UI Solution for Semantic Analysis using Machine learning: A comparative study

Arnav Sankhe
Dept. of physics and astronomy
University of Nottingham
Nottingham, England
ppxas5@nottingham.ac.uk

I. Introduction

"Semantics deals with the meaning of sentences and words as fundamentals in the world. Semantic analysis within the framework of natural language processing evaluates and represents human language and analyzes texts written in the English language and other natural languages with interpretations similar to those of human beings" [1]. "Sentiment analysis refers to the management of sentiments, opinions, and subjective text. The demand for sentiment analysis is raised due to the requirement of analyzing and structuring hidden information, extracted from social media in form of unstructured data" [2] The study of identifying people's emotions in written material, such as a product, service, or organisation evaluation, is known as semantic analysis. It also entails obtaining viewpoints based on user comments on YouTube or Twitter. Identifying these emotions in product reviews, tweets, and comments can lead to insights that can provide a company or individual with a competitive advantage.

This research holds all the findings of the previous work done on this topic. In the related works section, we are going to look at all the work done in the field of machine learning for semantic analysis, the work mostly consists of machine learning algorithms and their approach. This section also consists of an extended literature review on deep learning for semantic analysis. In the summary and prospect section, we look at hyperparameter tuning methods, more machine learning algorithms and ensemble approaches.

II. Related works

Sentiment Analysis research has expanded to include document-level classification, thanks to advances in Natural Language Processing (NLP) [3] to the classification of words and phrases [4]. Hatzivassiloglou and McKeown provided a method for retrieving semantic information from a big corpus. When the corpus is replaced, this approach isolates domain-dependent information and conforms to a new domain. Their technique focuses on adjectives, with the goal of identifying near-synonyms and antonyms.

For improving the model's efficiency and accuracy [5] For sentiment analysis, the ensemble framework was employed. They used movie reviews as well as multi-domain datasets derived from Amazon product reviews, such as reviews of books, electronics, DVDs, and kitchen appliances. By integrating several categorization algorithms and feature sets, they were able to frame the ensemble. To create the ensemble framework, They employed three types of classifiers: maximum entropy, Support Vector Machines, and Naive Bayes, as well as two types of feature sets: word-relations and part-of-speech information. Sentiment analysis was improved with the use of weighted combination, fixed combination, and meta-classifier ensemble approaches [5].

On social networking sites, people express their opinions on anything and everything. Recognizing various forms of data for training proved difficult. As a result, [6] a model was suggested to investigate sentiment using the hashtagged (HASH) data set, the iSieve data set, and the emoticon (EMOT) dataset. The authors used a range of feature extraction strategies to train their model, including lexical features, part-of-speech (POS) features, n-gram features, and microblogging features. They discovered that the POS function may not be useful in the world of microblogging, and that the Emoticon dataset's benefits are also lessened when microblogging aspects are included [6].

The paper's authors examined [7] social network analysis and how Twitter is a rich source for sentiment analysis, and they presented a methodology for implementing Twitter sentiment analysis using Twitter APIs. Their research is based on a variety of job search inquiries. Positive, negative, and neutral labels are included in the dataset. They found that neutral feelings are higher than positive or negative attitudes, indicating that Twitter sentiment analysis needs to be improved [7]. In the field of politics, Twitter has grown in popularity. The incumbent, former President Barack Obama, and the nine other opponents have been subjected to a real-time emotion analyzer [8]. They employed IBM's InfoSphere Streams platform (IBM, 2012) to route real-time data with speed and precision. They used the Twitter "firehouse" to find relevant tweets on candidates and events by creating logical keyword combinations. They had a 59 percent accuracy rate [8].

Some scholars have used Twitter messages to try to figure out what the public thinks about various topics such

as politics, movies, and news [9]. The paper's authors [10] used IMDB, a prominent Internet database with movie information, and Blippr, a 124 social networking site where users post reviews in the form of 'Iblips.' Using SVM, they were able to achieve an F-score of 0.9 and establish domain adaption as a promising technique for sentiment analysis. They developed the Relative Information Index (RII), a revolutionary feature reduction strategy that combines with another common technique, 'thresholding,' to provide a feature reduction technique that not only reduces features but also improves the F-score.[10].

TThe importance of sentiment analysis has grown to the point where it is now used in a variety of industries, including hotel management. In this sense, [11] divided public hotel reviews into positive and negative categories. They gathered 800 TripAdvisor reviews and used NLTK in Python to accomplish the preprocessing. Logistic Regression, Nave Bayes, Stochastic Gradient Descent Classifier, Random Forest, and Support Vector Machine were among the classifiers utilised.

According to their findings, the Nave Bays classifier performed the best, while the Stochastic Gradient classifier also performed well. The accuracy, recall, precision, and F1-score [11] findings were used in the analysis. [12] a paper with 94.4 accuracy that uses sentiment analysis and Linear SVC and Nave Bayes approaches to classify business reviews. We use Nave Bayes, which has a lower time complexity than other algorithms. If the number of features exceeds the number of samples, avoid over-fitting when selecting features.

n this study, [13] all the decision trees in an RF are combined to generate the RF prediction. To compute relevance scores between a prediction and all target classes, they use TFIDF weights of the terms in the semantic explanations. To determine its dependability, the class with the best relevance score is compared to the predicted class. The results of the experiments on 30 text datasets demonstrated that present DES techniques fail to improve RF's text classification performance, whereas (ii) the suggested method statistically considerably outperforms both traditional RF and existing DES methods [13].

Advanced text categorization systems based on back propagation neural networks are proposed in this research [14]. MBPNN overcomes the standard BPNN's slow training speed problem and can escape the local minimum. LSA not only dramatically reduces the dimension, However, it also overcomes the drawbacks of the commonly used vector space model method for text representation.

III. Summary and prospects

A. hyperparameter tuning

The hyperparameters of the RF randomness are controlled by the parameters mtry, sample size, and node size. They should be configured so that the separate trees have a reasonable strength without too much association between them (bias-variance trade-off). According to the literature and our own research, mtry is the most influential of these characteristics [15].

According to the publication [16], EDAs are the best methods for hyper-parameter adjusting SVM classifiers. It's vital to remember that the rest of the algorithms' performance is determined by the specific values of their user-defined parameters. These algorithms were scored much lower than those with one or no user-defined parameter while employing the best-reported values. Furthermore, using an optimization approach that involves more parameter settings than the issue dimensionality should be avoided.

B. Machine Learning Algorithms

This paper is about [17] SVM is treated to a unigram model, which generates a superior result than using it alone, and the nave byes technique outperformed the maximum entropy technique. When the semantic analysis WordNet is used in conjunction with the above method, the accuracy rises to 89.9 percent from 88.2 percent . The training data set, as well as WordNet for review summary, can be enlarged to improve the feature vector-related phrase recognition process. [17].

In this research [18] machine learning techniques were combined with a dictionary-based methodology under the lexicon-based Approach. Every product evaluation was subjected to sentiment analysis before being categorised using machine learning techniques such as NB and SVM. For Camera evaluations, the Naive Bayes classifier had 98.17 percent accuracy while the Support Vector Machine had 93.54 percent accuracy.

C. Ensemble approach

In this research [19] an ensemble strategy for automatic summary assessment that combines two of the most effective assessment techniques, LSA and n-gram co-occurrence is explained. There has also been a performance comparison of the suggested ensemble approach with other current techniques. In comparison to the best current technique, BLEU, which has an overall accuracy of 87 percent, the proposed approach has attained an overall accuracy of 96 percent [19].

The proposed ensemble method in this research [20] is more scalable and practical. Although experimental support limits the suggested strategy, the results suggest that machine learning has limitless potential in medical research.

References

[1] Salloum, S.A., Khan, R. and Shaalan, K. (2020) "A Survey of Semantic Analysis Approaches," in, pp. 61–70. doi:10.1007/978-3-030-44289-7_6.

[2] Ain, Q.T., Ali, M., Riaz, A., Noureen, A., Kamran, M., Hayat, B. and Rehman, A., 2017. Sentiment analysis using deep learning techniques: a review. Int J Adv Comput Sci Appl, 8(6), p.424.

[3] Pang, B. and Lee, L. (2008) "Opinion Mining and Sentiment Analysis," Foundations and Trends® in Information Retrieval, 2(1–2), pp. 1–135. doi:10.1561/1500000011

[4] Hatzivassiloglou, V. and McKeown, K.R. (1997) "Predicting the semantic orientation of adjectives," in Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics -. Morristown, NJ, USA: Association for Computational Linguistics, pp. 174–181. doi:10.3115/979617.979640.

[5] Xia, R., Zong, C. and Li, S. (2011) "Ensemble of feature sets and classification algorithms for sentiment classification," Information Sciences, 181(6), pp. 1138–1152. doi:10.1016/j.ins.2010.11.023

[6] Kouloumpis, E., Wilson, T. and Moore, J., 2011. Twitter sentiment analysis: The good the bad and the omg!. In Proceedings of the international AAAI conference on web and social media (Vol. 5, No. 1, pp. 538-541

[7] "Baweja, A. and Garg, P., 2019. Sentimental Analysis of Twitter Data for Job Opportunities. International Research Journal of Engineering and Technology (IRJET), 6(11).]

[8] Kamvar, S.D. and Harris, J. (2011) "We feel fine and searching the emotional web," in Proceedings of the fourth ACM international conference on Web search and data mining - WSDM '11. New York, New York, USA: ACM Press, p. 117. doi:10.1145/1935826.1935854.

[9] Neethu, M.S. and Rajasree, R. (2013) "Sentiment analysis in twitter using machine learning techniques," in 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT). IEEE, pp. 1–5. doi:10.1109/ICCCNT.2013.6726818.

[10] "Peddinti, V.M.K. and Chintalapoodi, P., 2011, August. Domain adaptation in sentiment analysis of twitter. In Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence.

[11] Gupte, A., Joshi, S., Gadgul, P., Kadam, A. and Gupte, A., 2014. Comparative study of classification algorithms used in sentiment analysis. International Journal of Computer Science and Information Technologies, 5(5), pp.6261-6264.

[12] Salinca, A., 2015, September. Business reviews classification using sentiment analysis. In 2015 17th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC) (pp. 247-250). IEEE.

[13] Islam, M.Z. et al. (2019a) "A Semantics Aware Random Forest for Text Classification," in Proceedings of the 28th ACM International Conference on Information and Knowledge Management. New York, NY, USA: ACM, pp. 1061–1070. doi:10.1145/3357384.3357891.

[14] Wang, W. and Yu, B. (2009) "Text categorization based on combination of modified back propagation neural network and latent semantic analysis," Neural Computing and Applications, 18(8), pp. 875–881. doi:10.1007/s00521-008-0193-3.

[15] Probst, P., Wright, M.N. and Boulesteix, A. (2019) "Hyperparameters and tuning strategies for random forest," WIREs Data Mining and Knowledge Discovery, 9(3). doi:10.1002/widm.1301.

[16] Rojas-Dominguez, A. et al. (2018) "Optimal Hyper-Parameter Tuning of SVM Classifiers With Application to Medical Diagnosis," IEEE Access, 6, pp. 7164–7176. doi:10.1109/ACCESS.2017.2779794.

[17] Gautam, G. and Yadav, D. (2014) "Sentiment analysis of twitter data using machine learning approaches and semantic analysis," in 2014 Seventh International Conference on Contemporary Computing (IC3). IEEE, pp. 437–442. doi:10.1109/IC3.2014.6897213.

[18] Jagdale, R.S., Shirsat, V.S. and Deshmukh, S.N. (2019) "Sentiment Analysis on Product Reviews Using Machine Learning Techniques," in, pp. 639–647. doi:10.1007/978-981-13-0617-4_61.

[19] He, Y., Hui, S.C. and Quan, T.T. (2009) "An Ensemble Approach for Semantic Assessment of Summary Writings," in 2009 International Conference on Computer Engineering and Technology. IEEE, pp. 490–494. doi:10.1109/ICCET.2009.90.

[20] Lu, J. et al. (2020) "Machine learning for assisting cervical cancer diagnosis: An ensemble approach," Future Generation Computer Systems, 106, pp. 199–205. doi:10.1016/j.future.2019.12.033.