.

[7] Shah, K., Munshi, N., Reddy, P. (2013). Sentiment analysis and opinion mining of microblogs. *University of Illinoise at Chicago, Course CS.* 2.3, 4.2

[8] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R. (2011). Sentiment Analysis of twitter data. In *Proceedings of the workshop on Languages in Social Media*, (pp.30-38).

[9] Kouloumpis, E., Wilson, T., Moore, J. (2011). Twitter Sentiment Analysis: The good the bad and the OMG! In *International Conference on Weblogs and Social Media (ICWSM), 2.3, 3.3, 4.2*.

[10] Go, A., Huang, L., Bhayani, R. (2009). Twitter sentiment analysis.

[11] Go, A., Huang, L., Bhayani, R. (2005). Twitter sentiment classification using distant supervision.

[12] 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>.

[13] Ortigosa, A., M. Martín, J., M. Carro, R. (2014). Sentiment analysis in Facebook and its application to e-learning. In *Journal of Computers in Human Behavior,* (pp. 527-541).

[14] Gimpel, K., Schneider, N., O'Connor B., Das D., Mills D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., Smith, N. (2011). Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, (pp. 42–47).

[15] Hajmohammadi, M., Ibrahim, R., Othman, Z., (2012). Opinion Mining and Sentiment Analysis: A Survey.