ILLINOIS STATE UNIVERSITY

SENTIMENT ANALYSIS OF TRENDING TWITTER HASHTAGS

by

Alberta Savage

PREDICTIVE ANALYTICS COMPETITION

April, 2019

Declaration

This essay was written in the Department of Mathematics, Illinois State University from October 2018 to March 2019, in partial fulfillment of the requirement for the completion of the Predictive Analytics Competition

Abstract

The importance of understanding data, particularly text data has recently been recognized as a vital part of service improvement for most businesses. However, improving the quality of service by understanding real user and customer opinions, complaints and suggestions remains a major challenge and companies have been leveraging the power of data to enable them gain meaningful insights into constantly improving goods and services offered. Journalism would greatly benefit from knowing the perceived reactions of stories in the daily news cycles. Analyzing textual data from tweets with attached hashtags would provide information on how the twitter community reacts to news stories and major breaking news around the country and the world. The focus of this project is using the power of deep learning and intelligent classifiers such as Sentiment Analysis to draw meaningful conclusions from text data on Twitter using the very powerful Vader package in Python. This project focuses on the use of information contained in tweets to categorize the sentiment attached to a hashtag (particularly trending hashtags) contained in those tweets.

Table of Contents

Chapter 1

Background of the study

1.1 Introduction

Twitter is an American online news and social networking service with over 25 offices around the world that allows users to post and interact with messages that are known as "tweets". From its creation in March 2006, Twitter has rapidly gained worldwide popularity and even proved to be the largest source of breaking news on the day of the 2016 US Presidential Elections with about 40 million election tweets sent by 10:00pm (Eastern Time).

One interesting feature of Twitter is the use of hashtags which make it possible to find messages with a specific theme or content and these hashtags are simple created by placing the pound sign, also known as the hash character in front of a string of alphanumeric characters. Trending Hashtags are hashtag-driven topics that immediately become popular at a time and e-commerce businesses especially capitalize on the current conversation of what might be holding consumer interest.

Hashtags have become an important tool for sentiment analysis which is a process of identifying and categorizing opinions expressed in a piece of text to determine whether the attitude of a writer towards a topic or product is positive, negative or neutral.

1.2 Literature Review

Natural language processing is a subfield of computer science, information engineering, and artificial intelligence concerned with the interactions between computers and human(natural) languages and how to program computers to process and analyze large amounts of natural language data Sentiment analysis or opinion mining is an area of natural language processing that is useful to a wide range of problems that are of interest to human-computer interaction practitioners and researchers, as well as those from fields such as sociology, marketing and advertising, psychology, economics and political science. Sentiment analysis builds systems that try to identify and extract opinions within text such as

1. Polarity: how positive or negative the speaker is
2. Subject: the focus of the speaker
3. Opinion holder: the speaker or person expressing opinion

Just like many other Natural Language Processing problems, sentiment analysis can be modeled as a classification problem with two parts

1. Subjectivity Classification: Classifying a sentence as subjective or objective

2. Polarity Classification: classifying a sentence as expressing a positive, negative or neutral opinion

A very large number of sentiment analysis approaches rely heavily on an underlying sentiment (or opinion) lexicon which refers to a list of lexical features like words which are labelled according to their semantic orientation as either positive or negative. Some pre-existing manually constructed lexicons, 100% curated by humans include LIWC, GI and Hu-Liu04

LIWC uses a proprietary dictionary of about 4500 words organized into one or more of 76 categories but one main disadvantage of this lexicon is that it does not include considerations for lexical items such as acronyms, emoticons or slang. LIWC is also unable to account for the differences in sentiment intensity of words.

General Inquire(GI) is designed for content analysis and this lexicon contains more than 11k words classified into one or more of 183 categories. This lexicon suffers from the same shortfall that the LIWC suffers from; lack of coverage of sentiment-relevant features and ignorance of intensity differences in sentiment-bearing words.

The Hu-Liu04 lexicon contains about 6800 words and is more attuned to sentiment expressions in social media text and product reviews but it is still unable to capture sentiment from emoticons or acronyms. VADER on the

other hand leverages the advantages of parsimonious rule-based modeling to construct a computational sentiment analysis engine

1. that works well on social media style text and can also be extended to multiple domains
2. requires no training data, but is constructed from a generalized, valence-based, human-curated gold standard sentiment lexicon
3. fast enough to be used online with streaming data
4. does not severely suffer from a speed-performance trade-off

This lexicon is constructed by examining already existing well-established sentiment word-banks and incorporating a full list of Western-style emoticons, sentiment-related acronyms and initialisms as well as commonly used slang with sentiment value. The resulting gold standard lexical features are about 7500.

Machine Learning Approach

Manually creating and validating a comprehensive sentiment lexicon is labor intensive, therefore automated means of identifying sentiment-relevant features in text have been explored. These machine-learning approaches "learn" the sentiment-relevant features of text and they include

1. Naive Bayes(NB) classifier which assumes that feature probabilities are independent of one another

2. Maximum Entropy, a general-purpose machine learning technique belonging to a class of exponential models that uses multinomial logistic regression

3. Support Vector Machines; non-probability classifiers which operate by separating data points in space using hyperplanes

These machine learning approaches require training data, which are can be very hard to obtain. They also depend on the training set to represent as many features as possible (which often, they do not especially in the case of short, sparse social media text). They are computationally expensive (restricts the ability to assess sentiment on streaming data). ML techniques often derive features "behind the scene" inside of a black box, not easily interpreted by humans and can be difficult to generalize to other domains. The applications of sentiment analysis are endless and may include the following

1. Predict the onset of depression in individuals based on social media text

2. Measure national happiness based on Facebook updates

3. Differentiating happy romantic couples from unhappy ones based on instant message communications

4. Characterizing the emotional variability of pregnant mothers from Twitter posts

Chapter 2

Data and Methods

2.1 Data

Real time tweets and trending hashtags would be streamed using the Twitter Application Program Interface (API). The VADER (Valence Aware Dictionary for sEntiment Reasoning) package would be used for the sentiment analysis. The trending hashtags as at April 1, 2019 that would be considered for this analysis are

1. $\#SHAZAM$
2. $\#NipseyHussle$
3. $\#cosn2019$
4. $\#wrestlemania$
5. $\#LeBron$

2.2 Method

1. A developer account is created on twitter to allow access to tweets. Unique credentials are generated to give user permission to stream tweets

2. Tweets are live-streamed based on trending hashtags

3. A sentiment classification system is built to determine the general sentiment of a trending hashtag using the text data contained in the tweet in Python.

It should be noted that the tweets are real time and depict the natural language terminology that exist in the world of social media and some tweets may contain strong and harsh expressions.

Chapter 3

Analysis

The VADER package is used for the analysis because of its suitability for the analysis of social media text.

500 of the most recent tweets of each trending hashtag are streamed and sentiment analysis is performed on each tweet.

The sentiment score of a sentence is calculated by summing up the sentiment scores of each VADER-dictionary-listed word in the sentence

1. Individual words have a sentiment score between -4 and 4
2. Sentence sentiment score is between -1 to 1 due to normalization
3. The averages of each score in the various categories (negative, neutral, positive and compound) are calculated
4. The average compound sentiment is a good indication of the overall sentiment that people are associating with the hashtags

A demonstration of the power of sentiment analysis can be seen in the analysis of tweets containing $\#NipseyHussle$.
A compound score of 0.04% is produced which indicates that the general sentiment associated with this hashtag is negative. Further investigation shows that Nipsey Hussle was a popular artist who was pronounced dead on April 1 2019 and the low sentiment score is reflective of how badly people are feeling about the hashtag.

Some more analysis of other trending hashtags are provided in Table 1 in the appendix.

Chapter 4

Conclusion

VADER has proven to be the best sentiment analysis tool for social media text, even outperforming the already existing tools such as LIWC, ANEW, GI and Hu-Liu04. It proved to be very useful particularly in this study and worked very well with live tweets. I was successfully able to access the API of Twitter and perform sentiment analysis on a couple of trending hashtags.

For future studies, I would like to create an app that can be accessed by every user on twitter for sentiment analysis using a combination of the different lexicons available.

Bibliography

[1] https://twitter.com/

[2] https://en.wikipedia.org/wiki/Hashtag

[3] https://www.cs.cornell.edu/home/llee/papers/sentiment.pdf

[4] http://crowdsourcing-class.org/assignments/downloads/pak-paroubek.pdf

[5] https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-ap

[6] https://programminghistorian.org/en/lessons/sentiment-analysis

[7] https://www.analyticsvidhya.com/blog/2018/07/hands-on-sentiment-analysis-

[8] https://www.geeksforgeeks.org/twitter-sentiment-analysis-using-python/

[9] https://en.wikipedia.org/wiki/Natural-language-processing

[10] https://www.analyticsvidhya.com/blog/2018/02/natural-language-processing

[11] https://en.wikipedia.org/wiki/Wisdom-of-the-crowd

[12] https://monkeylearn.com/sentiment-analysis/

[13] https://www.uvm.edu/pdodds/teaching/courses/2009-08UVM-300/docs/others/e

Appendix

Table 1

Hashtag	negative	neutral	positive	Compound sentiment
SHAZAM	2.12%	87.79%	10.09%	18.03%
NipseyHussle	11.01%	75.73%	13.26%	00.04%
cosn2019	1.32%	84.41%	14.26%	35.64%
wrestlemania	3.56%	83.95%	12.49%	22.11%
LeBron	5.12%	85.27%	9.61%	12.03%
Grammy	2.81%	85.82%	11.36%	23.41%
FridayFeeling	1.77%	86.82%	11.42%	24.00%