IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.301

Volume 3, Issue 12, May 2023

Fake Reviews Detection using Support Vector Machine

Prof. A. G. Shahapurkar, Diksha Badgujar, Mahesh Khadse, Pruthviraj Thopte, Avdhut Patil
Department of Computer Engineering
Sinhgad Academy of Engineering, Pune, Maharashtra, India
Savitribai Phule Pune University, Pune, Maharashtra

Abstract: *Social media is an effective informational channel for sharing details about the goods and services offered by online retailers. Customers who have purchased the goods themselves offer this information. Analysis of customer-cited features and specifications based on their sentiment. These descriptions and reviews may be found on the Flipkart and Twitter websites. Reviews of features/specifications from the Twitter and Flipkart websites were considered for this study project. As a result, the work's analysis of customers' issues with purchasing high-quality goods was its focus. For the purpose of evaluating comments, this work automates the process of extracting semantic-based elements or features and their opinions.*

Keywords: Sentiment Analysis, Aspect, Fuzzy Logic, Ecommerce, Customer Reviews, Decision Making

I. INTRODUCTION

Nowadays, when customers want to draw a decision about services or products, reviews become the main source of their information. For example, when customers take the initiation to book a hotel, they read the reviews on the opinions of other customers on the hotel services. Depending on the feedback of the reviews, they decide to book room or not. If they came to positive feedback fromthe reviews, they probably proceed to book the room. Thus, historical reviews became very credible sources of information to most people in several online services. Since,reviews are considered forms of sharing authentic feedback about positive or negative services, any attempt to manipulate those reviews by writing misleading or inauthentic content is considered as deceptive action and such reviews are labelled as fake. Social media is an effective informational channel for sharing details about the goods and services offered by online retailers. Customers who have purchased the goods themselves offer this information. Analysis of customer-cited features and specifications based on their sentiment. These descriptions and reviews may be found on the Flipkart and Twitter websites. Reviews of features/specifications from the Twitter and Flipkart websites were considered for this studyproject. As a result, the work's analysis of customers' issues with purchasing high-quality goods was its focus. For the purpose of evaluating comments, this study automates the extraction of semantic based elements or traits and associated opinions.Sentiment analysis has for the past few years gained attentionand popularity as a focus area for research. This has been attributed due to large amounts of textual data created in a variety of social networks, the web and other information centric applications It has been studied on various topics like movie reviews, newspaper articles, products reviews, restaurant reviews etc. Sentiment analysis or opinion mining attitudes towards a certain issue or topic ranging from individuals, events, products, services etc. It is a technology for extracting opinions from unstructured human authored documents closely related to data mining and uses machine learning techniques to identify, analyze the sentiments expressed in a text. Sentiment analysis (SA) has many applications and can be applied in various fields such as political debates, marketing, ecommerce etc. SA helps governments in assessing their strengths and weaknesses by analyzing the opinions from the public. SA helps companies get new marketing strategies andimprove product features. Consumers can also use sentiment analysis to research products or services before making a purchase while opinion of their company and their products or analyze customer satisfaction. The last ten years have been instrumental in the growth of the research on sentiment analysis mainly due to massive growth of

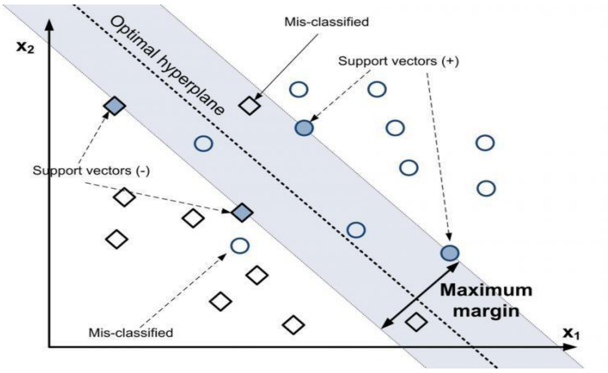
DOI: 10.48175/568

II. RELATED WORK

Opinions have always been important and integral part of human life influencing our behaviors. What other people think has always been an important piece of information during the decision-making process as both individuals and organizations seek opinions from friends and the public.

III. METHODOLOGY

Pre-processing of data Pre-processing will be carried out in the following steps: a) Tokenization - Tokenization involves splitting a stream of text into smaller units called tokens, usually, words or phrases. During tokenization, emoticons, twitter @usernames, URLs, and hashtags are considered and treated as individual tokens. Tokenization makes it easy to separate unnecessary symbols and punctuations and filter out only the words that can add valueto the sentimental polarity score of the text. b) Normalization - In this step, all upper-case words (For example: EDUCATION SYSTEM should be reformed) are converted into lower case characters. Abbreviations within atweet is noted and replaced by the actual meaning they represent (for example: OMG > Oh My God) while words having received my exams results!! Im reeeeaaallly happyyyy) are replaced by a single character. c) Part of Speech Tagging (POS): This is carried out by POStagger, which looks at various words in a tweet. In POS tagging, special tokens (such as hashtags URLs) are noted and Feature: Feature extraction is the process of transforming the input data into set of features. The performance of machine learning algorithm depends heavily on its features so it is crucial to choose exact Feature Extraction. On the other hand, our main goal is to apply several n-gram models which is unigram, bigrams, and trigrams to compare different n-grams schemes.



1) Term-weighting Scheme

The calculation of the term- weighting scheme plays a crucial role in extracting the most classical features as an input to the classifier. The more classical the features, the better performance of the classifier will be. The experiments applied several term- weighting schemes, consists of Term Frequency Inverse Document Frequency (TFIDF), Binary Occurrences (BO) and Term Occurrences (TO) for each n-gram scheme to create word vectors.

They are based on the following count:

fij the number of $total\ number$

$of\ documents\ in\ which\ term\ iappears$ at least once

Based on these counts, there are 3 classes available tomeasure the importance of term for document j, as denoted by vij:

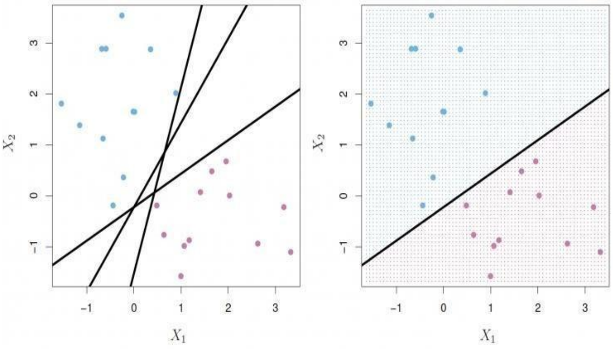
TFIDF - the tf/idf measure with $vij = fij\ fdj$ log ($|D|\ fti$),

where |D| is the total number of documents. The resulting vector for each document is normalized to the Euclidean unit length.Binary Occurrences (BO) - occurrences as a binary value Vij = {1, fij > 0 0, $else$

The resulting vector is not normalized.

Term Occurrences (TO) - the absolute number of occurrences of a term vij =fij the resulting vector is not normalized.

Classification using Support Vector Machine. Support Vector Machine (SVM) has been chosen for the classification in the experiments. The support-vector machines are a learning machine for two-group classification problems introduced. It is used to classify the texts as positives or negatives. SVM works well for text classification due to its advantages such as its potential to handle large features. Another advantage is SVM is robust when there is a sparse set of examples and because most of the problem are linearly separable. Support Vector Machine have shown promising result in previous researches in sentiment analysis. For classification, Support Vector Machine (SVM) with grid search and K- fold cross validation technique is used. Grid Search is basically a model for hyper parameter optimization. Hyper parameter tuning is an important task in SVM to extract more accurate results.



In Grid-Search, different models having different parameter values are trained and then evaluated using cross validation. For an RBF kernel, there are two parameters: C and ϒ. It cannot be ascertained in advance that whichvalues of C and ϒ are best suited fora given problem, so an optimized model is required which can identify the ideal pair of values for these parameters to achieve maximum accuracy. The process of 10-k cross validation is performed on each model of C and ϒ and the pair with optimum resultsis selected. Cross validation is a method used to test multiple models under a particular classifier with the subset of input data as explained by for K-fold cross validation, the training data is first divided into k subsets of same size. One subset is tested using the classifier on the remaining k-1 subsets. The cross- validation procedure can prevent the overfitting problem.

We have seen how the algorithm separates different datasets to classify them correctly. It maps the data to a higher dimensional space if the data is not separable in the ower dimensions. The algorithm just does not fit anyhyperplane, it tries to fit an optimal hyperplane.

As shown in the image above, we have blue and red data points, we want SVM algorithm to separate them so that it can learn and classify new and unseen points as red and blue. As seen in the left image, the SVM algorithm can pick any of the shown hyperplanes

IV. LITERATURE SURVEY

Aspect Based sentiment. summarization using Fuzzy Logic-Jenifer Jothi Mary and Dr. L. Arockiam- [1] The study of attitudes, feelings, and evaluations of individuals toward goods and events is known as sentiment analysis (SA).[2] For a range of uses, it attracted a lot of interest in the past from both business and academics.[3] People must make decisions; thus, opinions are important.[4] Not just forindividuals, but also for corporate entities, it is beneficial. Fuzzy logic can offer a speedy solution to the ambiguity prevalent in most natural languages.[5] The paper provides the case-based reasoning technology application elements of constructing a three-dimensional displacement control system. [6] In order to improve the system's accuracy when locating the investigated object in three dimensions, fuzzy

V. FUTURE SCOPE

Customers increasingly rely on reviews for product information. However, the usefulness of online reviews is impeded by fake reviews that give an untruthful picture of product quality. Therefore, detection of fake reviews is needed. In today’s generation this way of encouraging the consumers to write a review about a product has become a good strategy for marketing their product through real audience’s voice. Such precious information has been spammed and manipulated. Out of many researches one fascinating research was done to identify the deceptive opinion spam.

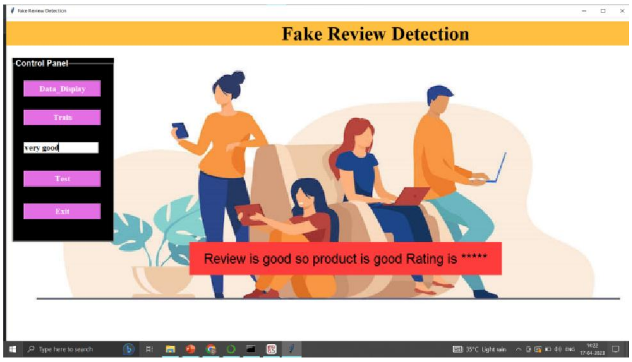
VI. RESULT

Sentiment analysis: This method involves analyzing the language used in a review to determine its overall sentiment. If a review seems overly positive or negative, it may be flagged as potentially fake.

DOI: 10.48175/568

Reviewer analysis: This approach involves examining the reviewer's history, including their review frequency, the number of reviews they have posted, and the types of products or services they review. If a reviewer has a suspicious history, their reviews may be flagged.

Machine learning: Machine learning algorithms can be trained on large datasets of real and fake reviews to learn to identify patterns and characteristics of fake reviews. These algorithms can then be used to identify new fake reviews based on these learned patterns.



very good

Review is good so product is good Rating is *****

VII. CONCLUSION

In this paper we proposed a fundamentally different approachto address the issue of multi-output for classification tasks. Previous approaches worked with the assumption that different classes need to be mutually exclusive in multi-class or multilabel classification tasks, due to discriminative learning of classifiers. This paper has proposed to transform a discriminative single-task classification problem into a generative multi-task classification problem. In other words, the class attribute, which is typically involved in a multi-classor multi-label classification task, needs to be transformed into several binary attributes, each of which is corresponding to one of the predefined class labels and could be independent or correlated to the other labels

ACKNOWLEDGMENT

Firstly, we wish to thank Prof. S. B. Ghawate our internal project guide who gave their co-operation timely and precious guidance without which this project would not have been a success. We thank them for reviewing the entire project with painstaking efforts and more of their profound ability to spot mistakes.

We would also like to thank our H.O.D Prof. S. N. Shelke for his continuous encouragement, support, and guidance at each and every stage of our project.

Lastly, we would like to thank all project stakeholders who were associated with the project and helped in its planning and execution. The project named “Sign Language Translation” would not have been possible without the extensive support of people who were directly or indirectly involved in its successful execution.

REFERENCES

[1] S. Das, S. Kar, and T. Pal, “Robust decision making using intuitionistic fuzzy numbers”, Granular Computing, vol. 1, pp. 1-14, 2016.

[2] Z. Xu and H. Wang, “Managing multi-granularity linguistic information in qualitative group decision making: an overview”, Granular Computing, vol. 1, pp. 21-35, 2016.

[3] B. Liu, “Sentiment Analysis and Opinion Mining”. Synthesis lectures on human language technologies, vol. 5, pp. 1-167, 2012.

[4] W. Perycz and S.M. Chen, Sentiment Analysis and Ontology Engineering: An Environment of Computational Intelligence. Springer, Heidelberg, 2016.

[5] J. Cendrowska, “PRISM: An algorithm for inducing modular rules”, International Journal of Man-Machine Studies, vol. 27, pp. 349-370, 1987.

[6] P. W. Frey and D. J. Slate, “Letter recognition using Holland-style adaptive classifiers”, Machine Learning, vol. 6, pp. 161-182, 1991

[7] M. Buscema, “MetaNet: The theory of independent judges”, Substance Use and Misuse, vol. 33, pp. 439-461, 1998.

[8] J. Wei, Q. Meng, and A. Badii, “Classification of human hand movements using surface EMG for myoelectric control”, Advances in Intelligent Systems and Computing, vol. 513, pp. 331-339, 2016.

[9] H. Binali, C. Wu, and V. Potdar, “Computational approaches for emotion detection in text”, in 4th IEEE International Conference on Digital Ecosystems and Technologies, pp. 172-177, 2010.

[10] Z. Teng, F. Ren, and S. Kuroiwa, “Emotion recognition from text based on the rough set theory and the support vector machines”, in International Conference on Natural Language Processing and Knowledge Engineering, pp. 36-41, 2007.

DOI: 10.48175/568