Master thesis proposal

Sentiment analysis of social media content in the context of learning environments

By

Seyed Ehsan Badakhshan

[Ehsan.badakhshan@rwth-aachen.de](mailto:Ehsan.badakhshan@rwth-aachen.de)

Supervisor

Anas Abdelrazeq M.Sc

Professor(s)

August 17, 2015

Contents

|  |  |
| --- | --- |
| 1. Abstract | 1 |
| 1. Introduction | 1 |
| 1. Related Work | 2 |
| 1. Social media sentiment analysis | 3 |
| 1. Proposed Timeline    1. Literature Review    2. Preparation/Initialization Phase    3. Data Collection    4. Filtering    5. Feature Selection    6. Classification    7. Evaluation and discussion    8. Documentation | 4 |
| 1. References | 7 |
|  |  |
|  |  |

Abstract

In the past few years, a great attention has been received by web documents as a new source of individual opinions and experience. This situation is producing increasing interest in methods for automatically extracting and analyzing individual opinion from web documents such as 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. (Some about each section . what I am going to go in abstract way)

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 contents. These contents (called tweets in twitter and status in Facebook) sometimes express opinions about different topics. This includes statements 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. With help of data mining and natural language processing techniques it could be detected and analyzed 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].

Sentiment analysis is got very popular in recent years and it mainly focuses on opinions which express or imply positive or negative sentiments.

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. Implementing sentiment analysis method or a combination of them in the learning environments of RWTH after tracking data over social networks.

The research field for this project will be natural language processing and data mining specifically we (passive) 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 papers 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 like using n-grams [3] [4] or Part of Speech (POS) tags and lexicons [3] [5] has been examined in the context of sentiment analysis. (mention more)

Researchers have also analyzed and compared data mining classification techniques like Naïve Bayes classifier, Support Vector Machine (SVM) and Maximum Entropy for classifying tweets [6] [7]. ( mention more as much as u can )

( add facebook)

Own problem

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

From October 1st, 2015 till March 31st, 2015 there were 16488 tweets related to the top-ranked universities in Germany (TU9¹) in both English and German language has been posted on twitter. If we subtract retweets from it, just 10189 original tweets in entire winter semester 2014/2015 has been collected. (not into detail for last study.) 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. Twitter sentiment classification using distant supervision] achieved around 80% accuracy rate. An Idea to increase sentiment analysis accuracy rate would be to considering different sentence features and testing other machine learning classification techniques( mention mother technique names and features name ).

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 more general and it is (giving us sentiment analysis ) of students about specific topics. Another Idea would be to classifying tweets based on topics such as university facility/sports/education and etc. for each university, then applying a sentiment analysis on each of the topics to come up with specific indicator such as education happiness (more details 🡪 help of classication of by students).

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.