Using Binary Classifier for Predicting Success of Technology Release Based Upon Sentiment Analysis of Social Media Data

Ali Akbar Jilani
Waterford Institute of Technology, Cord Road, Waterford, Ireland

Abstract

User Generated Content (UGC) over social media can be tapped to extract underlying public sentiment and take relevant actions. Proposed study uses sentiment analysis on twitter data to investigate its correlation with sentiments of same products as measured through Amazon user reviews. The study goes further into supervised machine learning techniques to train a binary classifier on the patterns of sentiment words occurring in the same twitter data to test its correspondence with online user reviews in the technology product domain.

Research Question

The purpose of this research is to measure public sentiments expressed by twitter users in the form of tweets posted about their experiences of technology, and to tally these sentiments with their equivalents of technology reviews found in sites such as Amazon. This research investigate the correlation between these sentiments as recorded through both the above sources. The underlying research question can therefore be defined as the following.

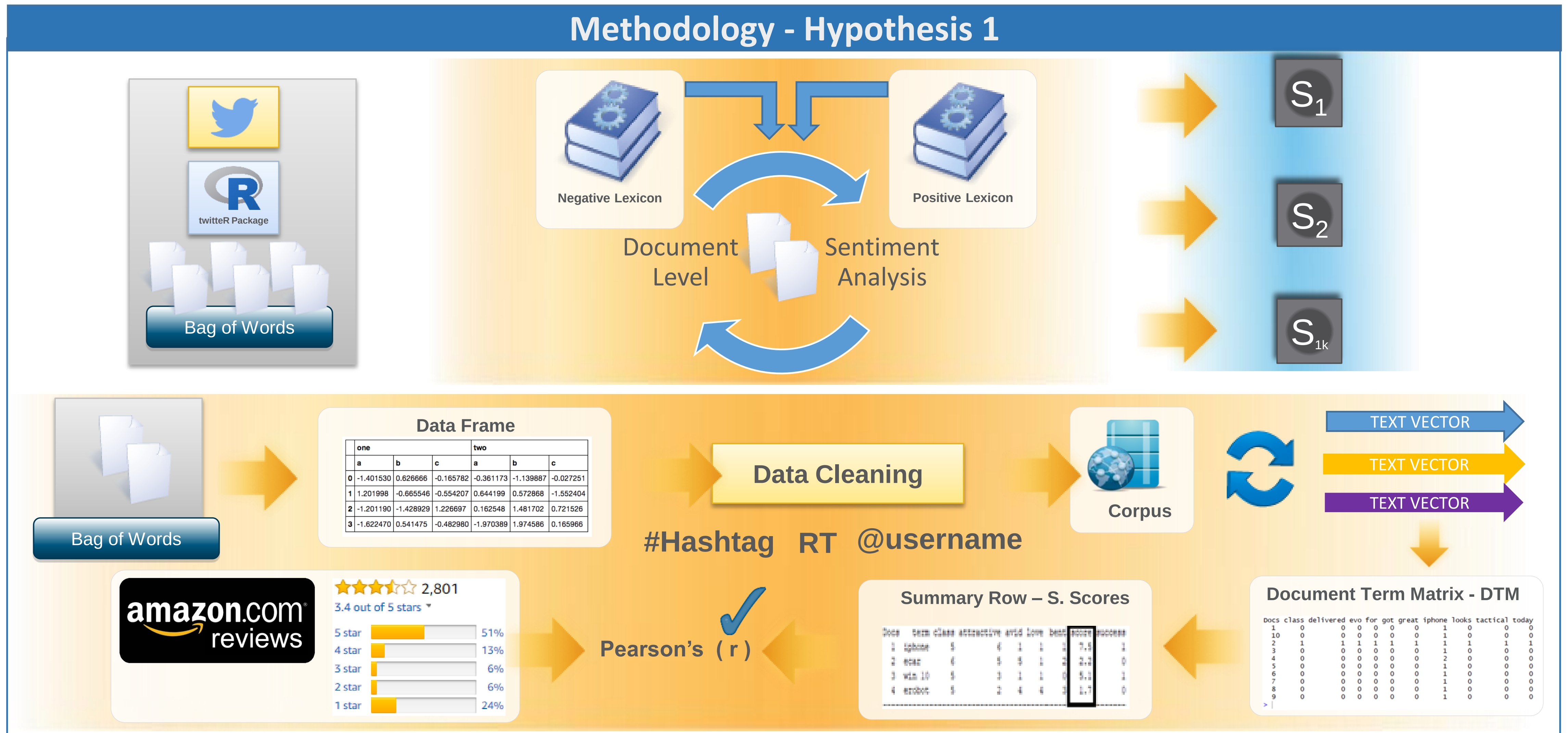
"Does aggregate social media sentiment agree with online user reviews of technology?"

Theory/ Hypotheses

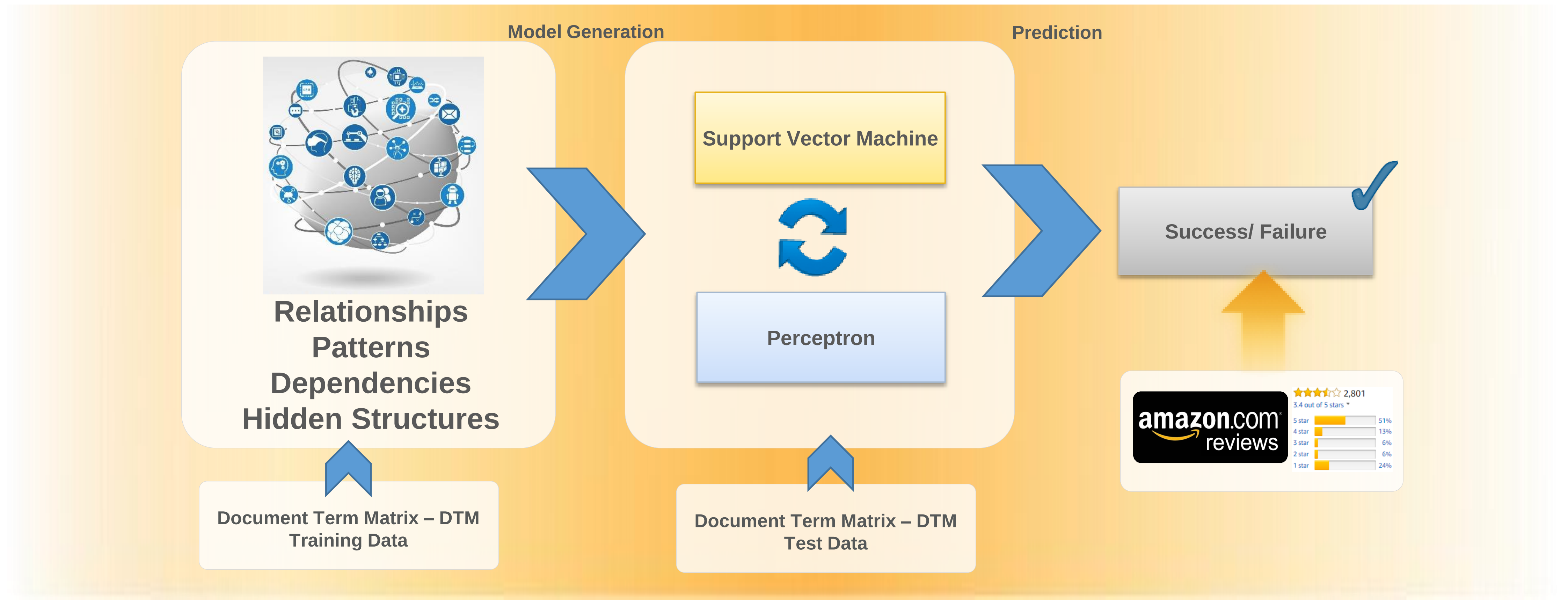
Hypothesis 1: Aggregate sentiment analysis score on social media will correlate to aggregate online user reviews of technology.

Hypothesis 2: Binary Classification could be used to classify social media sentiment as positive or negative in a way that corresponds with online user reviews of technology.

Methodology - Hypothesis 1



Methodology - Hypothesis 2



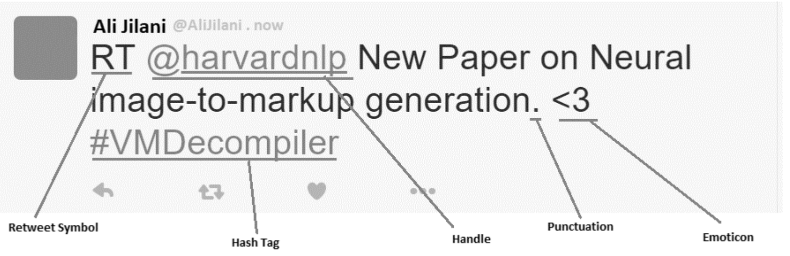
Data Gathering

Data is gathered using Twitter APIs for R imported as a package called twitteR. All data will be current and will be pulled lived at the time of study as time of day or day of week or even day of month have no effect on the quality of data.

Language filters will be applied to get only English results.

Keywords related to technology products e.g. iPhone, Nissan Leaf, etc.

Data Cleaning Activities


Ali Jilani @AliJilani . now
RT @harvardnlp New Paper on Neural image-to-markup generation. <3 #VMDecompiler

Harvested tweets will be cleaned after conversion to data frame since it makes is easier to manipulate data. Cleaning activities include removal of @handle labels, #Hashtags, Retweet Symbol (RT), URLs, Tiny URLs, stop words, Emoticons, All special characters other than English language letters, All extra spaces, Numbers and Punctuations along with other techniques like removal of stem words. All data is converted to lower case to gain homogeneity

Verification of Results

Hypothesis 1: Pearson Correlation Coefficient (r) will be used to measure the strength of a linear association between our aggregate sentiment scores and Amazon review star ratings on the other hand.

Hypothesis 2: The resulting prediction will be compared with the Amazon average star rating to calculate error which will be taken as the discrete difference between what the classifier is expected to predict and what it actually predicts.

Contact Information

Ali Jilani
MSc. In Computing, Enterprise Software Systems
Waterford Institute of Technology,
Department of Science & Computing
Email:20078735@mail.wit.ie

References

1. FELDMAN, R. 2013. Techniques and applications for sentiment analysis. *Commun. ACM,* 56, 82-89.
GIACHANOU, A. & CRESTANI, F. 2016. Like It or Not: A Survey of Twitter Sentiment Analysis Methods. *ACM Computing Surveys,* 49, 28-28:41.
2. STEPHEN, H. & REBECCA SCOTT (2017) 'Developing an Approach to Harvesting, Cleaning, and Analyzing Data from Twitter Using R', Information Systems Education Journal (ISEDJ), v15 n3 (May 2017), P42-54.

Acknowledgements

All my lecturers especially the names listed below.

1. Dr. Peter Carew
2. Dr. Bernard Butler
3. Dr. Kieran Murphy