International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2024: 7.101

Language Sentiment Analysis for Sunrail Transportation

Kunal Sonalkar[1], Shilpa Goel[2]

Department of CISE, University of Florida, Gainesville, FL, USA

Abstract: ***This paper presents a method to automatically extract sentiment (positive or negative) from a tweet along with the various challenges appear that in the field. Indeed, the task of automatic sentiment recognition in online text becomes more difficult for all the aforementioned reasons like limited size of character i.e. 140, unlimited spelling mistakes, slang words and different languages.***

Keywords: opinion, machine learning, tweets, semantic analysis, Naïve Bayes Classifier

1. Introduction

Sentiment Analysis involves determining the notion of a piece of text. For example, if we avail a service then our feedback can be positive, negative or neutral. Over the past decade there has been an exponential growth in the usage of social networking sites like Twitter, Facebook, Instagram, etc. Twitter is a popular micro-blogging service, with a rich source of information for sentiment analysis as the tweets sometimes express opinions and reviews about different topics. The sentiment of the tweets of a particular subject has multiple usages like including stock market analysis of a company, movie and product reviews, views in a political debate, in psychology to analyze the mindset of people that has a variety of applications, and so on. This project involves classification of tweets into neutral, positive and negative sentiments for a transportation company in Orlando. To classify broadly, a sentiment can be positive, negative, neutral, extremely positive, extremely negative, etc. Currently the data being used for training is being labeled by humans. While preprocessing we removed all the stop words from consideration and stemmed the words to create a proper feature vector.

A regular tweet has got its unique features like it is restricted to 140 characters, we can use RT feature which means retweet or the tweet was forwarded from a previous post; "@user" – implies that this message is a reply to the person with twitter handle as "user"; "#obama" - hashtag just tries to make a keyword (primary) in a tweet and embedding a link "http://bit.ly/9p "- redirects us to an external source.

Given the uniqueness of the tweets, they also present us with a new set of challenges. They are limited in length and usually span one sentence or even less. They might include incorrect spellings, colloquial phrases which might be difficult to interpret and random abbreviations. These also include hashtags which are used to facilitate search. Hence we need to make sure a certain level of sophistication while performing sentiment analysis.

Our present project performs mainly 2 tasks:
1. Predicting the sentiment of informal messages/tweets.
2. Predicting the sentiment of a word within the tweet.

One of our primary techniques which we have used is to pull tweets from the Twitter web API, comparing each word to positive and negative word bank, and then using a basic algorithm to determine the overall sentiment.

Twitter API - Data Source:
The REST APIs support short-lived connections and are rate-limited (one can download a certain amount of data but not more per day). REST APIs allow access to Twitter data such as status updates and user info regardless of time. However, Twitter does not make data older than a week or so available. Thus REST access is limited to data Twittered not before more than a week. Therefore, while REST API allows access to these accumulated data, Streaming API enables access to data as it is being Twittered.

We collected tweets manually using twitter's search api and a python tool named tweepy. Currently we have pulled over 1000 tweets over the period and labeled them manually. We use the consumer_key, consumer_secret, access_token, and access_token_secret to create the OAth access.

After getting the required permissions from twitter to access data, we add a search query as follows:
tweepy.Cursor(api.search, q=query).items(max_tweets)

The query used to search tweet used some keywords like: '#sunRail , ExpandSunRail , sunRail, Sunrailriders, RideSunRail ,#RideSunRail ,#Ridesunrail +exclude:retweets' where hashed words are considered a complete word as opposed to an individual word. With this query we are also avoid all the retweets as it populates the training data with repeated information. Keywords and hashtags were used to identify and collect messages relevant to the selected topic.

Along with the twitter data, the project also required other datasets like stopwords1, a dictionary of negative and positive words2, an emoticon dictionary3 and an acronym dictionary for twitter slang words4. The use of these are described in the next section.

A written text can be classified into a subjective text (a fact) or an opinionated text. We are interested in determining the opinionated text and classifying them into negative and positive sentiments.

Feature Extraction:

The report also describes the various preprocessing steps required to achieve high accuracies like the following:

1) Remove punctuations
2) Convert the sentence to a lower case
3) Remove stop words using a given stop word list
4) Stem words using Porter Stemmer algorithm.
5) Make all negated combinations a single feature word like “not happy”, “never complete”, etc using a hard coded list of negative words.
6) Normalizing elongations like arrrgggghhh -> arrgghh where repeated characters are placed just twice in a word.
7) Removing URLS and punctuations as after some analysis of data we found that their occurrence was quite less.
8) Removal of words which were not a noun, adjective, adverb or verb.

The resulting tweets after the above normalization techniques are shown as below:

old : I bicycle and take the sunrail & busses. It's not too bad tbh

https://t.co/Jq3YBldlOb new : bicycl sunrail amp buss 's bad honest

old : SunRail's $225M link to Orlando International Airport gains steam... https://t.co/85frL7qier https://t.co/YSIjSk6BmF no1

new : sunrail link orlando intern airport gain steam no1

old : @RideSunRail @OrlandoCitySC are we going to be able to take the sunrail to downtown to go to the citrus bowl for #FillTheBowlAgain ??

new : abl sunrail downtown citru bowl fillthebowlagain

URL:

People use twitter not only for expressing their opinions but also for sharing information with others. Given the short maximum length of tweets, one way of sharing is using links. Tweets include various links or URLs and these do not contribute to the sentiment of the tweet.

Stop words removal:

There are some words in a tweet like and, while, the, etc which are a part of every class during classification and do not contribute towards the sentiment of the tweet. These words are removed from the data so as to avoid using them as features. It will be seen later in the results that removing stop words from tweets would not make much difference in the accuracy as the tweets are already shortened due to which people tend to use less stop words in sentences.

Stemming:

Stemming is the process of reducing a word to its root form thereby reducing the feature space. Thus this helps in improving the accuracy of classification. As we will see in the results section, stemming gives a good increase in accuracy. By stemming, different derived words are mapped to their root words and this allows more matching between the tweets in the test and training set.

Repeated Letters:

Twitter contains a lot of texts with repeated characters as described above. In order to not lose information about such textual information, we use a preprocessing technique to make all such elongated duplications with more than 2 consecutive characters are reduced to a single letter. Thus, the number of independent words, denoting the same semantics, is reduced. For example, the sentence "I‟m soooooooo happyyyyyyyy!!!!!!" is more positive polar than the statement "I‟m so happy!".

Remove Nouns and Prepositions:

Given a tweet token, we identify the word as a noun word by looking at its part-of-speech tag assigned by the tokenizer. If the word is noun, we discard the word. Noun words do not carry sentiment and thus are of no use in our experiment. Similarly we remove prepositions too.

Remove Punctuations from Hashtags:

Hashtags represent a concise summary of the tweet, so it is useful to replace them with the exact same word without the hash. In order to capture the relevant information from hashtags, all special characters and punctuations are removed before using them as a feature. E.g. #nike replaced with 'nike'.

Remove everything that is not a string character:

In order to refine the tweet content a bit more, we remove all the possible characters that are not strings like numbers, etc as they do not carry any sentiment value with them.

2. Related Work

There is a lot of research work going on in this area. A very broad view of the existing work was presented by Alexander Pak and Patrick Paroubek. In their paper they discuss about collecting a corpus and training it using a sentiment classifier which was able to identify the positive, negative and neutral sentiment of a tweet. According to them Naive Bayes classifier gave better results as compared to SVM classifier. The best performance was achieved when a bigram was used.

Sentiment analysis on twitter data has been done previously by Shachi H Kumar, where she analysed on a bigger dataset that SVM had a better accuracy as compared to Naive Bayes approach. On the features, they have used Unigram, Bigram, along with stemming. They also perform some pre-processing of the data that was used in modeling the pre-processing techniques used in this project. The text processing they perform includes filtering out URLs, username references and repeated characters in words and stopwords.

Miles Osborne, Ting Wang, Zhunchen Luo proposed a standard machine learning approach to learn a ranking function for tweets that uses the available social features and opinionated feature. They constructed an opinionated lexica from sets of tweets matching specific patterns indicative of opinionated

messages. According to them, if a user is listed many times, it means that his tweets are interesting to a larger user population. They used a feature that measures how many times the author of a tweet has been listed for tweet ranking.

Few people have also devised a sentiment analysis technique called "Manual Annotation Scheme" where contextual polarity judgments are added to existing annotations in the (Multi perspective Question Answering (MPQA) Opinion Corpus which is available and is described at nrrc.mitre.org/NRRC/publications.htm) namely to the annotations of subjective expressions. A subjective expression is any word or phrase used to express an opinion, emotion, evaluation, stance, speculation. For this work, major focus is on sentiment expressions– positive and negative expressions of emotions, evaluations, and stances. As these are types of subjective expressions, to create the corpus, we just needed to manually annotate the existing subjective expressions with their contextual polarity.

In particular, an annotation scheme was developed for marking the contextual polarity of subjective expressions. Annotators were instructed to tag the polarity of subjective expressions as positive, negative, both, or neutral. The positive tag is for positive emotions (I'm happy), evaluations (Great idea!), and stances (She supports the bill). The negative tag is for negative emotions (I'm sad), evaluations (Bad idea!), and stances (She's against the bill). The both tag is applied to sentiment expressions that have both positive and negative polarity. The neutral tag is used for all other subjective expressions: those that express a different type of subjectivity such as speculation, and those that do not have positive or negative polarity.

The annotators were asked to judge the contextual polarity of the sentiment that is ultimately being conveyed by the subjective expression, i.e., once the sentence has been fully interpreted.

3. Our Approach

Natural Language Processing/Symbolic Technique (Unsupervised):

Much of the research in unsupervised sentiment classification using symbolic techniques makes use of available lexical resources. In "bag-of-words" approach, relationships between the individual words are not considered and a document is represented as a mere collection of words.

Machine Learning Approach - Supervised Learning:

A number of machine learning techniques have been adopted to classify the reviews. Machine learning techniques like Naive Bayes (NB), maximum entropy (ME), and support vector machines (SVM) have achieved great success in text categorization. The other most well known machine learning methods in the natural language processing area are K-Nearest neighbourhood, ID3, C5, centroid classifier, winnow classifier, and the N-gram model.

The primary issues in unsupervised techniques are classification accuracy, data sparsity and sarcasm, as they incorrectly classify most of the tweets with a very high percentage of tweets incorrectly classified as neutral. Majority of the researches employ Support Vector Machines or Naive Bayes classifiers because they usually obtain the best results. These classifiers perform very well in the domain that they are trained on, but their performance drops when the same classifier is used in a different domain.

Naive Bayes Classifier:

The Naïve Bayes (NB) classifier is based on Bayes rule, a practical Bayesian learning model that is easy to understand and implement. The Bayes rule allows us to determine this probability of any event. It is the probabilistic approach to the text classification. Here the class labels are known and the goal is to create probabilistic models, which can be used to classify new texts. It is specifically formulated for text and makes use of text specific characteristics. The NB classifier is based on the assumption that all the attribute values are conditionally independent given the target value of the instance.

Most Informative Features

contains(late) = True negati : neutra = 27.2 : 1.0 contains(sigh) = True negati : neutra = 14.7 : 1.0 contains(increas) = True positi : eutra = 11.2 : 1.0
contains(plan) = True negati : neutra = 10.5 : 1.0 contains(huh) = True negati : neutra = 10.5 : 1.0 contains(look) = True negati : neutra = 10.5 : 1.0 contains(remind) = True negati : neutra = 10.5 : 1.0 contains(apolog) = True neutra : positi = 10.2 : 1.0 contains(continu) = True positi : neutra = 9.1 : 1.0
contains(run) = True negati : positi = 8.0 : 1.0 contains(favorit) = True positi : neutra = 7.1 : 1.0 contains(due) = True neutra : positi = 6.8 : 1.0
contains(delay) = True neutra : positi = 6.6 : 1.0 contains(time) = True negati : neutra = 6.3 : 1.0 contains(stuck) = True negati : neutra = 6.3 : 1.0 contains(freez) = True negati : neutra = 6.3 : 1.0 contains(ave) = True negati : neutra = 6.3 : 1.0 contains(worst) = True negati : neutra = 6.3 : 1.0
contains(tap) = True negati : neutra = 6.3 : 1.0 contains(inform) = True negati : neutra = 6.3 : 1.0 contains(huge) = True negati : neutra = 6.3 : 1.0 contains(stand) = True negati : neutra = 6.3 : 1.0 contains(scream) = True negati : neutra = 6.3 : 1.0 contains(commut) = True negati : neutra = 6.3 : 1.0
contains(smoothest) = True negati : neutra = 6.3 : 1.0
contains(serious) = True negati : neutra = 6.3 : 1.0
contains(shut) = True negati : neutra = 6.3 : 1.0 contains(dethron) = True negati : neutra = 6.3 : 1.0
contains(bound) = True negati : neutra = 6.3 : 1.0 contains(debari) = True neutra : positi = 6.0 : 1.0 contains(amp) = True positi : neutra = 5.7 : 1.0 contains(nice) = True positi : neutra = 5.7 : 1.0

Limitations of Naive Bayes Approach:

Some of the tweets were misclassified as shown below:
Tweet → On @RideSunRail to @DrPhillipsCtr for Dirty Dancing with @itsbrick5!!□□□□❤❤ https://t.co/lCuZcnTVuB
Reason -> This was classified as neutral tweet, although with the emoticons it's a positive tweet. This happened because we

are not taking the emoticons into consideration.

Tweet → Liking the upper deck #SunRail @RideSunRail https://t.co/ygWRtswVzq This was classified as a neutral tweet although with the word "liking", it clearly is a positive tweet.

Jeffrey Breens Algorithm for Sentiment Analysis

It shows how to pull tweets from the Twitter web API, comparing each word to positive and negative word bank, and then using a basic algorithm to determine the overall sentiment.

We considered primary Keywords like "abortion", "#Genocide", etc to get a large number of tweets and then applied the matching algorithm. This technique tries to classify each word from the tweet into either a positive or a negative element. The major drawback of this technique is that it ignores the contextual relevance of the words. Therefore the scores assigned to the tweets may not be as accurate.

The following snapshot shows the result:

29	-4	As the President mourns the tragic lost children in #Newtown, will he also grieve those who have lost their lives by way of #abortion?
30	-1	2012 Victories: State Hospitals Can't Force Nurses to Abort Children http://t.co/gYv8lf via @ACLJ #prolife #prochoice #abortion
31	1	#guncontrol feels like my generation's #abortion. Starting to think the two sides will argue about this for my lifetime.
32	-1	James Dobson: God 'Has Allowed Judgment To Fall Upon Us' For #Gay Marriage, #Abortion http://t.co/o9lOlOsv #religion #LGBT
33	0	RT @LiveActionFilms: Yet another in the opposition to #abortion-loving #ObamaCare! Business owners would like to keep their consciences. http://t.co/TPQD29ou
34	0	'W/' the way Obama was acting talking about Sandy, you'd think he would be crying outside #Abortion clinics #war #tcot #UNYHBT
35	0	RT @Vote4Wallace: #Prenatal Testing: A Double-Edged Sword Leading to #Abortion [then #Genocide] http://t.co/tUL2dvwH #Life #Prolife #ObamaCare #Medical #HHS
36	-1	@ShantaCovington where do Dobson and others get this stuff? What in the world do those babies have to do with #ssm and #abortion? #ridiculous
37	1	"They Want Us to Be Silent" ~ Luke Robinson http://t.co/mRNRrhuK #Abortion #NewYorkGenocide #ProLife
38	-1	RT @jcbyoung: An estimated 1.2million babies were aborted in the #US in 2011. Is it finally time for them to review their #abortion law? #america #prolife
39	-1	An estimated 1.2million babies were aborted in the #US in 2011. Is it finally time for them to review their #abortion law? #america #prolife
40	-1	@GOPLeader Perhaps you can get Congress to have a moment of silence at the end of the year for all the little children lost to #abortion?
41	0	Yet another in the opposition to #abortion-loving #ObamaCare! Business owners would like to keep their consciences. http://t.co/TPQD29ou
42	-2	@LtRedX no child died in the name of the second amendment but in the name of #murder, akin to #abortion.

It pulls the tweets from twitter and assigns a particular score to every tweet (shown in second column). Positive/Negative scores indicate the positive/negative emotion. These scores are according to the match found in the positive/negative word bank.

```
pos.matches = match(words, pos.words)
neg.matches = match(words, neg.words)
score = sum(pos.matches) - sum(neg.matches)
```

The following command assigns the scores to text sentences and we store the scores and tweets in a csv file.

scores.df = data.frame(score=scores, text=sentences)

Support Vector Machine:

Support vector machines represent an extension to nonlinear models of the generalized portrait algorithm developed by Vladimir Vapnik. The SVM algorithm is based on the statistical learning theory and the Vapnik Chervonenkis (VC) dimension introduced by Vladimir Vapnik and Alexey Chervonenkis. Support vector machines (SVM) are a group of supervised learning methods that can be applied to classification or regression. This classification creates a hyper-plane or a set of hyperplanes in a high-dimensional space such that the separation is maximum.

In a three-classed classification (as we have here), there will be three pairwise classification. That is positive-negative, negative-neutral, and positive - neutral. Support vector machines and the maximum hyperplane that divides the training space into classes as far apart from each other as possible. SVM is computationally expensive since it involves all the discretization, normalization and repetitive dot products operations.

In this study we use the machine learning approach due to the promising results obtained in previous works. To build our model we employed Support Vector Machines (SVM) as the supervised machine learning algorithm, as it has been proved to be effective on text categorization tasks and robust on large feature spaces.

Results on applying SVM:
Gets a hit ratio of 71% -

Used RBF kernel with gamma = 0 and a penalty parameter: 1000000

The test data is a part of the training data(20%) which was broken down using SCIKIT learn python tool. In order to obtain the feature vector, we have used TF IDF vectorizer. Currently keeping tweets in the training dataset after removing stop words and words which do not belong to the set – Noun, adverb, adjective or verb.

Also, tried to apply Naïve Bayes and Max Entropy Classifier, but didn't obtain good results as most of the tweets were classified as neutral.

References

[1] Twitter opinion mining framework using hybrid classification scheme.
[2] Twitter as a Corpus for Sentiment Analysis and Opinion Mining by Alexander Pak, Patrick Paroubek
[3] Twitter Sentiment Analysis by Shachi H Kumar
[4] https://jeffreybreen.wordpress.com/2011/07/04/twitter-text-mining-r-slides/
[5] https://jeffreybreen.wordpress.com/tag/sentiment-analysis/
[6] http://allthingsr.blogspot.com/2012/01/updated-

sentiment-analysis-and-word.html
[7] https://sites.google.com/site/miningtwitter/questions/sentiment/analysis
[8] http://thinktostart.com/sentiment-analysis-on-twitter/
[9] http://andybromberg.com/sentiment-analysis/
[10] http://www.youtube.com/watch?v=adIvt_luO1o
[11] http://jeffreybreen.wordpress.com/2011/07/04/twitter-text-mining-r-slides/
[12] IndiSent Analysis in Twitter using Machine Learning Methods by Neelima and Dr. Ela Kumar.
[13] Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis By Theresa Wilson, Janyce Wiebe, Paul Hoffmann.
[14] Opinion Retrieval in Twitter by Zhunchen Luo, Miles Osborne,Ting Wang