A Study on Text Mining on Twitter: Identifying Opinion and Detecting Different Forms of Speech Using Writing Patterns

August 2019

Mondher Bouazizi

A Thesis for the Degree of Ph.D. in Engineering

A Study on Text Mining on Twitter: Identifying Opinion and Detecting Different Forms of Speech Using Writing Patterns

August 2019

Graduate School of Science and Technology
Keio University

Mondher Bouazizi

別表６
（３）

Keio University

Thesis Abstract

No.

Registration Number	■ "KOU" □ "OTSU" No. *Office use only	Name	Mondher Bouazizi

Thesis Title

A Study on Text Mining on Twitter:
Identifying Opinion and Detecting Different Forms of Speech Using Writing Patterns

Thesis Summary

Over the last two decades, online user-generated content has been exponentially increasing. With its increase, a proportionally increasing interest has been attributed to this data from the research community. While several works have been targeting different types of user-generated media such as photos, videos and audio content, text has always attracted most of the attention for several reasons. To begin with, due to the unique properties of natural languages, the analysis of such data presents several challenges. Nevertheless, hitherto, average internet users still use text more than any other type of media to interact with one another.

The studies performed on online generated text cover a wide range of types of analysis. These include but are not restricted to the analysis of motivations of users to share information, the evaluation of interests in events, the identification of prominent users, etc. Sentiment analysis, in particular, presents nowadays a hot topic of research. Sentiment analysis, also known as opinion mining, refers to the automatic identification and aggregation of opinions of people towards specific topics by analyzing their online written texts and publications. Sentiment analysis has several applications, ranging from product analytics to market analysis and public opinion orientation towards events such as elections, etc. Nevertheless, it is a field that is yet to be explored, with several of its challenges are yet to be dealt with. Instances of these include fine-grained sentiment analysis, evolution of sentiments over time, aspect-based sentiment analysis, etc.

On a related context, over the last decade or so, the focus of sentiment analysis has shifted from review websites, such as movie reviews websites, or online shops such as amazon etc., towards social media and microblogging websites. This is because these (i.e., social media and microblogging websites) have become the top attraction of online users, and the most visited and consulted platforms on the internet today. Twitter, in particular, has attracted a lot of attention, due to the ease of access to its data and the nature of the relationships between its users. That being the case, in our work, our experiments will be mostly conducted on data collected from Twitter.

This dissertation explores several of the challenges of sentiment analysis on social media, notably fine-grained sentiment analysis and sarcasm detection.

Chapter 1 introduces the concept of sentiment analysis on social media, its applications and challenges. We present several of the existing work which dealt with this task. We focus mainly on works on Twitter. However, relevant works which were performed on other social media or online websites will be presented as well. This chapter also summarized the scope and contribution of this dissertation.

Chapter 2 tackles a common challenge that has always been difficult to perform, yet very important to enhance the performance of sentiment analysis systems, i.e. the identification of sarcasm on social media. We use machine learning and the concept of patterns to identify sarcastic statements on Twitter. We run our experiments on a data set of texts posted on Twitter (i.e., tweets) and compare the performance of our proposed method to that of some conventional works. We also show how the identification of such statements can enhance the performance of sentiment analysis.

Chapter 3 focuses on a different task: multi-class sentiment analysis. As yet, most of the core of research on this field has been interested in the binary and ternary classification of texts. These refer to the classification of texts into positive and negative, and into positive, negative and neutral, respectively. Instead of limiting ourselves to such a coarse-grained classification, we go into a further level of granularity and classify texts into multiple sentiments. We re-use the concept introduced in the previous chapter, i.e., patterns, to perform this task. Alongside, we introduce SENTA (SENTiment Analyzer); a tool we have built that allows to extract, out of a wide variety of features, ones that can be used for applications such as sentiment analysis or sarcasm detection, through an easy-to-use graphical user interface.

Chapter 4 discusses in more details the results obtained in the previous one, explains the limitations of the task of multi-class classification which make it inherently difficult, and in some extreme cases impossible and describes the relation between sentiments and how correlated ones can be with some others. This chapter also offers possible solutions to overcome the limitations of multi-class sentiment analysis.

Chapter 5 presents a substitution to multi-class classification, which we refer to as Sentiment Quantification. Sentiment quantification refers to the identification of multiple sentiments expressed in a text, and attributing different scores to them to reflect their importance and weight within that text. In our proposed approach we use patterns and special type of unigrams to attribute scores to different sentiments to rank them and identify which ones are present in a given text, and which are not.

Finally, Chapter 6 concludes this dissertation highlighting its key points and the contribution made within, and proposes possible venues for future research of the topic of sentiment analysis.