PUNE INSTITUTE OF COMPUTER TECHNOLOGY,
DHANKAWADI PUNE-43.

A Seminar Report

On

Twitter Sentiment Analysis on Tourism using Lexicon Based Approach

SUBMITTED BY

NAME: Vikas Kodag
ROLL NO: 3373
CLASS: TE-3

GUIDED BY
PROF. M. S. Takalikar

COMPUTER ENGINEERING DEPARTMENT

Academic Year: 2017-18

PUNE INSTITUTE OF COMPUTER TECHNOLOGY,
DHANKAWADI PUNE-43.

CERTIFICATE

This is to certify that Mr. ***Vikas Kodag*** , *R*oll No**. 3373** a student of T.E. (Computer Engineering Department) Batch 2017-2018, has satisfactorily completed a seminar report on " **Twitter Sentiment Analysis on Tourism using Lexicon Based Approach**" under the guidance of Prof. M. S. Takalikar towards the partial fulfillment of the third year Computer Engineering Semester II of Pune University.

Prof. M. S. Takalikar	Dr. R.B.Ingle
Internal Guide	**Head of Department,**
	Computer Engineering

Date:

Place:

Twitter Sentiment Analysis on Tourism Using Lexicon Based Approach

Contents

List of Figures

List of Tables

Abstract:

Sentimental Analysis is reference to the task of Natural Language Processing to determine whether a text contains subjective information and what information it expresses i.e., whether the attitude behind the text is positive, negative or neutral. It is also known as emotion extraction or opinion mining. This is a very popular field of research in text mining. The basic idea is to find the polarity of the text and classify it into positive, negative or neutral. It helps in human decision making. To perform sentiment analysis, one has to perform various tasks like subjectivity detection, sentiment classification, aspect term extraction, feature extraction etc.

Keywords*: Sentiment Analysis, Opinion Mining, Social Networks, Classifiers, Supervised learning, Unsupervised learning.*

1 INTRODUCTION

Sentiment Analysis (also referred as opinion mining) is the study of affective states and subjective information in the customer data (such as reviews and survey responses, online and social media) by using natural language processing and data mining techniques [1]. Sentiment analysis aims to determine the attitude of a subject with respect to some topic or the overall contextual polarity or emotional reaction to some object, such as a document, interaction, or event. The attitude may be a judgment or evaluation, affective state, or the intended emotional communication.

For opinion mining or sentiment analysis some methods are applied like – Naive Bayes Machine Learning Classifier, Sentiwordnet, Support Vector Machine. Here we have used Lexicon based approach of Sentiment Analysis.Sentiment lexicon is used in the lexicon based approach. Sentiment lexicon is a collection of known and defined words. A specific sentiment is assigned to each word in the collection. The lexicon based approach is divided into dictionary based approach and corpus based approach[2].

Sentiment analysis task is divided into three categories; Aspect level, Sentence level, Document level [3]. Aspect level analysis deal with the aspects of items. It can also be considered as phrase level analysis. In Sentence level, each sentence is considered as an entity. Summation method is used to provide overall result of the document. In document level, the whole document is considered as a single entity.

1.1 Motivation

Twitter Sentiment Analysis was thoroughly dealt by Alec Go, Richa Bhayani and Lei Huang, Computer Science graduate students of Stanford University. They used various classifiers, including Naive Bayes, Maximum Entropy as well as Support Vector Machines to classify the tweets. The feature extractors used by them were both unigrams and bigrams combined. Parts of speech tag was used because same word may have different meaning depending on its usage. The data-set used by them was huge, comprising 1.6 million tweets divided equally into positive and negative classes.

We have chosen to work with twitter since we feel it is a better approximation of public sentiment as opposed to conventional internet articles and web blogs. The reason is that the amount of relevant data is much larger for twitter, as compared to traditional blogging sites. Moreover the response on twitter is more prompt and also more general (since the number of users who tweet is substantially more than those who write web blogs on a daily basis). Sentiment analysis of public is highly critical in macro-scale phenomena like predicting the needs of tourist and their opinions on the tourism spot. This could be done by analysing overall public sentiment towards the place with respect to time for finding the correlation between public sentiment and the place of interest. The government can also estimate the changes to be made, facilities to be provided to attract more tourists in the future and in which a negative response was registered since twitter allows us to download stream of geo-tagged tweets for particular locations. Other applications of Sentiment Analysis includes the review of movies and products, popularity of an event. Predicting the results of popular political elections and polls is also an emerging application to sentiment analysis. One such study was conducted by Tumasjan et al. in Germany for predicting the outcome of federal elections in which concluded that twitter is a good reflection of offline sentiment

1.2 Literature Survey:

Sentiment analysis has been studied in wide area of domain such as movie review, teaching review [4], product review, e-learning, hotel review and many more. A small number of studies have focused on applying machine learning techniques in the tourism sector.

A study [5] aimed to create a system that would assist users in understanding tourism opinions on the web by finding and extracting subjective information from reviews in tourism websites. Aspect extraction was performed with the use of frequent nouns and the opinion was determined.

Estela Marine-Roig et al.[6] addressed the problem of finding out the frequently occurring trends of different tourist places from tourist opinions. The authors proposed a trends extraction framework that consisted of five phases i.e. semi automatic downloading, arranging, cleaning, debugging, and analyzing. Trends extraction framework is better than previous method Liu (2011) in trends extraction because two extra phases of cleaning and debugging has been added up to eliminate the noise present in the tourist's opinions. The limitations of the work are that i. Method does not classify the derived frequent trends into positive and negative trends ii. method extracts same trend in one opinion sentence multiple times that create the reputation of trends iii. Method extracts many irrelevant and meaningless trends during classification.

In another way to enhance the performance of opinion sentences extraction Shimada, K.. [7] used support vector machine for sentences classification. The authors addressed the problem to identify whether tweet on-site are more likelihood or tweet off-site. The authors proposed a method to evaluate on-site likelihood. Firstly, this method takes tweets and identifies tourism related tweets. Secondly, extracts tourism related tweets and deletes the remaining ones. Lastly classifies the extracted tweets on the basis of different features of tourist places using SVM. The finding of this paper is that classification has improved by applying the method of on-site likelihood filtering method. The same fact is shown in the results i.e. without applying this filtering method Recall=58.2% and Precision = 75.0% and after applying the filtering method Recall=65.0% and Precision=80.5% . If there is a location name at the start of any tweet then it is high onsite likelihood tweet. The limitations of the work are that i. mostly the comments of authors on tweets are more than any other person which are mostly positive or negative that create nose in sentiment analysis ii. method extracts some sentences in which no opinion about targeted tourist place is given that creates noise during classification of reviews.

1.3 Applications

1.3.1 Support in decision making:

Decision making is a very important field of our life. Opinions extracted from reviews helps us in making various decisions like “which books to buy” , “which hotel to go” , “which movie to watch” etc.

1.3.2 Business application:

In today’s world of competition, every company wants to satisfy its customers requirements by creating new innovative products. Assessments of individuals are an essential angle today with the goal that organizations can get an input from clients and can roll out sought improvements in their item. Google Product Search is one illustration.

1.3.3 Predictions and trend analysis:

Sentiment analysis enables one to predict market trends by tracking views of public. It is also helpful in elections where candidates wants to know the expectations of people from them. It is also used to follow the events that are trending right now.

2 PROPOSED MATHEMATICAL MODEL

Let S be the set which defines the system.

S = { s, e, X, Y, Fm, DD, NDD, Su, Fl}

s = Start State

e = End State

X = Input Set = { Manually annotated sentiments, Input sentence}

Y = Output Set = { Polarity of the input sentence}

Fm = Set of functions = { F1, F2, F3}

Where,

F1() = { Tokenizing the input sentences into words}

F2() = { classifies sentence as subjective and objective}

F3() = { Determines the polarity of the sentence along with polarity value and normalized value}

DD = Deterministic Data

= { X}

NDD = Non-Deterministic Data

= { Y}

Su= Success case ={ Sentiment got classified according to the correct polarity }

Fl=Failure case={ Polarity of the sentence is reversed }

3 DESIGN AND ANALYSIS OF SYSTEM

3.1 Process Pipeline

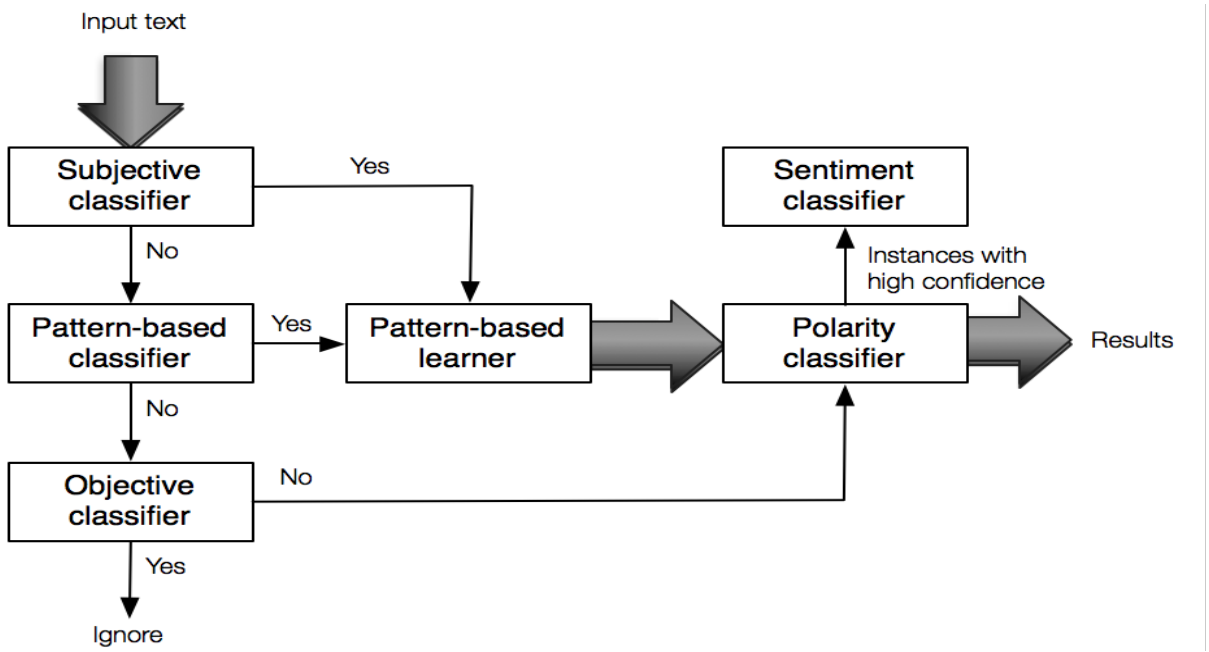

Figure 1: Process Pipeline

The current pipeline that is implemented in sentiment.py is depicted in the following diagram. Initially, the input text is split into sentences and each sentence is fed to a high precision subjectivity classifier. If the sentence is classified as subjective then syntactic patterns are learned from this instance. In case that the sentence is not detected as such then it is fed to the pattern-based classifier. The pattern-based classifier outputs the class of the sentence based on the learned patterns so far. If the instance is subjective then again more patterns are learned from it, otherwise it is fed to a high precision objectivity classifier. If the sentence is classified as objective, then it is ignored, otherwise it is fed to the polarity classifier. Finally, the polarity classifier estimates the numerical sentiment and normalized sentiment values and outputs the result.

3.2 Subjective Classifier

The high precision subjectivity classifiers are used to classify the sentences as “Subjective” . The high-precision classifiers use lists of lexical items. Many of the subjective clues are from manually developed resources. The subjectivity clues are divided into those that are strongly subjective and those that are weakly subjective, using a combination of manual review and empirical results. A strongly subjective clue is one that is seldom used without a subjective

meaning, whereas a weakly subjective clue is one that commonly has both subjective and objective uses.

The high-precision subjective classifier classifies a sentence as subjective if it contains two or more of the strongly subjective clues.

3.3 Objective Classifier

The high-precision objective classifier takes a different approach. Rather than looking for the presence of lexical items, it looks for their absence. It classifies a sentence as objective if there are no strongly subjective clues and at most one weakly subjective clue in the current, previous, and next sentence combined.

The similar approach for objectivity classification is not taken as the sentences containing objective clues does not readily lead to high precision objective classification. Add sarcasm or a negative evaluation to a sentence about a dry topic such as stock prices, and the sentence becomes subjective. Conversely, add objective topics to a sentence containing two strongly subjective words such as odious and scumbag, and the sentence remains subjective.

The bootstrapping process for subjective and objective classifier is given below

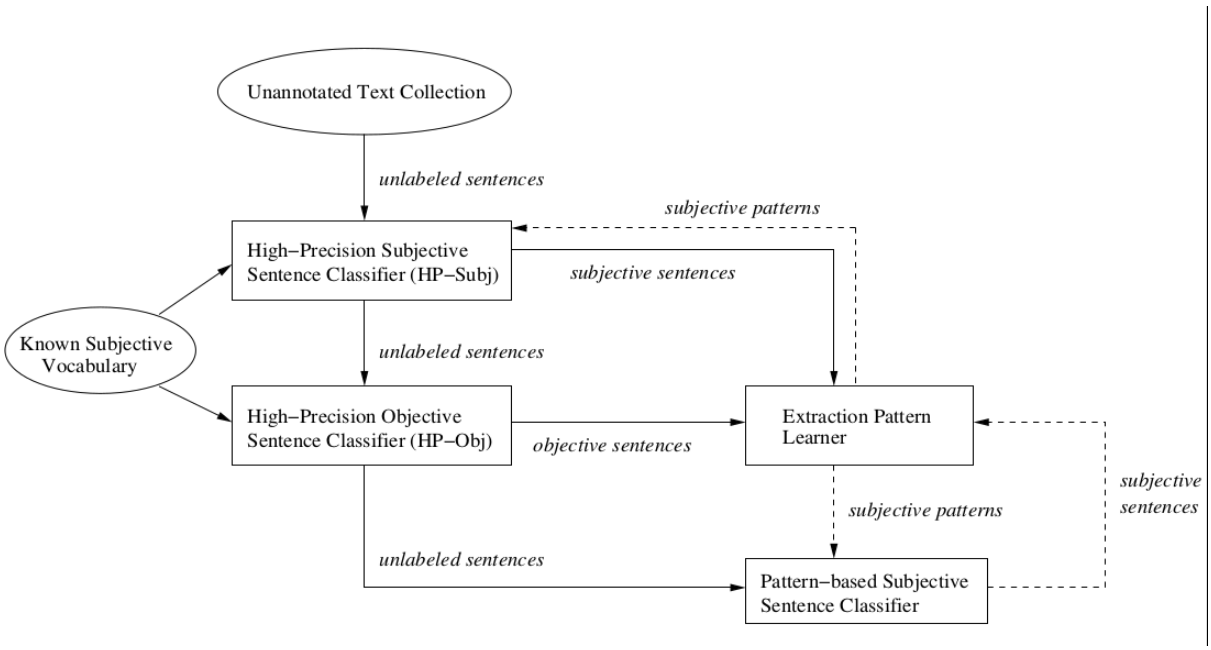

Figure 2: Bootstrapping Process

3.4 Polarity Classifier

A two-step approach was adopted to contextual polarity disambiguation. For the first step, we concentrate on whether clue instances are neutral or polar in context (where polar in context refers to having a contextual polarity that is positive, negative or both). For the second step, we take all clue instances marked as polar in step one, and focus on identifying their contextual polarity.

The three polarity features(modifies polarity, modified by polarity, conj polarity) look in a window of four words before, searching for the presence of particular types of polarity influencers. General polarity shifters reverse polarity (e.g., little truth, little threat). Negative polarity shifters typically make the polarity of an expression negative (e.g., lack of understanding). Positive polarity shifters typically make the polarity of an expression positive (e.g., abate the damage).

4 DISCUSSION ON IMPLEMENTATION RESULTS

The proposed system was applied to the tweets gathered from twitter account of "GujratTourism" . The final result are displayed in Table 1.

Positive Sentiments	681
Negative Sentiments	50
Neutral Sentiments	259

Table 1: Results from Twitter Sentiment Analysis

The system was also tested against random dataset comprising of 420 manually annotated sentences. The results are displayed in Table 2.

	Manual Results	Analysis Results
Positive Sentiments	152	199
Negative Sentiments	202	205
Neutral Sentiments	66	16

Table 2: Sentiment Analysis on Random Sentences

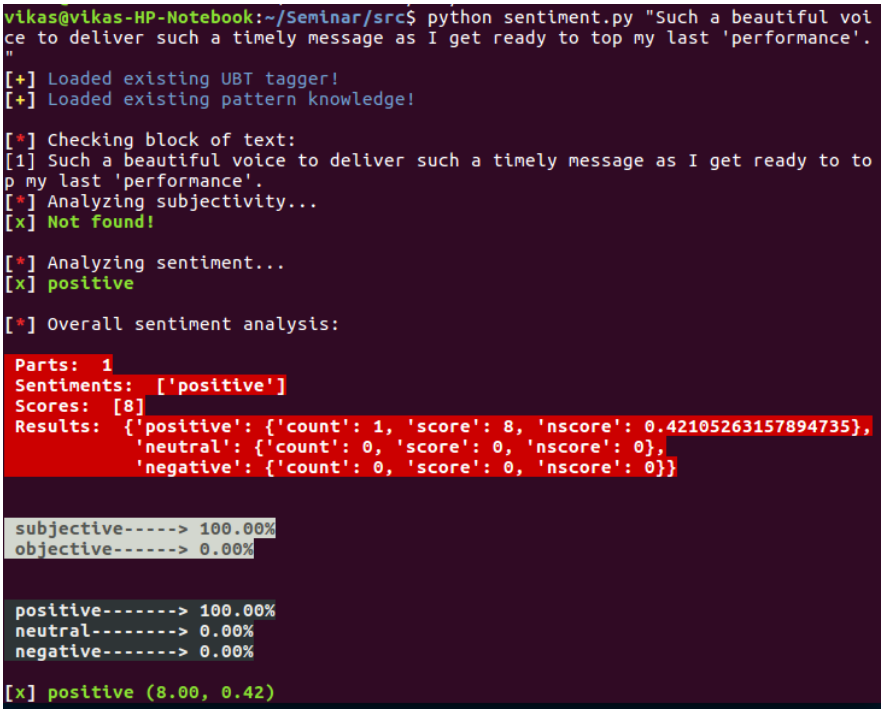

Figure 3: Snapshot of working(Positive Sentiment)

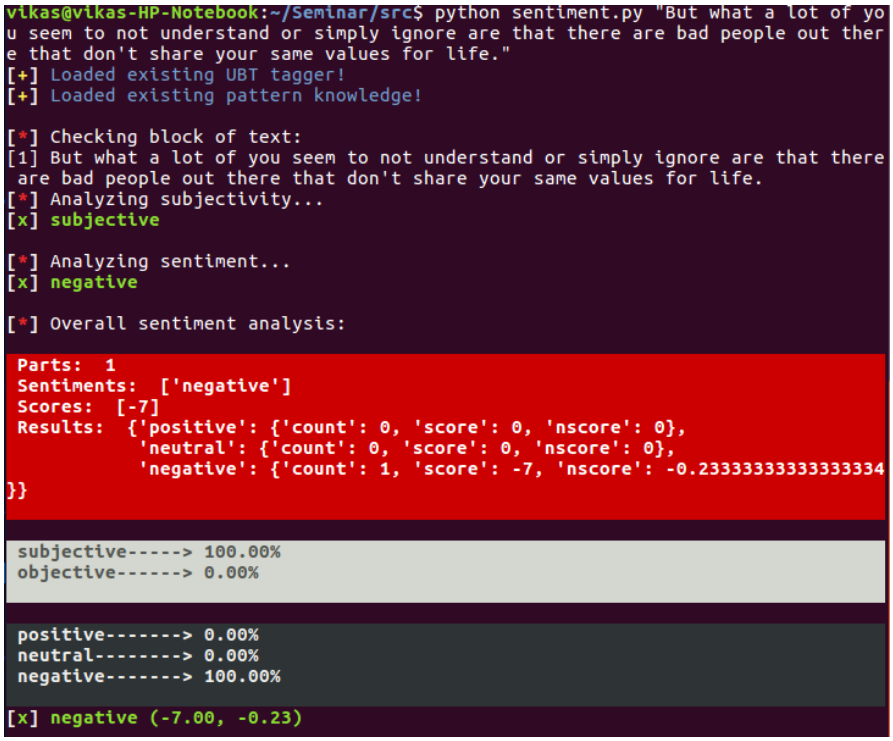

Figure 4: Snapshot of working(Negative Sentiment)

5 CONCLUSION AND FUTURE ENHANCEMENT

We studied the tweets from twitter account of “GujaratTourism”. The lexicon-based approach was used to to build an extensive sentiment lexicon. The high precision subjectivity classifiers and polarity classifiers were used to generate the results. The overall accuracy that system was able to achieve was 63% . The system showed higher accuracy in classifying negative sentiments which was dropped while classifying positive sentiments.

As a future work, we have identified multiple works that can be carried out. Firstly, fine-grained level methods can be adopted to identify whether there is a polarity towards search term or not. Secondly, WSD methods can be implemented to increase the correctness of SentiWordNet analysis since WSD handles identification of word context problem. Thirdly, sarcasm detection can be applied as addressed in experiments and results section to improve the accuracy of results. Lastly, machine learning methods may be applied to understand the accuracy of them with respect to lexicon-based methods.

References

[1] Marrese-Taylor, E., Velasquez, J. D., & Bravo-Marquez, F. (2013). Opinion Zoom: A Modular Tool to Explore Tourism Opinions on the Web (pp. 261–264). IEEE. doi:10.1109/WI-IAT.2013.193

[2] W. Medhat, A. Hassan and H. Korashy, "Sentiment analysis algorithms and applications: A survey," Elsevier, Ain Shams Engineering Journal, vol. 5, Issue 4, pp. 1093-1113, December 2014.

[3] R. Xia, F. Xu, C. Zong, Q. Li, Y. Qi, and T. Li, "Dual Sentiment Analysis: Considering Two Sides of One Review," IEEE Transactions on Knowledge and Data Engineering, vol. 27, Issue 8, pp. 2120-2133, August 2015.

[4] A. El-Halees, "Mining opinions in user- generated contents to improve course evaluation," Software Engineering and Computer Systems, pp. 107-115, 2011.

[5] Marrese-Taylor, E., Velasquez, J. D., & Bravo-Marquez, F. (2013). Opinion Zoom: A Modular Tool to Explore Tourism Opinions on the Web (pp. 261–264). IEEE. doi:10.1109/WI-IAT.2013.193

[6] Colhon, M, Badica, C, & Sendre, A (2014). Relating the Opinion Holder and the Review Accuracy in Sentiment Analysis of Tourist Reviews. In Knowledge Science, Engineering and Management (pp.246-257). Springer International Publishing.

[7] Himada, K., Inoue, S., & Endo, T (2012, September). On-site likelihood identification of tweets for tourism infonnation analysis. In Advanced Applied lnformatics (11A1AA1), 2012 IIAl International Conference on (pp. 117-122). IEEE.

[8] H. Kaur, V. Mangat and Nidhi, "A survey of sentiment analysis techniques," 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC) , Palladam, 2017, pp. 921-925. Doi: 10.1109/I-SMAC.2017.8058315

[9] B. K. Bhavitha, A. P. Rodrigues and N. N. Chiplunkar, "Comparative study of machine learning Techniques In sentimental analysis," 2017 International Conference On Inventive Communication and Computational Technologies (ICICCT), Coimbatore, 2017, pp. 216-221. Doi: 10.1109/ICICCT.2017.797

[10] https://www.safaribooksonline.com/library/view/natural-language-annotation/9781449332693/ch01.html