Text Classification & Sentiment Analysis

Literature Review

*by*

Ian Santillan  
Student #: 500866649

Instructor: Ceni Babaoglu

Friday, June 15, 2018

Ryerson University

# Abstract

We examine sentiment analysis on Twitter data. In today’s world, micro-blogging sites has become a platform for individuals or organizations across the world to express their opinions, sentiment and experience in the form of tweets, status updates, blog posts, etc. This platform has no political and economic restrictions. This paper discusses an approach where a published stream of tweets from Stantec are then subjected to preprocessing and classified based on their emotional content as positive, negative and neutral. The performance of the unsupervised algorithm is then analyzed.

# Introduction

In recent years, a huge number of people have been attracted to social-networking platforms like Facebook, Twitter and Instagram. Most use social sites to express their emotions, beliefs or opinions about things, places or personalities.

Twitter is a social networking and microblogging service that lets its users post real time messages, called tweets.

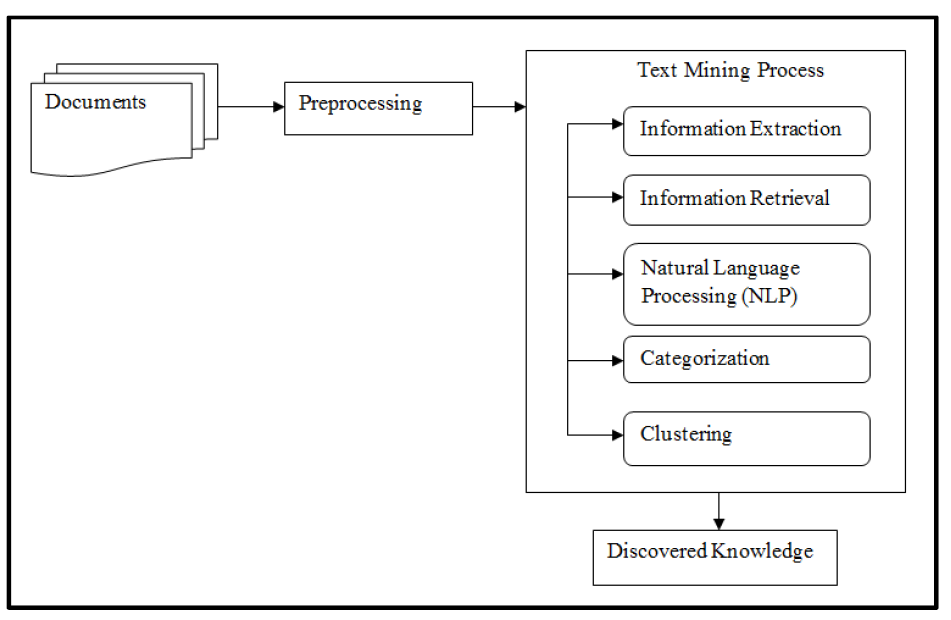
Tweets have many unique characteristics, which implicates new challenges and shape up the means of carrying sentiment analysis on it as compared to other domains.

Following are some key characteristics of tweets:

* *Message Length:* The maximum length of a Twitter message is 280 characters (was originally 140 chars). This is different from other sentiment classification research that focused on classifying longer texts, such as product and movie reviews.
* *Writing style:* The occurrence of incorrect spellings and cyber slang in tweets is more often in comparison with other domains. As the messages are quick and short, people use acronyms, misspell, and use emoticons and other characters that convey special meanings.
* *Availability:* The amount of data available is immense. More people tweet in the public domain as compared to Facebook (as Facebook has many privacy settings) thus making data more readily available. The Twitter API facilitates collection of tweets for training.
* *Topics:* Twitter users post messages about a range of topics unlike other sites which are designed for a specific topic. This differs from a large fraction of past research, which focused on specific domains such as movie reviews.
* *Real time:* Blogs are updated at longer intervals of time as blogs characteristically are longer in nature and writing them takes time. Tweets on the other hand being limited to 140 letters and are updated very often. This gives a more real time feel and represents the first reactions to events.

# Text Classification

### Text Mining Steps



Text categorization (a.k.a. text classification) is the task of assigning predefined categories to free-text documents. It can provide conceptual views of document collections and has important applications in the real world. For example, news stories are typically organized by subject categories (topics) or geographical codes; academic papers are often classified by technical domains and sub-domains; patient reports in health-care organizations are often indexed from multiple aspects, using taxonomies of disease categories, types of surgical procedures, insurance reimbursement codes and so on. Another widespread application of text categorization is spam filtering, where email messages are classified into the two categories of spam and non-spam, respectively.

# Sentiment Analysis

Sentiment analysis is a type of data mining that measures the inclination of people’s opinions through natural language processing (NLP), computational linguistics and text analysis, which are used to extract and analyze subjective information from the Web - mostly social media and similar sources. The analyzed data quantifies the general public's sentiments or reactions toward certain products, people or ideas and reveal the contextual polarity of the information.

Sentiment analysis is also known as opinion mining.

# Conclusion

Sentiment analysis is an evolving field with a variety of use applications. Although sentiment analysis tasks are challenging due to their natural language processing origins, much progress has been made over the last few years due to the high demand for it. Not only do companies want to know how their products and services are perceived by consumers (and compare to competitors), but consumers want to know the opinions of others before making buying decisions.

The growing need for product insights – and the technical challenges currently facing the field –will keep sentiment analysis and opinion mining relevant for the foreseeable future. Next-generation opinion mining systems need a deeper bind between complete knowledge bases with reasoning methods inspired by human thought and psychology. This will lead to a better understanding of natural language opinions and will more efficiently bridge the gap between unstructured information in the form of human thoughts and structured data that can be analyzed and processed by a machine.

The result: intelligent opinion mining systems capable of handling semantic knowledge, making analogies, continuous learning, and detecting emotions — leading to highly efficient sentiment analysis.

# References

Ali Hasan, S. M. (2018). Machine Learning-Based Sentiment Analysis for Twitter Accounts. *Mathematical and Computational Applications*.

Ankit Pradeep Patel, A. V. (2017). Literature Survey on Sentiment Analysis of Twitter Data using Machine Learning Approaches. *IJIRST –International Journal for Innovative Research in Science & Technology | Volume 3 | Issue 10*.

Janani, D. S. (2016). *Text Mining: Open Source Tokenization Tools - An Analysis.* Advanced Computational Intelligence: An International Journal (ACII), Vol.3, No.1, January 2016.

Joachims, Y. Y. (2008). *Text\_categorization*. Retrieved from Scholarpedia: http://www.scholarpedia.org/article/Text\_categorization

Kastrenakes, J. (2018, February 8). *Twitter says people are tweeting more, but not longer, with 280-character limit*. Retrieved from theverge: https://www.theverge.com/2018/2/8/16990308/twitter-280-character-tweet-length

Katrekar, A. (n.d.). *An Introduction to Sentiment Analysis.* GlobalLogic.

Sebastian, A. K. (2012). Sentiment Analysis on Twitter. *International Journal of Computer Science Issues | Vol. 9 | Issue 4 | No. 3*.