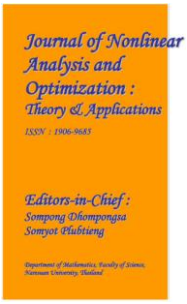
Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 15, Issue. 1, No.3 : 2024
ISSN : **1906-9685**



OPINION MINING ON CUSTOMER REVIEWS THROUGH CAT-CUK BASED FEATURE SELECTION AND NB-SVM BASED CLASSIFICATION

Dr M. Malathi, Department of Computer Science, Assistant Professor, Nallamuthu Gounder Mahalingam College, Pollachi, Coimbatore, Tamilnadu.

1. INTRODUCTION

Today's world is greatly dependent on the online life. Conventionally it was the "Word of Mouth" that decides the endurance of a product life cycle even though smart promotional strategies helps but only in initial stages of the product. That is slowly taken over by Customer Reviews of products in online shops and e-commerce websites. The potential consumers are greatly influenced by the reviews from the other customers. A new user bases is not only about his or her emotion on a product studying the reviews but also learns uses or alternate uses or hacks about the products from the experiences of existing customers. Studying such impact is also known as sentiment analysis, which is a form of National Language Processing (NLP) and computational linguistics technique for detecting and extracting subjective information from source material.

The reviews by the consumer may be positive, negative or neutral, and Opinion Mining aims to determine the opinions, beliefs and conclusions of the writer towards the product. The challenge of acquiring, analyzing and concluding from the huge repository of structured and unstructured data available online is a huge task but not impossible. To enable that, Opinion Mining is systematized into three phases. The first phase of pre-processing stage is about planning how and where cleaning process takes place. The second phase of feature extraction. Third Phase is Classification. This research work concentrates on developing a new feature extraction algorithms and new classification algorithms based on Cuckoo Search, Cat Swarm Optimization, mRMR, NB (Naive Bayes Classification) and SVM (Support Vector Machine) algorithms which is used to extract Amazon product review data.

2. LITERATURE SURVEY

In the past few years, a great attention has been received by web documents as a new source of individual opinions and experience. This situation is producing increasing interest in methods for automatically extracting and analyzing individual opinion from web documents such as customer reviews, weblogs and comments on news. Oumaima Oueslati, Erik Cambria et al., carried out an in-depth qualitative study of the most important research works in this context by discussing strengths and limitations of existing approaches. In particular, they made a survey on both approaches that leverage machine translation or transfer learning to adapt English resources to Arabic and approaches that stem directly from the Arabic language.

Ali Hasan, Sana Moin et al., contributed the adoption of a hybrid approach that involves a sentiment analyzer that includes machine learning. Moreover, this paper also provides a comparison of techniques of sentiment analysis in the analysis of political views by applying supervised machine-learning algorithms such as Naïve Bayes and support vector machines (SVM). Muhammad Taimoor Khan, MehrDurrani et al., reviewed a sentiment analysis techniques and highlight the need to address natural language processing (NLP) specific open challenges. Without resolving the complex NLP challenges, ML techniques cannot make considerable advancements. The open issues and challenges in the area are discussed, stressing on the need of standard datasets and evaluation methodology.

Dipti Sharma and Munish Sabharwal proposed a hybrid feature selection which is a combination of Particle swarm optimization (PSO) and cuckoo search. Due to the subjective nature of social media reviews, hybrid feature selection technique outperforms the traditional technique. The performance

factors like f-measure, recall, precision, and accuracy tested on twitter dataset using Support Vector Machine (SVM) classifier and compared with convolution neural network. Experimental results of this paper on the basis of different parameters show that the proposed work outperforms the existing work.

3. OBJECTIVES OF THE STUDY

- To study and analyze various feature extraction and classification techniques for opinion mining.
- To find out an effective solution to enhance the performance of classification.
- To analyze Cat Swarm Optimization algorithm, Cuckoo Search algorithm, mRMR algorithm for feature extraction, and SVM (Support Vector Machine) algorithm, NB (Naïve Bayes) algorithm for classification.
- To propose new feature extraction algorithm CAT-CUK using existing algorithms such as Cat Swarm Optimization algorithm, Cuckoo search algorithm, mRMR algorithm and hybridized algorithms M-CAT, M-Cukoo.
- To propose new classification algorithm called NB-SVM using existing classification algorithms such as NB and SVM.
- To compare and analyze the results of M-CAT and NB based opinion mining, M-CAT and SVM based opinion mining, M-Cuckoo and NB based opinion mining, M-Cuckoo and SVM based opinion mining, CAT-CUK and NB based opinion mining, CAT-CUK and SVM based opinion mining and CAT-CUK and NB-SVM based opinion mining to decide which is the best combination to provide better result.
- To prove that the combination of CAT-CUK based feature extraction and NB-SVM based classification performs better accuracy on opinion mining.
- To test and examine the efficiency of proposed new algorithms such as M-CAT, M-Cuckoo, CAT-CUK and NB-SVM on Amazon customer product data set.

4. PROPOSED RESEARCH

Opinion mining and sentiment analysis, the words that are used interchangeably these days is a field of text data mining that involves extracting opinions from evaluative texts and classifying the polarity of the opinion as positive, neutral or negative based on the orientation of the text results after the computational treatment of opinions expressed towards the main features. Opinion mining process is a three phase process. First phase is Pre-processing where cleaning process takes place. In this phase various techniques used for pre-processing such as Text Tokenization, Spelling Check, Stop Words Removal, Punctuation Removal, Word Stemming, and Letter Replacement.

Second phase is feature extraction phase, which is very important phase for better result. In this research work M-CAT, M-Cuckoo and CAT-CUK algorithms are proposed for feature extraction based on mRMR (minimum redundancy and maximum relevance) algorithm, Cuckoo algorithm Cat Swarm Optimization algorithm. The minimum redundancy and maximum relevance (MRMR) based feature selection algorithm iteratively selects features that are maximally important to the prediction task and minimally redundant with the collection of features already selected, unlike univariate feature selection methods that return a subset of features without accounting for redundancy between the selected features.Cuckoo is a parasite of the best known brood. Cat swarm optimization algorithm is inspired by cats' habits of resting and tracing. It seems that cats are lazy and spend much of their time sleeping. However, their knowledge is very strong during their rest and they are very conscious of what is happening around them. So, intelligently and intentionally, they are continually watching the world and when they see a target, they begin to move quickly towards it.

In third phase classification process takes place. In this research work Support Vector Machine, Naive Bayes Classification algorithms are used to propose new classification algorithm called NB-SVM.The Support Vector Machine (SVM) is a collection of directed learning techniques used for regression classification and analysis that analyzes data and recognizes patterns.The Naive Bayes Classifier is a well-known supervised method of machine learning. It is Thomas Bayes' probabilistic classifier. This technique of classification assumes that the presence or non-existence of any feature in the file is independent of any other feature's existence or non-appearance. The classifier of Naïve Bayes claims

that a file is a bag of words and believes that the possibility of a word in the file is independent of its position in the file and the existence of another word.

5. FRAMEWORK AND RESEARCH METHODOLOGY

Opinion mining works on three phases. First phase performs pre-processing work on Amazon Product Review dataset. Text Tokenization, Spelling Check, Stop Words Removal, Punctuation Removal, Word Stemming, Letter Replacement are the pre-processing techniques. Text Tokenization is the process of breaking a string sequence into pieces called tokens, such as words, keywords, phrases, symbols and other components. Tokens may be words, phrases or even whole sentences that are individual. Some characters, such as punctuation marks, are discarded in the process of tokenization. In NLP text processing, removing stop words is a crucial phase. It comprises deleting high-frequency keywords like to, at, for, is, and others from a sentence that contribute little or no semantic significance. Word stemming is a technique which is used to extract the root of a given word, it is emphasized. For the information retrieval, the current variety in word morphology is a major concern. In addition to having words of the same origin, duplicity of words is avoided. The improvement in results is due to the stemming process. Because stemming is in charge of managing this variance, the dictionary and storage space are both decreased. Stemmers are algorithms that are aware of stemming. Other than English, there are different types for certain letters. The multiple forms of each of these letters have therefore been substituted by some previous research into the default type. Most punctuation marks, such as commas and full stops, are not helpful for polarity detection. The process of removing these known as Punctuation Removal.Data coming from social media contains many spelling mistakes. This step removes those spelling mistake for better result.

The output of this first phase is taken as an input for second phase. Second phase is a feature extraction phase. In this phase the proposed algorithms such as M-Cuckoo algorithm, M-CAT algorithm, CAT-CUK algorithm performs feature extraction process and provides result. The result shows that the drawback of Cuckoo Search algorithm solved by M-Cuckoo and the drawback of Cat Swarm Optimization algorithm solved by M-CAT and finally proved that the result of CAT-CUK performs better compared with M-Cuckoo and M-CAT.

Then output of this second phase is taken as an input for third phase. Third phase is a classification phase. M-CAT, M-Cuckoo and CAT-CUK perform feature extraction process. NB, SVM and NB-SVM perform classification process. In this phase M-CAT and NB based opinion mining, M-CAT and SVM based opinion mining, M-Cuckoo and NB based opinion mining, M-Cuckoo and SVM based opinion mining, CAT-CUK and NB based opinion mining, CAT-CUK and SVM based opinion mining, CAT-CUK and NB-SVM based opinion mining are performed. Finally, when compared the performance of all these result, this research work learned that CAT-CUK and NB-SVM provides better result on various factors compared to all combinations.

6. EXPERIMENTAL RESULTS AND ANALYSIS

This research work experienced with Amazon product review dataset and is implemented in ORIGIN software. Our proposed algorithm works on second and third phase in opinion mining. In first phase, dataset is implemented on various pre-processing techniques. Then in second phase dataset's features have been extracted using M-CAT, M-Cuckoo and CAT-CUK algorithms. Among these feature extraction process CAT-CUK provides best result. The output of this second phase is implemented for classification in third phase. NB, SVM, NB-SVM are proposed classification algorithms. Among these three classification algorithms, NB-SVM provides best result.

This research work experimented with 7 combinations of feature extraction and classification algorithms for opinion mining such as M-CAT and NB based opinion mining, M-CAT and SVM based opinion mining, M-Cuckoo and NB based opinion mining, M-Cuckoo and SVM based opinion mining, CAT-CUK and NB based opinion mining, CAT-CUK and SVM based opinion mining and CAT-CUK and NB-SVM based opinion mining. Finally, these experiments reveal that CAT-CUK and NB-SVM based opinion mining depicts far better performance on various factors such as f-measure, recall, precision, and accuracy compared with all other 6 combinations of feature extraction and classification algorithms.

7. CONCLUSION

Opinion Mining or Sentiment Analysis is a task in the processing of natural language to find the customers' mood about buying a specific product or subject. It involves developing a framework in many online shopping sites to gather and review opinions about the product made. Opinion mining is a sub-field of the mining of web content. Opinions are statements that reflect the opinion or sentiment of individuals. Therefore this research proposes a hybrid feature extraction and classification algorithm for effective opinion mining. This research work proposed new feature extraction algorithm called CAT-CUK using existing algorithms such as Cuckoo search algorithm, CAT swarm optimization algorithm, mRMR algorithm and proposed algorithms M-CAT, M-Cuckoo. Also proposed new classification algorithm called NB-SVM using existing classification algorithms such as NB and SVM. The performance factors like f-measure, recall, precision, and accuracy tested on Amazon product review dataset using all combinations M-CAT based NB, M-CAT based SVM, M-Cuckoo based NB, M-Cuckoo based SVM, CAT-CUK based NB, CAT-CUK based SVM and CAT-CUK based NB-SVM. So finally the research work concluded that CAT-CUK based feature extraction and NB-SVM based classification provides best result for opinion mining based on above given factors.

8. REFERENCES

[1]OumaimaOueslati, Erik Cambria et al., "*A review of sentiment analysis research in Arabic language*", Elsevier, Future Generation Computer Systems, 112, pp. 408-430, 2020.

[2]G. Dharani Devi, Dr. S .Kamalakkannan, "*Literature Review on Sentiment Analysis in Social Media: Open Challenges toward Applications*", International Journal of Advanced Science and Technology, Vol. 29, No. 7, pp. 1462-1471, 2020.

[3]Harshit Sanwal, Sanjana Kukreja, "*Design Approach for Opinion Mining in Hotel Review using SVM With Particle Swarm Optimization (PSO)*", International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 8 Issue 09, pp. 519-522, Sept-2019.

[4]Ali Hasan, Sana Moin et al., "*Machine Learning-Based Sentiment Analysis for Twitter Accounts*", Mathematical and Computational Applications, 23,11, pp. 1-15, 2018.

[5]Dipti Sharma, Munish Sabharwal, "*Sentiment Analysis for Social Media using SVM Classifier of Machine Learning*", International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278-3075, Vol.8, Issue.9S4, July 2019.

[6]Harshal Thakre, Vaibhav Pate et l., "*Sentiment Analysis Using Fusion Cuckoo Search Technique for Social Media Text*", International Journal of Advanced Technology & Engineering Research (IJATER), ISSN No: 2250-3536, pp.1-12, Vol. 8, Issue 6, Nov. 2018.