International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 08 | August - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

A Review on Data Driven Approaches for Opinion Mining and Sentiment Analysis

Priyanka Yadav[1] Prof. Ashish Tiwati[2]

***Abstract*- Due to the emergence of social media becoming a common platform for sharing of views and day to day activities, more researches are aimed at extracting data out of social media data mining, one subset of which is text mining of social media data such as twitter, facebook, Whatsapp etc. Due the enormous amount of data available with the websites, they become a natural choice for text mining and/or opinion mining. This paper presents the necessity for text cum opinion mining based Sentiment analysis and its various associated techniques. It is expected that the paper will pave a path for future researchers to carry forward their research in a direction which best suits their application**

Keywords- Opinion Mining, Text Mining, Opinion Mining, Clustering, Artificial Intelligence (AI), Machine Learning (ML)

I. INTRODUCTION

Since Tim Berners Lee proposed the Web for the very first time in 1989, it has been continually evolving towards the current Social Web or Web 2.0. In the Social Web, users have become both content consumers and producers. These contents, either inherited from static Web 1.0 pages or user-generated in the Web 2.0, may range from merely plain text documents to more complex Web resources which present some kind of structure. While traditional Web pages were static interlinked hypertext documents which basically contained raw text or multimedia files, new Web documents are substantively different from prior Web ones, allowing all users to freely contribute. This new scenario presents some interesting points:

- Dynamic and permanently updated content enriches the user experience.
- Information flows bidirectional between site owners and site users by means of evaluation, review, and commenting.
- Site users may add content for others to see, comment, modify and improve (crowd sourcing).
- Web 2.0 sites develop APIs to allow automated usage (e.g. by individual apps or by sites that gather data from different resources and build an aggregated mash up).
- One of the most important key features of the Web 2.0 is that users have the ability to collectively classify information by means of tags, developing free taxonomies of information called folksonomies.

Data mining is an application of data processing in which expert patterns and information is extracted. This extracted information is consumed using applications and actual time programs for making choices.

II. PREVIOUS WORK

The section enunciates and exemplifies the different approaches for using mined data for sentiment analysis and their applications.

Obiedat et al. proposed a hybrid approach by combining the Support Vector Machine (SVM) algorithm with Particle Swarm Optimization (PSO) and different oversampling techniques to handle the imbalanced data problem. SVM is applied as a machine learning classification technique to predict the sentiments of reviews by optimizing the dataset, which contains different reviews of several restaurants in Jordan. Data were collected from Jeeran, a well-known social network for Arabic reviews. A PSO technique is used to optimize the weights of the features, as well as four different oversampling techniques, namely, the Synthetic Minority Oversampling Technique (SMOTE), SVM-SMOTE, Adaptive Synthetic Sampling (ADASYN) and borderline-SMOTE were examined to produce an optimized dataset and solve the imbalanced problem of the dataset.

Shehu et al. proposed three data augmentation techniques namely Shift, Shuffle, and Hybrid to increase the size of the training data; and then we use three key types of deep learning (DL) models namely recurrent neural network (RNN), convolution neural network (CNN), and hierarchical attention network (HAN) to classify the stemmed Turkish Twitter data for sentiment analysis. The performance of these DL models has been compared with the existing traditional machine learning (TML) models. The performance of TML models has been affected negatively by the stemmed data, but the

performance of DL models has been improved greatly with the utilization of the augmentation techniques. Based on the simulation, experimental, and statistical results analysis deeming identical datasets, it has been concluded that the TML models outperform the DL models with respect to both training-time (TTM) and runtime (RTM) complexities of the algorithms; but the DL models outperform the TML models with respect to the most important performance factors as well as the average performance rankings.

M. Estrada et al. presented a comparison among several sentiment analysis classifiers using three different techniques – machine learning, deep learning, and an evolutionary approach called EvoMSA – for the classification of educational opinions in an Intelligent Learning Environment called ILE-Java. Authors develop two corpora of expressions into the programming languages domain, which reflect the emotional state of students regarding teachers, exams, homework, and academic projects, among others. A corpus called sentiTEXT has polarity (positive and negative) labels, while a corpus called eduSERE has positive and negative learning-centered emotions (engaged, excited, bored, and frustrated) labels

D. Dhou et al. analyzed two important subtasks in this field, stance detection and product aspect mining, both of which can be formalized as the problem of the triple (target, aspect, opinion) extraction. In this paper, we first introduce the general framework of opinion mining and describe the evaluation metrics. Then, the methodologies for stance detection on different sources, such as online forum and social media are discussed. After that, approaches for product aspect mining are categorized into three main groups which are corpus level aspect extraction, corpus level aspect, and opinion mining, and document level aspect and opinion mining based on the processing units and tasks

Lijuan Zheng et al. proposed that while several data sources are available on the internet to be mined, yet a judicious use of web mining is to be done prior to any system design model is to be used. The critical factor is also the feature selection from the raw data to be included in the analysis of the data as a whole. The accuracy of the system was evaluated to evaluate the performance of the system.

Jitendra Kumar Rout et al. proposed that unstructured text mining approach is often used and the text is to be replaced with suitable tokens or numerical counterparts prior to training any designed mechanism for the classification of the text data. While data as a whole can consist of more than textual data, hence pre-processing of the data is of topmost priority. This approach could outperform simple token based approaches used previously.

Asha S Manek et al. used the Support Vector Machine (SVM) classifier as a tool for the prediction of movie success based on the sentiment analysis classification of social media data. The technique used the use of feature extraction of user database for providing ratings to movies. The user reviews were fed to the SVM as the hyper-plane concept of the SVM was used for the final classification.

Md Rakibul Islam et al. focussed on automated classification of sentiment based classification can be leveraged in several applications which need an automated mechanism for sentiment classification. The major challenge in this section is the proper training of the automated system as the training accuracy would yield high classification accuracy later.

Hassan Saif et al. proposed the use of SentiWord technique which would yield significantly close results of user opinions that are based on sentiments exhibited by users using platforms such as social media applications, blogs etc. This data can subsequently become quintessential in manoeuvring the changes in fields affected by it. The accuracy attained was 85%.

Duyu Tang et al. proposed the domain where such applications can be used are sentiment embedding systems where sentiment data can be embedded onto smart human machine interfaces that would behave in a similar fashion replicating the human nature. This would lead to more sophisticated AI based platforms replacing human intervention altogether in applications where the data is complex and cumbersome for humans to analyze with high accuracy.

Xing Fang et al. used the concept of sentiment analysis for product review. The product review was based on the sentiment analysis classification of social media data. The technique used the use of feature extraction of user database for providing rating for the products by different users. The sentiment data was used to train a network which would predict the review score of a product.

Basant Agarwal et al. proposed a Dependency-Based Semantic Parsing for level based sentiment analysis. This approach used the sentiment parsing approach wherein the sentiment data (often in the form of text data) was parsed to attain intelligible data to the use of sentiment analysis. The approach proposed that unstructured text mining approach is often used and the text is to be replaced with suitable parsed sub-tokens or numerical counterparts prior to training any designed mechanism for the classification of the text data.

Emitza Guzman et al. proposed a fine grained sentiment analysis paradigm for app reviews. The authors cited the fact that the data to be mined is often unintelligible for actual use and the need for feature extraction was of utmost importance.

The authors showed that data-cleaning could enhance the accuracy with which the reviews could be predicted. The application of such a system can be used for multiple purposes and in this case, app review was to be used.

Geetika Gautam et al. proposed the use of machine learning for the analysis of twitter data. The authors showed that designing some training mechanism for the extracted data which would yield high accuracy of classification.First and foremost, the machine or artificial intelligence system requires training for the given categories. Subsequently, the neural network model needs to act as an effective classifier.

Erik Cambria et al. proposed the use of opinion mining for sentiment analysis purposes. The opinions mined directly from user databases were used for subsequent sentiment analysis. The technique used the use of feature extraction of user database for providing rating for the products by different users. The sentiment data was used to train a network which would predict the sentiment context from the mined opinion data set used to train a network.

Bjorn Schuller et al. proposed Knowledge-Based Approaches to Concept-Level Sentiment Analysis. The approach used the knowledge discovery approach. The approach was simple enough to extract features pertaining to the training of neural systems. The authors cited the fact that the data to be mined is often unintelligible for actual use and the need for feature extraction was of utmost importance and formed the knowledge (feature) based information discovery.

Hassan Saif et al. proposed the semantic analysis of twitter data for sentiment analysis. The authors showed that it is often difficult to estimate the context in which the statements are made. Words in textual data such as tweets can be used in different contexts leading to completely divergent meanings. The automated classification of sentiment based classification can be leveraged in several applications which need an automated mechanism for sentiment classification.

Bing Liu et al. presented a survey on opinion mining and sentiment analysis. The paper puts forth the various approaches that have been proposed and tested by different authors for sentiment analysis using opinion mining. It was shown that the sentiment extraction of users from large and complex data sets is however daunting. This is to be ensured that the context (semantics) is to be taken into account prior to reaching conclusions and implicit meaning has to be inferred correctly.

Alena Neviarouskaya et al. proposed a Secure SentiFul: A Lexicon for Sentiment Analysis for sentiment analysis. The authors showed that a training algorithm doesn't work upon textual data directly to find some pattern. It needs to be fed with numerical substitutes. Hence it becomes mandatory to replace the textual information with numerical information so as to facilitate the learning process of the numerical data processing system.

Jorge Carrillo de Albornoz et al. proposed a Feature Mining and Sentiment Analysis for Product Review Ratings mechanism. This approach was based on feature mining from raw data. It used the concept of sentiment analysis for product review. The product review was based on the sentiment analysis classification of social media data. The technique used the use of feature extraction of user database for providing rating for the products by different users.

Gang Li et al. proposed a clustering based approach for sentiment analysis. This approach was based on the clustering or segregation of data sets based on similarity index that was to be used further for the sentiment analysis of data. The clustering approach makes forms a cluster of similar sentiment data feed that is used to train a network that is often used as a standard model for subsequent classification. The fundamental challenge though is the logic behind the clustering approach that is to be used for the final classification of data.

Wei Wang et al. proposed a Semi-supervised topic sentiment mixture model that was to be used for product review purposes. The approach was not fully supervised in approach resembling a partial supervised learning mechanism like the particle swarm optimization or the student teacher based learning optimization problems. The approach was however based on expert view based weight adaptation. The authors cited the fact that the data to be mined is often unintelligible for actual use and the need for feature extraction was of utmost importance. The authors showed that data-cleaning could enhance the accuracy with which the product reviews could be predicted. The application of such a system can be used for multiple purposes and in this case product review was the chosen application.

Sentiment Analysis concept has been one of the most researched work areas for various researches since a long time. Different methodologies have been incorporated by different authors for making the analysis very accurate and robust. The machine learning approach has been considered to a great extent in many domains and it has shown good results. The various previous works of different authors have been discussed and explained in depth and their working methods put forward. Instances have been drawn from their work and also certain areas have been identified where better work can be carried out for improvement.

Motivation for Problem Domain

Sentiment Analysis has been one of the most interesting and widely used areas of research till data. But this body of work and all previous works in this domain has faced considerable challenges in terms of improved performance with regards to accuracy and low time complexity. Hence the motivation has been to design an effective and efficient system for the sentiment analysis based on Bayesian Classifier approach for improved performance and results

III. ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING USED FOR OPINION MINING ANALYSIS

Although different mechanisms can be utilized for text mining and subsequent sentiment analysis, the most effective technique that is coming up front is the use of artificial intelligence and machine learning for the classification of sentiment based data yielding to sentiment analysis. Artificial Intelligence is formally defined as the development of computer systems that would perform tasks generally needing human intervention. Such a system generally exhibits the following attributes:

1) Accepting data in a parallel manner.

2) Analyzing data

3) Finding regularities or patterns in the data

4) Producing an output based on the above steps.

The process is often referred to as machine learning wherein the machine is not explicitly programmed for a certain task due to the complexity of the data to be handled and the non-predictive nature of the data to be handled. Such systems mimic the human nature of performing the tasks at the outset with a randomly chosen mechanism and gradually adapting to the changes that yield lesser errors in the output. There are several challenges in sentiment analysis from textual data. The challenges can be stated as follows:

Contextual Analysis

It is often difficult to estimate the context in which the statements are made. Words in textual data such as tweets can be used in different contexts leading to completely divergent meanings.

Frequency Analysis

Often words in textual data (for example tweets) are repeated such as
##I feel so so so happy today!!
In this case, the repetition of the word is used to emphasize upon the importance of the word. In other words, it increases to its weight. However, such rules are not explicit and do not follow any regular mathematical formulation because of which it is often difficult to get to the actuality of the tweet.

Converting textual data into numerically weighted data
The biggest challenge in using an ANN based classifier is the fact that the any ANN structure with a training algorithm doesn't work upon textual data directly to find some pattern. It needs to be fed with numerical substitutes. Hence it becomes mandatory to replace the textual information with numerical information so as to facilitate the learning process of the neural network.

Challenges in Existing Systems
1) Data mining of relevant data which is exhaustive in nature so as to cover most of the cognitive parameters of tweets [2].
2) Making data suitable for analysis by requisite and effective pre-processing.
3) Extracting semantic parts from whole tweets which would in turn reduce the dimensionality of the enormous data size for training based on tokenization [8].
4) Attaining low time complexity in training an exhaustive set of data.
5) Designing some training mechanism for the extracted data which would yield high accuracy of classification. First and foremost, the machine or artificial intelligence system requires training for the given categories [10]. Subsequently, the neural network model needs to act as an effective classifier. The major challenges here the fact that sentiment relevant data vary significantly in their parameter values due to the fact that the parameters for each building is different and hence it becomes extremely difficult for the designed neural network to find a relation among such highly fluctuating parameters. Generally, the Artificial Neural Networks model's accuracy depends on the training phase to solve new problems, since the Artificial Neural Networks is an information processing paradigm that learns from its environment to adjust its weights through an iterative process [19]. The main challenge or shortcoming is to design the Artificial Neural Networks structure using a training algorithm that is:
a) Stable: The inference is the fact that using such an algorithm, the errors should monotonically decrease.
b) Fast: The algorithm should not have excess time complexity. To overcome this shortcoming, evolutionary algorithms are to be used to adopt the search algorithm to evolve the Artificial Neural Networks (ANN) connection weights, learning rules, architectures or the input features [10]. Moreover accurate feature extraction and structuring is necessary to train the Artificial Neural Networks (ANN) accurately.

IV. PERFORMANCE PARAMETERS

The performance parameters for evaluation of the performance of an algorithm, for sentiment analysis are:

Accuracy: It is defined as:

$$Ac = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Recall: It is mathematically defined as:

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

Precision: It is mathematically defined as:

$$Precisiosn = \frac{TP}{TP+FP} \quad (3)$$

F-Measure: It is mathematically defined as:

$$F - Measure = \frac{2.Precision.Recall}{Precision+Recall} \quad (4)$$

Here.
TP represents true positive
TN represents true negative
FP represents false positive
FN represents false negative

V. CONCLUSION

It can be concluded that sentiment analysis has emerged as a field which can have diverse applications in various fields such as banking, stock pricing, politics, social media, advertising, academics etc. While several techniques are available for social media text mining, but Artificial Intelligence and Machine Learning have emerged as the most effective techniques for sentiment analysis of social media data. Also the performance metrics should be kept in mind while designing of any sentiment classification technique.

References

[1] R. Obiedat et al., "Sentiment Analysis of Customers' Reviews Using a Hybrid Evolutionary SVM-Based Approach in an Imbalanced Data Distribution," in IEEE Access, 2022, vol. 10, pp. 22260-22273.

[2] H. A. Shehu et al., "Deep Sentiment Analysis: A Case Study on Stemmed Turkish Twitter Data," in IEEE Access, 2021, vol. 9, pp. 56836-56854.

[3] MLB Estrada, RZ Cabada, RO Bustillos, "Opinion mining and emotion recognition applied to learning environments", Journal of Expert Systems, Elsevier 2020.

[4] R. Wang, D. Zhou, M. Jiang, J. Si and Y. Yang, "A Survey on Opinion Mining: From Stance to Product Aspect," in IEEE Access, vol. 7, pp. 41101-41124, 2019

[5] Lijuan Zheng,Hongwei Wang ,Song Gao," Sentimental feature selection for sentiment analysis of Chinese online reviews", SPRINGER 2018

[6] Jitendra Kumar Rout,Kim-Kwang Raymond Choo,Amiya Kumar Dash,Sambit Bakshi ,Sanjay Kumar Jena,Karen L. Williams," A model for sentiment and emotion analysis of unstructured social media text", SPRINGER 2018

[7] Asha S Manek ,P Deepa Shenoy,M Chandra Mohan,Venugopal K R, "Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier", SPRINGER 2017

[8] Md Rakibul Islam ; Minhaz F. Zibran, , "Leveraging Automated Sentiment Analysis in Software Engineering", IEEE 2017

[9] Hassan Saif, Yulan He, Miriam Fernandez, Harith Alani, "Contextual Semantics for Sentiment Analysis for Twitter", Elsevier 2016

[10] Duyu Tang , Furu Wei , Bing Qin ,Nan Yang ,Ting Liu, Ming Zhou," Sentiment Embeddings with Applications to Sentiment Analysis", IEEE 2016

[11] Xing Fang ,Justin Zhan," Sentiment analysis using product review data",SPRINGER 2015

[12] Basant Agarwal ,Soujanya Poria,Namita Mittal,Alexander Gelbukh,Amir Hussain, "Concept-Level Sentiment Analysis with Dependency-Based Semantic Parsing: A Novel Approach", SPRINGER 2015

[13] Emitza Guzman ; Walid Maalej," How Do Users Like This Feature? A Fine Grained Sentiment Analysis of App Reviews", IEEE 2014

[14] Geetika Gautam ; Divakar Yadav," Sentiment analysis of twitter data using machine learning approaches and semantic analysis", IEEE 2014

[15] Erik Cambria , Björn Schuller ,Yunqing Xia ,Catherine Havasi," New Avenues in Opinion Mining and Sentiment Analysis",IEEE 2013

[16] Erik Cambria ,Björn Schuller ,Bing Liu ,Haixun Wang ,Catherine Havasi, " Knowledge-Based Approaches to Concept-Level Sentiment Analysis", IEEE 2013

[17] Hassan Saif,Yulan He,Harith Alani," Semantic Sentiment Analysis of Twitter", SPRINGER 2012

[18] Bing Liu ,Lei Zhang, "A Survey of Opinion Mining and Sentiment Analysis", SPRINGER 2012

[19] Alena Neviarouskaya , Helmut Prendinger , Mitsuru Ishizuka, "Secure SentiFul: A Lexicon for Sentiment Analysis", IEEE 2011

[20] Jorge Carrillo de Albornoz, Laura Plaza, Pablo Gervás, Alberto Díaz, "A Joint Model of Feature Mining and Sentiment Analysis for Product Review Ratings", SPRINGER 2011

[21] Gang Li, Fei Liu, "A clustering-based approach on sentiment analysis", IEEE 2010

[22] Wei Wang, "Sentiment analysis of online product reviews with Semi-supervised topic sentiment mixture model", IEEE 2010

[23] Erik Boiy, Marie-Francine Moens, "A machine learning approach to sentiment analysis in multilingual Web texts", SPRINGER 2009

[24] Songbo Tan, Xueqi Cheng, Yuefen Wang, Hongbo Xu, "Adapting Naive Bayes to Domain Adaptation for Sentiment Analysis", SPRINGER 2009