Application of Subjectivity and Sentiment Score in Sentiment Analysis

Dr. Parul Verma
Amity Institute of Information Technology
Amity University Uttar Pradesh, Lucknow
pverma1@lko.amity.edu

Abstract

Sentiment Analysis is done to assess the sentiments of the end users behind their views that are posted on any public domain. The researchers are working hard from long time to assess the notion of the views automatically through machine. Many researchers worked and given solutions for the accurate assessment of the sentiments of the end users. One aspect of sentiments is subjectivity which refers to the evaluation of expression, feelings and speculation of the end users. We must identify whether a text is subjective or objective during sentiment analysis, on the basis of it we can decide the polarity of the text. The paper is addressing the classification of sentiments using subjectivity and polarity. The two popular classification models are used to classify the sentiments and their accuracy is evaluated- Multinomial Naïve Bayes and Decision Tree. The accuracy of the sentiment classification is evaluated. It is observed that out of the two classifiers Multinomial Naïve Bayes outperformed the two.

Keywords- Sentiment Analysis, Subjectivity, Sentiment Score,Multinomial Naïve Bayes, Decision Tree

Introduction

In current scenario the need and requirement of Sentiment Analysis is growing very fast. Opinion mining is another name of Sentiment Analysis. The basic purpose of sentiment analysis is to process and analyze the text content posted on the social media. People now days are ready to put their views and thoughts on the various social media platforms. Their views may per personal or some reviews on the products they use. Sentiment Analysis helps the organizations to assess the customer satisfaction with respect to their products. Though the focus of sentiment analysis is to identify the sentiment in terms of positive, negative or neutral called as polarity of the sentiment. However, an extension of it to find out the exact sentiment of the text as unhappy, happy, angry, joyful etc.

Aspect based sentiment analysis is also used for the analysis of the feedback of the customer that relates the feedback with the emotions of the customers. Subjective sentiment refers to the meaning or sound of a particular text in the context to artificial intelligence and natural language processing.Sentiment is subjective so we do not have an objective way of evaluating sentiment inherently.All current techniques are defined by human input.Analysis of sentiment facilitates the design of a method to gather and evaluate the emotional tone behind words.This is important because it helps users to gain an understanding of the individuals' attitudes, opinions and emotions in the data. Using TextBlob method this paper demonstrates the subjectivity detection for the analysisofsentiment. Words are used as bigram.

Literature Review

Sentiment analysis is the process of extraction and classification of sentiment by using NLP techniques, text analytics and various computational techniques. The sentiment analysis is playing key role in various domain where it is giving ideas to the business organizations to work on their strategic planning with respect to their products based on the analytics performed on the sentiment classification. The work of few researchers that worked in the area of Sentiment Analysis is discussed here-

TheMohsenGhorbaniet. al.utilized neural networks that is very popular and robust machine learning models for their work for sentiment analysis. They used Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM) networks to identify the polarity of sentiments on a dataset of Movie reviews. They used Deep Learning algorithm with Word Embeddings in their work and got highly accurate results [1].

Wei Zhanget. al. in their work they classified sentiments into happiness, hope, disgust and anxiety. They used a novel approach that is purely based on the conjunction of the support vector machine and latent semantic analysis. Their experimental work given positive outcome for fine grained computing of sentiments on online reviews. They concluded that their work helped the company in understanding customer sentiments and improving their branding.

DoaaMohey El-Din emphasized on negative polarity sentiments on online medium. According to him more than 60% of the sentiment on online platform face the challenge of negative polarity. They introduced five levels for classification of negative polarity sentiments. They outperformed with the average accuracy for all level of negative sentiments with 87.4%. For negative precision their technique indicates 10% increase in performance. [3].

A corpus-based approach to reverse the analysis by constructing a dictionary of pseudo-antonyms.Test review is classified by two sides of review through dual prediction. The researchers developed an algorithm that totally depends upon dictionaries, the algorithm is an unsupervised algorithm, and it can manage issue of polarity change. They use Restaurant review dataset for the same.[4]

Sandeep Nigamet. al. demonstrated their work on sentiment analysis on Sentiment140 dataset. They applied various machine learning models on the

mentioned dataset and observed that logistic regression outperforms as compared to the rest of the machine learning models[5].

The researchers implemented different neural network models for sentiment analysis and classification as well. The good performance is observed by the researchers with respect to sentiment classification using deep learning models[6].

The authors of this paper suggested a new approach of Hierarchical Knowledge and Multi-Pooing. In their method knowledge information can be gathered from three different levels – character level, local level and global level as well. They utilized this information to solve the 'weak features' problem. They also utilized multi-pooling approach that is used to extract and evaluate multiple features of the sentiments [7]. Authors used PSOGO sentiment analysis approach to improve the performance of sentiment analysis with IG for feature selection and SVM because of the learning engine. The tested and validated their model on two different datasets of different fields.[8]

Authors used polarity distribution in their suggested framework. Their framework increased the accuracy of the polarity consistency. Their experiments are conducted to suggest inter- and intra-dictionary inconsistencies on five sentiment dictionaries and WorldNet[9].

Subjectivity detection is the method of extracting subjective statements from results. Sentiment analysis is utilized to automate the analysis of such results. Sindhu Chandra Sekharanet a. put their work to find opinionated data based on its polarity. Their work can give companied opinionated feedback for their product and they can further make improvements in it [10].

The researchers presented a language independent model for the prediction of polarity, positive or negative opinion on any subject given in a natural language text. The assignment of weights is done for attributes, individual words etc. based on their position and its probability of being subjective in nature. The subjectivity of each attribute is calculated in a two-step process. The results of the assessment on a regular movie review dataset show 89.85 percent accuracy of classification [11].

The Multimodal Opinion level Sentiment Intensity dataset ((MOSI), the first opinion-level annotated corpus of sentiment and subjectivity analysis in online videos, is introduced in this paper. Subjectivity, sentiment intensity, per-frame and per-opinion annotated visual features, and per-milliseconds annotated audio features are all labeled in the dataset [12].

The literature for subjectivity detection is reviewed in this article, which includes both hand-crafted and automatic versions. It focuses on the key assumptions that these models make, the results they produce, and the problems that remain to be investigated in order to enhance our understanding of subjective sentences. Finally, the benefits and disadvantages of each strategy are weighed. Hand-craft is a wide category that includes a range of techniques[13].

Methodology

We had used subjectivity and polarity for the sentiment identification. Text Blob is open-source library that is being used for variouas Natural Language Processing tasks. Words are used in the form of bigrams. The range of polarity score is **-1.0 - 1.0 .**The range of subjectivity is *0.0 - 1.0.*Sentiment score is calculated using Review Text and Sentiment Column for test set and train set both. Finally, we applied machine learning algorithm on reviews and sentiment score column and calculated accuracy for the various models.

Dataset

The dataset we used is take from Kaggle.com is based on "Women's Clothing E-Commerce Reviews". The dataset is a comma-separated (.csv) file.This dataset includes 23486 rows and 10 feature variables. Our work is based on Review Text and Rating column. The Rating column stores the customer rating from 1 (Worst) to 5 (Best).

Text Pre-processing

Few reviews are selected from the dataset to explain the process of preprocessing followed in this work -

1. "Absolutely wonderful - silky and sexy and comf..."
2. "Love this dress! it'ssooo pretty. ihappene..."
3. "I had such high hopes for this dress and reall... "

The preprocessing starts with the cleaning of the raw data which is further converted into tokens followed by its normalization. The results of the same are displayed in Table number 1-5

Cleaning Raw Data-The reviews collected may have some words/characters that are not significant with respect the meaning of the text is being removed.

Lowering case-This step will convert the unqiue words into lowercase. This step will remove the data sparsity and and it will also reduce the size of dataset as well.

Removal of special characters-This step will remove some special characters mentioned in the review. For example-!,$,% . It will make processing of the reviews easier.

sub (actual pattern, replacing pattern, data)
This allows us to substitute the second argument in the data for the first argument.

Reviews	***Clean_review***
Absolutely wonderful - silky and sexy and comf.. Love this dress! it'ssooo pretty. ihappene...	absolutely wonderful silky and sexy and comfo... love this dress its sooo pretty i happened t...

TABLE 1: ***Cleaned Raw Data***

Removal of stopwords-Stop words of a language that help you in sentence formation. However, while processing the sentiments, stop words in general do not make any contribution in detection of the sentiments. Hence, we need to remove these stopwords like ‘a', 'an', 'the', 'is', ‘what’ etc. After removing stop words, we had calculated the word count for the left words in each review.

Review Row	**Removal of Stop word**	**Word Count**

Absolutely wonderful - silky and sexy and comfort…	absolutely, wonderful, silky, sexy, comfortable	5
Love this dress! it'ssooo pretty. ihappene...	love, dress, sooo, pretty, happened, find, st...	31
I had such high hopes for this dress and reall...	i, high, hopes, dress, really, wanted, work, ...	48

TABLE 2: ***Stop Word Removal***

Removal of URLs- The dataset reviews may contain some URLs, we need to remo ve them from the review for processing it effectively as the URLs do not contribute in sentiment assessment of the reviews.

Removal of HTML Tags- In case we had collected data through web scraping then the data might consist of some HTML Tags which we need to remove before processing the dataset.

Calculated most common words by rating- Using 'Rating Column' and their words-counts we had calculated most common words from reviews. A table is shown the most common words-

1	2	3	4	5
(dress, 372)	(dress, 739)	(dress, 1394)	(dress, 2291)	(dress, 5654)
(I, 353)	(I,706)	(I,1281)	(size,2164)	(love,5342)
(like, 348)	(like, 691)	(top, 1119)	(I, 2037)	(I, 5009)

Review Row	POS
Absolutely wonderful - silky and sexy and comfort…	(Absolutely, RB), (wonderful, JJ), (silky, JJ...
Love this dress! it'ssooo pretty. ihappene...	(Love, NNP), (dress, NN), (sooo, NN), (pretty..
I had such high hopes for this dress and reall...	[(I, PRP), (high, VBP), (hopes, NNS), (dress, ..
Review Row	**POS**
Absolutely wonderful - silky and sexy and comfort…	(Absolutely, RB), (wonderful, JJ), (silky, JJ...
Love this dress! it'ssooo pretty. ihappene...	(Love, NNP), (dress, NN), (sooo, NN), (pretty..
I had such high hopes for this dress and reall...	[(I, PRP), (high, VBP), (hopes, NNS), (dress, ..

Table 3: ***Most common words for 'Rating Column'***

Identifying Parts of Speech (POS)-The next step is to classify the tokens of the reviews lexically. It means depending upon the POS the tokens are classified as per their usage in the dataset review. POS – tagger associates a tag to each keyword in a review. Table 4 display results of POS tagging.

TABLE 4:***Identification of Parts of Speech***

Using POS tag over 'Rating column' we had extracted adjective and noun for the given reviews. A table has showed the noun and adjective words with their numbers.

Adjective				
1	2	3	4	5
(top, 204)	(top, 422)	(top, 830)	(top, 1386)	(great, 3848)
(small, 162)	(small, 340)	(small, 664)	(great, 1240)	(top, 2957)
(fabric, 108)	(large, 213)	(large, 457)	(small, 1134)	(small, 2259)
Noun				
dress, 301)	(dress, 586)	(dress, 1104)	(size, 2164)	(size, 4847)
(size, 188)	(size, 482)	(size, 1009)	(dress, 1861)	(dress, 4532)
(fabric, 122)	(color, 249)	(color, 567)	(color, 978)	(color, 2463)

TABLE 5: ***Identification of Adjectives and Nouns***

Calculated Bigrams- N-gram feature is used popularly for language modelling purpose. TextBlob facilitates us to access ngrams using inbuilt function "ngrams". The function returns a tuple of successive words. In the modle bigrams are created from the rating column and finally calculated frequency using in built function FreqDist(). Results displayed in Table 6.

1	**2**	**3**	**4**	**5**
((going, back), 44)	((wanted, love), 96)	((wanted, love), 146)	((true, size), 243)	((true, size), 916)
(looks, like), 50)	((going, back), 98)	((I, love), 154)	((I, love), 319)	((I, love), 1143)

((looked, like), 44)	((looked, like), 89)	((I, ordered), 114)	((This, dress), 215)	((This, dress), 532)

TABLE 6: ***Bigrams with their frequency***

Sentiment Calculation- We had calculated sentiment on the basis of the "Review column". The method of evaluating the writer's attitude or emotion, whether positive, negative, or neutral, is known as sentiment analysis. The sentiment function of text blob returns two properties, polarity, and subjectivity.Polarity is a kind of float value that lies in the range of -1.0 to +1.0. Here 0 signifies neutral, +1 signifies very positive and -1 very negative.

Subjectivity is a float value between 0.0 and 1.0, with 0.0 being the most objective and 1.0 being the most subjective. Objective sentences are factual, while subjective sentences convey personal feelings, views, beliefs, opinions, allegations, wishes, assumptions, and speculations.Table 7displays the sample result of subjectivity and polarity. This table tells which reviews shows positive word shows higherpolarity, lower subjectivity and which reviews has negative sense shows lower polarity and higher subjectivity. Then we had calculated sentiment score on rating column.Using this sentiment score we had calculated the accuracy of the various models.

Reviews	**Subjectivity**	**Polarity**	**Sentiment Score**
Review1	0.933333	0.633333	2
Review2	0.725000	0.339583	1
Review3	0.625000	0.550000	2

TABLE 7 :*Results forSubjectivity,Polarity and Sentiment*

RESULTS AND DISCUSSION

The work showcased an identified subjectivity model using a pre labeled dataset. Our worked is based on the 'Review, Rating column.The first step is data preprocessing.Three sample reviews from the dataset are taken to demonstrate the working of preprocessing steps and for sentiment calculation. The preprocessing covers the cleaning, tokenizing and removing stopwords from the reviews. Finally,

parts of speech(PoS) tagging process is performed and also calculated adjective and noun and also calculated frequency distribution of word over rating column, the intermediate result of three reviews has been shown in the Table2-6.We calculated subjectivity and polarity of a sentiment using TextBlob which is shown in table 7. Finally, we had calculated sentiment score over rating column.

The feature set is generated using BoW with the help of CountVectorizer function. The dataset is classified into 70 percent training and 30 percent test data. Two popular models are used to evaluate accuracy MultinomialNB and DTR. Table 8 showcases the results of the 02 classifiers-

Model	**Accuracy Score**	**Precision Score**	**Recall Score**	**f1 Score**
Multinomial NB	**0.769927**	0.583356	0.352107	0.326443
Decision Tree	0.754471	0.514419	0.490711	0.500629

TABLE 8:*Results*

Conclusion

In this paper, using a textblob, we identified sentiment analysis using subjectivity. Text Blob is a Python library with a simple API for communicating with its methods and performing simple NLP tasks. General text analytics operation using NLTK tokenization, noun and adjective phrase extraction, stops word removal, POS tagging-grams and sentiment extraction etc. performed on the text. In this research subjectivity and polarity is used. We had calculated sentiment score for sentiment identification. Bi-grams words are used in this work. We need to improve the sentiment by using sentence level or phrase level analysis. Our model performs the best result with an accuracy of 76.9%. The accuracy improvement can be done by correcting misspelled word to some extent.

REFERENCES

[1]Ghorbani M., Bahaghighat M., Xin Q. and Özen F. "ConvLSTMConv network: ADeepLearning Approach for Sentiment Analysis in Cloud Computing" Ghorbani et al. Journal of Cloud Computing: Advances System and Applications (2020) 9:16 https://doi.org/10.1186/s13677-020-00162-1

[2]] Wei.Zhang1·Sui-xi Kong1·Yan-chun Zhu2·Xiao-le Wang3,"Sentiment classification and computing for online reviews by a hybrid SVM and LSA based approach", Received: 19November2017/Revised: 20December2017/Accepted: 29December2017 ©SpringerScience+BusinessMedia, LLC, part of SpringerNature2018, https://doi.org/10.1007/s10586-017-1693-7

[3]El-Din MoheyDoaa (2017), "Negative Polarity Levels for Sentiment Analysis ", d.mohey@alumni.fci-cu.edu.eg February 2017

[4] Mohan .I, Rani D.,G Pranathi.M .J .A "Corpus Based Dual Sentiment Analysis" 07 June 2019,stats and author profiles for this publication at: https://www.researchgate.net/publication/333651351

[5] Nigam S., , Das A.K., BalabantarayC.R.,"Machine Learning Based Approach To Sentiment Analysis "a116018@iiit-bh.ac.in, ajit@iiit-bh.ac.in, rakesh@iiit-bh.ac.in ,International Conference on Advances in Computing, Communication Control and Networking (ICACCCN2018) ISBN: 978-1-5386-4119-4/18/$31.00 ©2018 IEEE

[6] Kalaivani A, ThenmozhiD."Sentimental Analysis using Deep Learning Techniques", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878 Volume-7, Issue-6S5, April 2019, Retrieval Number:1070476S519/19©BEIESP

[7]. WANG H., HOU M., LI F., AND ZHANG Y (2020)." Chinese Implicit Sentiment Analysis Based on Hierarchical KnowledgeEnhancement and Multi-Pooling "Received June 1, 2020, accepted July 7, 2020, date of publication July 13, 2020, date of current Version July 21, 2020.Digital Object Identifier 10.1109/ACCESS.2020.3008874

[8] Li X, Li J, Wu Y (2015) "A Global Optimization Approach to Multi-Polarity Sentiment Analysis".PLoSONE10 (4):e0124672.doi:10.1371/ journal.pone.0124672

[9]C.SUJITHRA AND A.ARUNKUMAR "POLARITY CONSISTENCY CHECKING FOR MULTIPOLARITY BASED DOMAIN INDEPENDENT SENTIMENT DICTIONARIES" International Journal of Scientific & Engineering Research Volume 8, Issue 5, May-2017 ISSN 2229-5518

[10] C. Sindhu , B. Sasmal, R. Gupta , J. Prathipa," Subjectivity Detection for Sentiment Analysis on Twitter Data" Department of Computer Science and Engineering, SRM Institute of

Science and Technology, Chennai, India e-mail: sindhucmaa@gmail.com Notes in Networks and Systems 130, https://doi.org/10.1007/978-981-15-5329-5_43

[11] VeselinRaychev and PreslavNakov" Language-Independent Sentiment Analysis Using Subjectivity and Positional Information" Department of Mathematics and Informatics Sofia University "St KlimentOhridski" 5, James Bourchier Blvd.,1164 Sofia, Bulgaria {veselin.raychev, preslav.nakov}@fmi.uni-sofia.bg[1911.12544v1] (arxiv.org)

[12]Zadeh A., Zellers R., Pincus E.," MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos".

[13] ItiChaturvedi, Erik Cambria, Roy E. Welsch, Francisco Herrera, "Distinguishing Between Facts and Opinions for Sentiment Analysis": Survey and Challenges, Information Fusion (2017), doi: 10.1016/j.inffus.2017.12.006