Sentiment Analysis of Data in Social Networking Sites using Machine Learning Approaches

[1]***Yarasu Madhivi Latha,*** *Associate Professor, RISE Krishna Sai Gandhi Group of Institutions, Ongole,*
[2]***Lavanya Baviri,*** *Associate Professor, RISE Krishna Sai Gandhi Group of Institutions, Ongole,*
[3]***Yezarla Ashok,*** *Assistant Professor, RISE Krishna Sai Gandhi Group of Institutions, Ongole.*

ABSTRACT:

The emergence of various social networking platforms has made it incredibly easy for individuals to create, express, and share their ideas, thoughts, opinions, and emotions with a global audience. With the rapid advancement of technology, miniature computers and Smartphone have become ubiquitous, allowing people to easily convey their thoughts on social media platforms like Facebook, Twitter, Wikipedia, LinkedIn, Google+, Instagram, and more.

Thanks to the exponential growth in both population and communication technologies over the past decade, the use of social networks has surged, and they are now employed for a multitude of purposes. One promising application that has garnered attention is the analysis of users' posts to detect signs of depression.

In this paper, we explore how it is possible to gauge the level of depression in an individual by observing and extracting emotional cues from their text-based content. We achieve this through the utilization of emotion theories, machine learning techniques, and natural language processing tools across various social media platforms.

Keywords—Sentiment Analysis, Social Networking Sites (SNS), Depression Measurements

INTRODUCTION

Sentiment analysis (SA) can be defined as a computational approach to assess people's sentiments, evaluations, and viewpoints regarding individuals, issues, entities, topics, events, and products, as well as their attributes. Its primary objective is to automatically uncover the underlying attitudes held towards a particular entity. This analytical technique holds significant importance in comprehending user opinions, which finds valuable applications in commercial endeavors such as product reviews, political campaigns, product feedback, marketing analysis, and public relations. Moreover, it holds potential in addressing more critical concerns like security threats by monitoring discussions related to terrorism.

In the context of the Fourth Industrial Revolution (4IR), unstructured data, such as social media text, has emerged as a pivotal source of information. This data is readily available through individual user-generated content on platforms like Facebook and microblogs. However, understanding and extracting meaningful insights from text data pose formidable challenges, and no definitive solution has been proposed, neither by researchers nor by industry stakeholders.

This study centers on one of the most prevalent techniques used in SA, namely Supervised Machine Learning (SML). In recent years, SML approaches have consistently produced the best SA results. However, most of these achievements are based on in-domain datasets, where training and testing data share similar characteristics.

While a majority of research in this field focuses on enhancing SA outcomes, there is a notable gap concerning the identification of instances where the proposed SML model encounters limitations. Assessing the model's suitability when applied to real datasets, especially in real-time scenarios, is a crucial concern. It would be invaluable to know when the model's performance begins to degrade. Nonetheless, to date, there has been a lack of studies addressing this specific issue. Prior research has highlighted how model performance deteriorates when unfamiliar features are introduced into the testing dataset, often due to semantic complexities in sentiment-related words, where the same terms can represent different sentiments.

PROPOSED METHODOLOGY

There are two primary sentiment classification techniques: binary classification and multi-class sentiment classification. In binary classification, each document (denoted as di in D, where D = {d1, d2, d3 ...dn}) is categorized into one of two classes: either "Positive" or "Negative." On the other hand, in multi-class sentiment classification, each document (di) is classified into one of five categories: "Strong Positive," "Positive," "Neutral," "Negative," or "Strong Negative."

In our proposed methodology, we incorporate four key components: preprocessing, feature extraction, meta-learning, and training data. To enhance the accuracy of sentiment analysis in product reviews, we have developed the SLCABG model by combining the strengths of sentiment lexicons, Convolutional Neural Network (CNN) models, Gated Recurrent Unit (GRU) models, and attention mechanisms.

Here's a breakdown of the SLCABG model's architecture:

1. Sentiment lexicon is employed to boost the sentiment-related features within the reviews.
2. CNN and GRU networks are utilized to extract essential sentiment features and contextual information from the reviews.
3. The attention mechanism is employed to assign weights to these extracted features.
4. The weighted sentiment features are subsequently subjected to classification.

The SLCABG model comprises six layers: an embedding layer, a convolutional layer, a pooling layer, a Bidirectional GRU layer, an attention layer, and a fully connected layer. This architecture is designed to effectively capture sentiment information and improve sentiment analysis accuracy, particularly in the context of product reviews.

1.Support Vector Machine (SVM*):* SVM, a machine learning classifier, is a versatile tool applicable to both classification and regression tasks. The Single Kernel SVM is a prevalent choice for data analysis across various domains, including social media. In the realm of text classification, Linear SVM is renowned for its high-performance capabilities.

In this algorithm, we represent each data point as a distinct point in an n-dimensional space. Each feature's value corresponds to the coordinate in this space, with n denoting the number of features at our disposal.

2. *Naive Bayes (NB):* Naive Bayes is a machine learning model that finds applications in both classification and regression tasks. In various domains, including social media, the Single Kernel SVM is a frequently employed technique for data analysis.

In this algorithm, we represent each data item as a point within an n-dimensional space. In this space, the value of each feature corresponds to the value of a specific coordinate. Here, n represents the total number of features at our disposal.

P (label | features) = P (label) ∗ P (features | label) / P (features) (1)

In the equation above, P(label) represents the prior probability of the label's occurrence. It signifies the likelihood that a random feature set will belong to that specific label. This probability is calculated by considering the number of training instances with the label relative to the total number of training instances.

P(features | label) denotes the prior probability of a given feature set being classified under a particular label. This probability takes into account the association between features and labels observed in the training dataset.

P(features) represents the prior probability of a given feature set's occurrence. This probability measures the likelihood of a randomly selected feature set matching the specified feature set. It is derived from the frequency of observed feature sets in the training data.

P(label | features) informs us about the probability that the given features should be associated with a particular label. A high value for this probability indicates a reasonable level of confidence that the label is appropriate for the given set of features.

3.*Maximum Entropy:* Maximum Entropy is alternatively referred to as a conditional exponential classifier or a logistic regression classifier. This classifier operates by transforming labeled feature sets into vectors through encoding. These encoded vectors are subsequently employed to calculate weights for each feature. These feature weights are then combined to ascertain the most probable label for a given feature set..

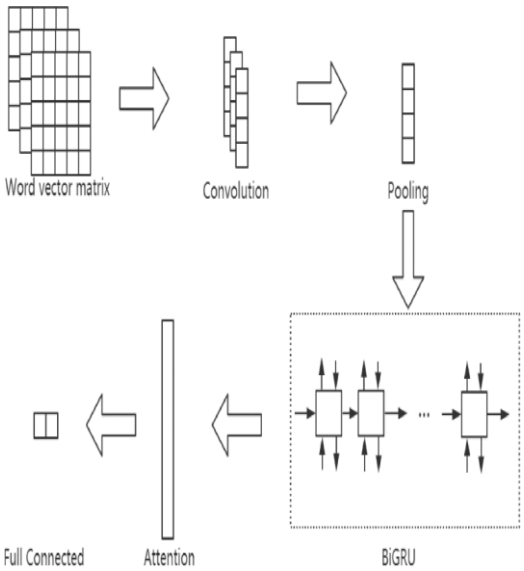

Fig: Sentiment Analysis for the Measurement of Depression.

Suppose the input text statement is $S = \{w1,w2,\cdots,wi,\cdots,wn\}$, where wi represents a word in S, and the task of our model is to predict the sentimental polarity P of the statement S.

$H(X) = -\Sigma P(X)\log 2P(X)$ (2)

$H(X)= \Sigma P(X)\log 21/P(X)$ (3)

$H(X) = E[\log 21/P(X)]$ (4)

The range of entropy is based on the number of outcomes that is $0 \leq H(X) \leq \log(n)$

PERFORMANCE EVALUATION

It gives a brief overview on the evolutionary time-line of sentimental recognition methods given by the researchers across the globe along with the datasets. There are various algorithms of machine learning such as support vector machine, naïve bayes and Entropy approach.

	Accuracy		
Classifier Name	Accuracy	Precision	Recall
Support Vector Machine	91%	83%	79%
Naïve Bayes	83%	88%	83%
Maximum Entropy	80%	84%	80%

Table: Performance Evolution Table

CONCLUSION

In this paper we have made a comparison among SVM, NB and ME classifiers regarding sentence level sentiment analysis for depression measurement. We have adopted voting model and feature selection technique. We examined the performance of our proposed methods on two datasets, twitter dataset and 20newsgroups. Our experiment indicates that SVM shows superior result as compare to Nave Bayes and Maximum Entropy classifiers. We observed that the accuracy of SVM is 91 %, the accuracy of Nave base is 83 % and the accuracy of Maximum Entropy is 80 %.

REFERENCES

[1] Banitaan, Shadi, and Kevin Daimi. ”Using data mining to predict possible future depression cases.” International Journal of Public Health Science (IJPHS) 3.4 (2014): 231-240.

[2] Yousefpour, Alireza, Roliana Ibrahim, and Haza Nuzly Abdel Hamed. ”Ordinal-based and frequency-based integration of feature selection methods for sentiment analysis.” Expert Systems with Applications 75 (2017): 80-93.

[3] Hussain, Jamil, Maqbool Ali, Hafiz Syed Muhammad Bilal, Muhammad Afzal, Hafiz Farooq Ahmad, Oresti Banos, and Sung young Lee. ”SNS based predictive model for depression.” In International Conference on Smart Homes and Health Telematics, pp. 349-354. Springer, Cham, 2015.

[4] S. Tanwar, T. Ramani, and S. Tyagi, ‘‘Dimensionality reduction using PCA and SVDin bigdata: A comparative case study, ’’in Future Internet Technologies and Trends, Z .Pateland S.Gupta, eds. Cham, Switzerland: Springer, 2018, pp. 116–125.

[5] R. S. Jadhav and P. Ghadekar, ‘‘Content based facial emotion recognition model using machine learning algorithm,’’ in Proc. Int. Conf. Adv. Comput. Telecommunication (ICACAT), Dec. 2018, pp. 1–5.

[6] A. Mollahosseini, D. Chan, and M. H. Mahoor, ‘‘Going deeper in facial expression recognition using deep neural networks,’’ in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), LakePlacid,NY,USA, Mar.2016, pp. 1–10.

[7] T. Ojala, M. Pietikainen, and T. Maenpaa, ‘‘Multi resolution gray-scale and rotation invariant texture classification with local binary patterns,’’ IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 7, pp. 971–987, Jul. 2002.

[8] R. Gupta, S. Tanwar, S. Tyagi, and N. Kumar, ‘‘Machine learning models for secure data analytics: A taxonomy and threat model,’’ Comput. Commun., vol. 153, pp. 406–440, Mar. 2020.

[9] T. L. Paine, P. Khorrami, W. Han, and T. S. Huang, ‘‘an analysis of unsupervised pre-training in light of recent advances,’’ 2014, arXiv: 1412.6597. Available: https://arxiv.org/abs/1412.6597