Building Sentiment analysis Model using Graphlab

First Mona Mohamed Nasr, Second Essam Mohamed Shaaban, and Third Ahmed Mostafa Hafez

Abstract —Sentiment analysis is called opinion mining which is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes. Starting from the importance of the sentiment analysis generally for individuals and more specifically for gigantic organizations, we started digging in this paper. Graphlab was used to build the sentiment models. Many algorithms were used along with text features selection techniques to predict the positive and negative sentiments like "SVM", "logistic regression" and "boosted trees". The mentioned classifiers were applied to a Hotel reviews dataset got from Trip Advisor website to emulate real customer opinions. The results showed that using SVM classifier along with N-grams features selection technique was superior to others.

Keywords—Classification, Feature Selection, Support Vector Machine (SVM), Logistic Regression, Decision trees.

1 Introduction

The revolution of social media, e.g.(reviews, forum discussions, blogs, microblogs, Twitter, and social networks)makes it easy to know the reviews of any product. Hence the need for analyzing sentiments (reviews) has emerged.Sentiment analysis, also called opinion mining, is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes [1].In recent years many researchers built sentiment models to analyze product reviews and classify them to positive and negative sentiments. Ortigosa et al[2]proposed a hybrid approach that combines lexical-based and machine-learning techniques. The results showed that it is feasible to perform sentiment analysis in Facebook with high accuracy (83.27%).Parkhe and Biswas[3] focused on aspect-based sentiment analysis of movie reviews in order to find out the aspect specific driving factors. These factors are the score given to various movie aspects and generally, aspects with high driving factors direct the polarity of the review the most. They depend on Lexicons, POS, A Naïve Bayes and SVM classifier. The results showed that by giving high driving factors to Movie, Acting and Plot aspects of a movie, we obtained the highest accuracy in the analysis of movie reviews about 79.372%.Nagamma et al[4]applied sentiment analysis for studying the relationship between the online reviews for a movie and the movies box office revenue performance. They useda hybrid approach that combines Term Frequency (TF) and Inverse Document Frequency (IDF) values as features along with Fuzzy Clustering and Support Vector Machine (SVM) Classifier for predicting the trend of the box office revenue from the review sentiment. The results showed that using reviews based on clustering has helped to show an improvement in the accuracy from 62% to 89.65% on SVM classifier with and without clustering. While using NB classifier gave an accuracy of 72.41% under both conditions.Hegde & Padma[5]applied a case study of Kannada SA for mobile product reviews .they used a lexicon-based method for aspect extraction.

Furthermore, the Naive Bayes classification model is applied to analyze the polarity of the sentiment due to its computational simplicity and stochastic robustness. Therefore, a customized corpus has been developed. Their preliminary results indicate that this approach is an efficient Technique performed with 65 % accuracy for Kannada SA.

In this paper sentiment model was built by using SVM, Decision trees, and Logistic Regression depending on Hotel reviews dataset crawled from Trip Advisorafter applying some modification and transformation from web form to CSV form. All models were built by using IPython Notebook with Graphlab module and SFrame package. The results show that the Sentiment Model-based SVM with N-grams features is superior toothers.

2 Implementation Package

During the implementation phase; Ipython notebook with GraphlabCreate are used to scale much larger data than other available resources like Pandas.

IJSER

2.1 Ipython Notebook

A powerful interactive Python shell; One of Python's most useful features is its interactive interpreter. It allows fast testing of ideas without the overhead of creating test files as is typical in most programming languages. However, the interpreter supplied with the standard Python distribution is somewhat limited for extended interactive use. The goal of IPython is to create a comprehensive environment for interactive and exploratory computing. To support this goal, IPython has three main components. An enhanced interactive Python shell.A decoupled two-process communication model, which allows multiple clients to connect to a computation kernel, most notably the web-based notebook provided with Jupyter.An architecture for interactive parallel computing.All of IPython is open source (released under the revised BSD license) [6].

2.2 Graphlab Create

GraphLab Create provided from Turi Company. Graphlab is a Python library, backed by a C++ engine, for quickly building large-scale, high-performance data products. Some key features of GraphLab Create are Analyze terabyte scale data at interactive speeds, on your desktop,a single platform for tabular data, graphs, text, and images,state of the art machine learning algorithms including deep learning, boosted trees, and factorization machines. [6].

2.3 SFrame

Sframe is a component of Graphlab that is a tabular, column-mutable data frame object that can scale to big data. The data in SFrame is stored column-wise on the GraphLab Server side, and is stored on persistent storage (e.g. disk) to avoid being constrained by memory size. Each column in SFrame is a size-immutable SArray, but SFrames are mutable in that columns can be added and subtracted with ease. SFrame essentially acts as an ordered dict of Sarrays [6].

2.4 CLASSIFICATION ALGORITHMS

Classification algorithms use an existing dataset with predefined categories to build a predictive model for future classification.

2.4.1 Support Vector Machine (SVM)

SVM is a supervised learning model. SVM classification technique analyzes data and recognizes patterns from them. SVM uses a very small sample set and generate pattern from that[7]. SVM represents a powerful technique for general (nonlinear) classification, regression and outlier detection with an intuitive model representation. Itincludes linear, polynomial, radial basis function, and sigmoidal kernels[8]. The mainsignificance of the SVM is that it is less susceptible for over fitting of the feature input from the input items, this is because it is independent on feature space. SVM is fast accurate while training as well as during testing[9].

2.4.2 Logistic Regression

Logistic regression is workhorse of statistic and it can be used for binary classification or for predicting the certainty of binary outcome [10]. Logistic regression is a probabilistic statistical classification method, which has been widely applied to two class classification tasks [11]. The logistic regression model has been used for many years to explain a binary response variable Y through a vector of explanatory variables ($X1,2, \ldots, Xp$). Which can be quantitative, qualitative or both. This model is more flexible than the linear regression model since it does not have the requirements of the independent variables to be normally distributed, linearly related, nor equal variance within each group [12].

2.4.3 Decision Trees

Decision tree is one of the most widely used classification tools. As its name implies, a decision tree can be viewed as a classifier in the form of a tree structure, in which each node is either a leaf node or decision node. All decision nodes have splits, testing the values of some functions of data attributes. Each branch from the decision node corresponds to a distinct outcome of the test. Each leaf node has a class label attached to it. The most crucial issues in decision tree classification modeling are selecting the splits and determining the size of the tree[13].

2.5 Feature Selection

2.5.1 Term Frequency–Inverse Document Frequency (TF-IDF)

The prototypical application of TF-IDF transformations involves document collections, where each element represents a document. Documents are represented in a bag-of-words format, i.e. a dictionary whose keys are words and whose values are the number of times the word occurs in the document[14].

2.5.2 N-Grams

N-gram based techniques are predominant in modern natural language processing (NLP) and its applications. Usually, they are used as features in representing vector space model and then the standard classification algorithms are applied for this model. N-grams are sequences of elements as they appear in texts. These elements can be words, characters, POS tags or any other elements as they appear one after another in texts. Common convention is that ''n'' in n-grams corresponds to the number of elements in a sequence [15].

IJSER

3 IMPLEMENTATION METHODOLOGY

The proposed sentiment model depends on a preprocessed Hotel reviews datasetafter applying features selection techniques then applying ML algorithms as shown in figure 1.

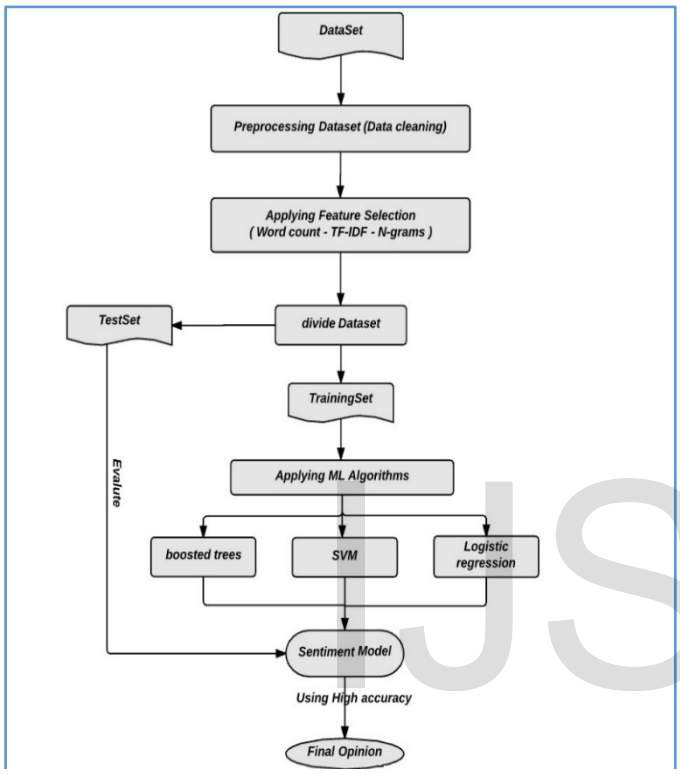

Fig. 1. The Proposed Sentiment Model

Hotel reviews dataset that wascrawled from Trip Advisorand was downloaded from DAIS repository [16] was used in the classification Algorithms. Meta data includes: Author, Content, Date, Number of Reader, Number of Helpful Judgment, Overall rating, Value aspect rating, Rooms aspect rating, Location aspect rating, Cleanliness aspect rating, Check in/front desk aspect rating, Service aspect rating and Business Service aspect rating. Ratings ranges from 0 to 5 stars, and -1 indicates this aspect rating is missing in the original html file [17]-[18].

Firstly the data were in HTML format. Using the Linux shell scripts the data was converted from HTML format to CSV format. Data cleaning was run in the CSV data to remove the empty rows and corrupted data.

Shell script "awk" command was used to organize the dataset and convert it to CSV format. Used the SFrame to get the required column that will be usedinthe study like "Hotel name, review ... etc."Table-1 contains a brief description about dataset after organization.

TABLE 1
DATASET DESCRIPTION

Attribute	Type	Description
Hotel	String	Hotel name
Review	String	Customer opinion
Rate	Integer	Overall rate values from 0 to 5
Sentiment	Binary	Negative sentiment equal 0 and positive sentiment equal 1

Datasetwas randomly divided into training set and test set. Training data contains 171406 reviews and the testing set contains 42675 reviews. Figure 2 shows the percentage of training set and test set.

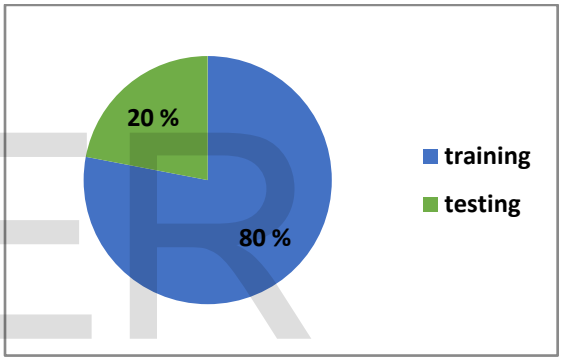

Fig. 2. Training and Testing set.

4 EXPERIMENTAL RESULTS

Inthe pre-processing/cleaningstage, empty cells and reviews were removed. Then the classifierswere applied on Hotel dataset as following:

Boosted trees classifier was able to correctly predict that 2665 reviews were negative and 34977 reviews were positivegiving the least performance and lowest accuracy of 88.21%.

Logistic regression classifier with word count feature given a higher performance and more accuracy of 91.64%. It gave the same results even when combining it withTF-IDF feature.Both of them correctly predicted that 33832 reviews were positive and 5273 reviews were negative. Logistic regression with bigrams feature gavea higher accuracy thanthe otherfeaturesselectionaround 92.75%. It was able to correctly predict 34368 positive reviews and 5211 negative ones.

Support vector machine classifier with word count features gave almost the sameperformance and accuracy

IJSER

compared to the previous classifiersaround91.42 %.It gave the same accuracy even with TF-IDF feature. Those models were able to correctly predict 34316 positive reviews and 4698 negatives. SVM with bigrams feature gavethe highest accuracy of93.50 %. It was able to correctly predict 34922 reviews were positive and 4980 reviews were negative. Table 2 shows more information about correctly and incorrectly predicted records moreover; Table 3 shows accuracy results and error rates for all used classifiers:

TABLE 2
CONFUSION MATRIX FOR THE CLASSIFIERS

Algorithm	Actual class	Actual prediction	
		Positive	Negative
boosted trees classifier	Positive	34977	425
	Negative	4608	2665
logistic regression classifier with word count or TF-IDF feature	Positive	33832	1570
	Negative	2000	5273
Logistic regression with bigrams feature	Positive	34368	1034
	Negative	2062	5211
SVM classifier with word count or TF-IDF feature	Positive	34316	1086
	Negative	2575	4698
SVM with bigrams feature	Positive	34922	480
	Negative	2293	4980

TABLE 3
DETAILS OF EXPERIMENTAL RESULTS ACCORDING TO ACCURACY

Classifier	Feature selection	Accuracy Rate	Error Rate
boosted trees	Word count	88.21 %	11.79 %
Logistic regression	Word count	91.64%	8.36 %
Logistic regression	TF-IDF	91.64%	8.36 %
Logistic regression	N-grams	92.75 %	7.25 %
SVM	Word count	91.42 %	8.58 %
SVM	TF-IDF	91.42 %	8.58 %
SVM	N-grams	93.50 %	6.5 %

The above resultsshow that using SVM with N-grams feature gavethe best performance and accuracy of 93.50 % which is superior to others as shown in figure 3.

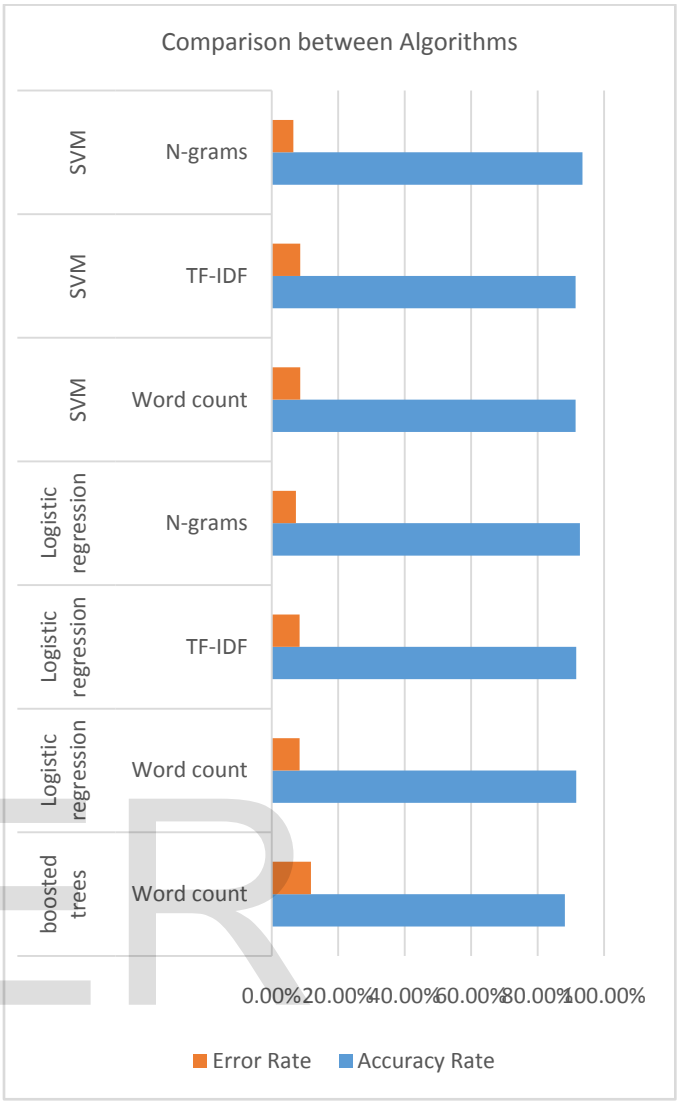

Fig. 3.Comparison between Algorithms with different features According To Accuracy Rate

The ROC curve "Receiver Operating Characteristic curve" is supported by Graphlab create for some classifiers like logistic regression and decision trees, but not supported for SVM. The ROC Curve gives us a statistical view of the performance of the created model across all possible thresholds. The ROC curve for supported classifiers is shown in figure 4.

IJSER

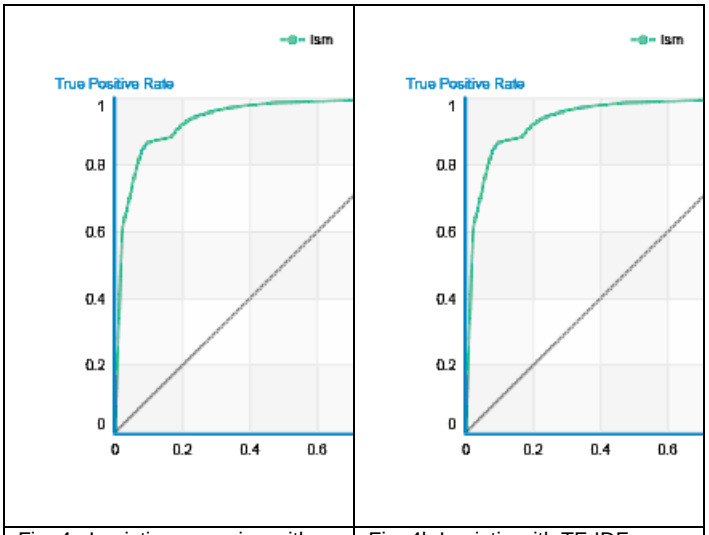

Fig. 4a.Logistic regression with word count feature

Fig. 4b.Logistic with TF-IDF feature

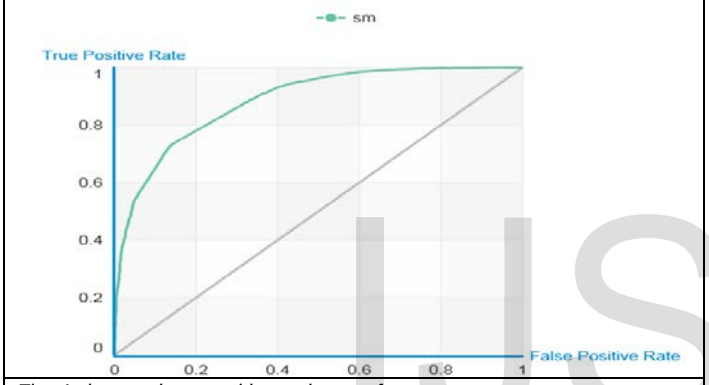

Fig. 4c.boosted trees with word count feature

Fig. 4. ROC curve for supported classifiers

5 CONCLUSION

Sentiment analysis became one of the most important thing for individuals and companies to get the customers' feeling about their products and services. In this paper sentiment analysis has been applied to Hotel reviews by using many classifiers with different feature selections techniques to get the best accuracy and performance. SVM classifier with N-grams feature gave best performance and accuracyof 93.50 %.This result is the highest amongst all researchers in this field mentioned in brief at the beginning of this paper. The highest previous researchers' accuracy was achieved by Nagamma of 89.65% while ours is 93.50 %.

REFERENCES

[1] B. Liu, Sentiment Analysis and Opinion Mining, no. May. 2012.
[2] A. Ortigosa, J. M. Martín, and R. M. Carro, "Sentiment analysis in Facebook and its application to e-learning," Comput. Human Behav., vol. 31, no. 1, pp. 527–541, 2014.
[3] V. Parkhe and B. Biswas, "Aspect based sentiment analysis of movie reviews: Finding the polarity directing aspects," Proc. - 2014 Int. Conf. Soft Comput. Mach. Intell. ISCMI 2014, pp. 28–32, 2014.
[4] P. Nagamma, H. R. Pruthvi, K. K. Nisha, and N. H. Shwetha, "An improved sentiment analysis of online movie reviews based on clustering for box-office prediction," Int. Conf. Comput. Commun. Autom. ICCCA 2015, pp. 933–937, 2015.
[5] Y. Hegde and S. K. Padma, "Sentiment Analysis for Kannada using Mobile Product Reviews," pp. 822–827, 2015.
[6] "Ipython API Documentation." [Online]. Available: http://ipython.readthedocs.io/en/stable/overview.html.
[7] R. Pandya, "C5 . 0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning," vol. 117, no. 16, pp. 18–21, 2015.
[8] D. Meyer, "Support Vector Machines," vol. 1, pp. 1–8, 2015.
[9] I. J. Of, "Research in Computer Applications and Robotics a Survey on Trust Based," vol. 4, no. 4, pp. 55–58, 2016.
[10] kaThomas P. Min, "Algorithms for maximum-likeihood logistic regression."
[11] W. Hu, Y. Qian, and F. K. Soong, "A new Neural Network based logistic regression classifier for improving mispronunciation detection of L2 language learners," Proc. 9th Int. Symp. Chinese Spok. Lang. Process. ISCSLP 2014, pp. 245–249, 2014.
[12] A. P. De Barros and F. D. A. Ten, "A Pattern Classifier for Interval-valued Data Based on Multinomial Logistic Regression Model," pp. 541–546, 2012.
[13] Y. Liu and G. Salvendy, "Design and evaluation of visualization support to facilitate decision trees classification," Int. J. Hum. Comput. Stud., vol. 65, no. 2, pp. 95–110, 2007.
[14] "turi user guide." [Online]. Available: https://turi.com/learn/userguide/feature engineering/tfidf.html.
[15] G. Sidorov, F. Velasquez, E. Stamatatos, A. Gelbukh, and L. Chanona-Hern??ndez, "Syntactic N-grams as machine learning features for natural language processing," Expert Syst. Appl., vol. 41, no. 3, pp. 853–860, 2014.
[16] "DAIS repository." [Online]. Available: http://times.cs.uiuc.edu/~wang296/Data/.
[17] H. Wang, Y. Lu, and C. Zhai, "Latent aspect rating analysis without aspect keyword supervision," Kdd, pp. 618–626, 2011.
[18] H. Wang, Y. Lu, and C. Zhai, "Latent aspect rating analysis on review text data," Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD '10, p. 783, 2010.

IJSER

Mona Nasr is currently Vice Dean of Faculty of Computers and Information, Helwan University for Community Service and Environmental and Ex-Vice Dean at the Canadian International College (CIC), El Sheikh Zayed Campus. She received the Ph.D. degree in Information Systems from Faculty of Computers and Information, Helwan University, Egypt, 2006. She awarded a trophy (The Best Senior Level STEM Executive 2015), full fellowship from the Bibliotheca Alexandrina (BA) in cooperation with the Arab Regional Office for the World Academy of Sciences for the advancement of science in developing countries (TWAS-ARO). And she awarded full fellowship from the Cyprus Institute, 2012. She has 26 journal publication, 30 conferences publications and 1 book chapter publication.

Essam Shaaban is now Assistant Professor, Information Systems Department Faculty of Computers and Information Beni-Suef University, and Assistant professor as a part timer Information Systems Department Faculty of Computers and Information Future University in Egypt (FUE), October 6 University, MUST University, and Helwan University. He received master degree in 2007, and PhD degree 2013. He has 3 journal publications and 1 conference publication.

IJSER

Ahmed Mostafa is now working as value added service engineer at Huawei. He received diploma degree at business information technology at faculty of computer and information 2013, and pre master courses in information system faculty of computer and information Helwan University 2014.