A DRUG RECOMMENDATION SYSTEM BASED ON SENTIMENT ANALYSIS OF DRUG REVIEWS USING MACHINE LEARNING

[1]K.TEJENDHAR GOUD, [2]ATHAR KHAN, [3]SHAIK AKHIL AHMED, [4]P.VINAYREDDY
[1, 2, 3, 4]Department of Computer Science and Engineering, [4]AssociateProfessor
[1, 2, 3, 4]Vijay Rural Engineering College, Manik Bhandar, Nizamabad-503003

Abstract: Predicting drug responses is a crucial difficulty in computational personalized medicine. a large number of machine learning methodologies, particularly those grounded on deep studying, had been proposed for this activity. Although, those techniques regularly describe pharmaceuticals as strings, which is not a natural representation of molecules. Moreover, the translation of factors like as mutations or replica number aberrations influencing remedy response has now not been comprehensively addressed. Procedures: this text introduces a unique approach, Graph DRP, utilizing a graph convolutional network to address the hassle. In Graph DRP, medicinal drugs were shown as molecular graphs that directly represent atomic bonds, even as cell traces have been characterized as binary vectors of genomic abnormalities. Convolutional layers learnt the consultant traits of medications and cellular traces, which have been sooner or later incorporated to symbolize every drug-cell line pair. The reaction fee for each drug-cell line pair was ultimately predicted by a fully connected neural network. 4 types of graph convolutional networks were employed to analyze the traits of prescribed drugs. Findings: Graph DRP surpasses tCNNS throughout all performance metrics in every trial performed. In conclusion, modeling prescribed drugs as graphs can enhance the efficacy of drugs response prediction.

"Index Terms - *Graph convolutional networks, drug response prediction, molecular graphs, genomic aberrations, machine learning, sentiment analysis, drug reviews".*

1. INTRODUCTION

The exponential increase of user-generated content on internet platforms has led users to depend greater on digital reviews for their decision-making procedures. This phenomena encompasses the healthcare enterprise, as consumers and healthcare specialists actively seek data from drug opinions to assess medicinal drug efficacy, destructive effects, and standard user happiness. The substantial quantity and diversity of these tests pose a big barrier in identifying meaningful and constant findings. Therefore, there's an increasing call for automated structures which could efficiently extract pertinent patterns and attitudes from this huge dataset.

A Drug recommendation system (DRS) utilising sentiment analysis and device gaining knowledge of presents a feasible alternative. Sentiment analysis, an vital subfield of natural Language Processing (NLP), allows algorithms to discern subjective information in textual information via recognizing high quality, poor, or neutral emotions conveyed in user reviews [1]. While applied in drug reviews, it serves as a powerful tool to gauge the general perspective and happiness of consumers approximately certain treatments. This facilitates medicinal drug recommendations based on public sentiment and assists physicians in making evidence-based totally judgments.

The incorporation of machine learning into this framework enables the popularity of critical language characteristics and styles that affect sentiment classification. Numerous machine learning models, which include logistic regression, decision trees, and deep learning architectures including LSTM, had been utilized to categorise sentiment with large accuracy [2][3]. The Drug review Dataset [4] is a vital aid for training and verifying those models, consisting of comprehensive, real-world review data from websites like drugs.com.

Prior research imply that fitness-related picks are gradually fashioned via virtual sources. A research via Fox and Duggan [5][7] found out that a full-size phase of the public consults online health forums prior to making medical decisions. This transition emphasizes the necessity to expand state-of-the-art algorithms which could extract relevant data from public boards and assessment systems.

Moreover, previous models such as GalenOWL [9] and probabilistic issue mining methodologies [8] sought to analyze pharmacological evaluations for the goal of hints. In spite of their efficacy, those models often exhibited deficiencies in personalizing and interpretability. Current breakthroughs, such as intelligent medication recommender structures and ontology-primarily based methodologies, display that integrating domain expertise with statistics-driven strategies can improve recommendation precision.

Consequently, a sentiment-driven DRS personalizes prescription recommendations while reducing risks linked to subjective interpretation mistakes via patients and physicians [6]. This approach utilizes a strong integration of NLP, machine learning, and sentiment analysis to decorate the accessibility, precision, and customization of drug recommendations in contemporary telemedicine and virtual health structures [10][11].

2. RELATED WORK

A range of state-of-the-art medical recommendation systems has been developed to resource sufferers and healthcare specialists in making educated remedy picks. Bao and Jiang [12] proposed a framework for a smart medicine recommendation system, emphasizing that the mixing of clinical facts and advice algorithms can also decorate remedy planning and patient care. The gadget exhibited the functionality of data mining and user profiling in improving the customization of prescription recommendations. Zhang et al. [13] delivered CADRE, a cloud-based medicinal drug recommendation tool for on-line pharmacies, similarly developing the belief of smart assistance. This idea employs cloud computing and a distributed structure to enhance the supply and scalability of pharmaceutical recommendation offerings, providing a user-friendly answer for remote and real-time medicine recommendations.

In the domain of social media and microblogs, deep learning methodologies were investigated to model user-generated content for recommendation targets. Li et al. [14] introduced a twitter modeling technique using long short-term memory (LSTM) networks for hashtag suggestion, demonstrating the potential of recurrent neural networks to seize temporal relationships and semantic contexts in short-text data. Zhang et al. [15] tested the bag-of-words (BoW) model from a statistical viewpoint, clarifying its fundamental importance in lots of text categorization and records retrieval endeavors. However its simplicity, the Bag-of-words model is still a well-known technique for text representation, regularly functioning as a benchmark for more state-of-the-art models.

Strategies for feature extraction and textual content vectorization, such as TF-IDF, have proved vital in sentiment evaluation and records retrieval. “Ramos et al. [16] employed TF-IDF to check word relevance in document searches”, highlighting its efficacy in keeping apart key phrases and minimizing noise in textual data. Conversely, Word2Vec, a neural embedding approach, affords a greater sophisticated comprehension of word semantics. Goldberg and Levy [17] provide a complete elucidation of the terrible-sampling approach hired in Word2Vec, demonstrating how this technique generates dense vector representations that encapsulate word semantics and syntactic hyperlinks. Bollegala et al. [18] more desirable the usage of word embeddings by means of introducing a technique for unsupervised cross-domain word representation learning, which improves the generalization of sentiment analysis models across various datasets and domains.

TextBlob, a Python module, is appreciably utilized for sentiment analysis attributable to its ease of use and effectiveness. The TextBlob API offers pre-educated sentiment classifiers and linguistic tools, rendering it suitable for initial exploratory data analysis in natural language processing tasks [19]. To enhance comprehension and visualization of high-dimensional data, such word embeddings or emotion ratings, van der Maaten and Hinton [20] developed the “t-distributed stochastic neighbor embedding (t-SNE) method”. This approach enables the visualization of clusters and linkages within data, improving the comprehension of model results.

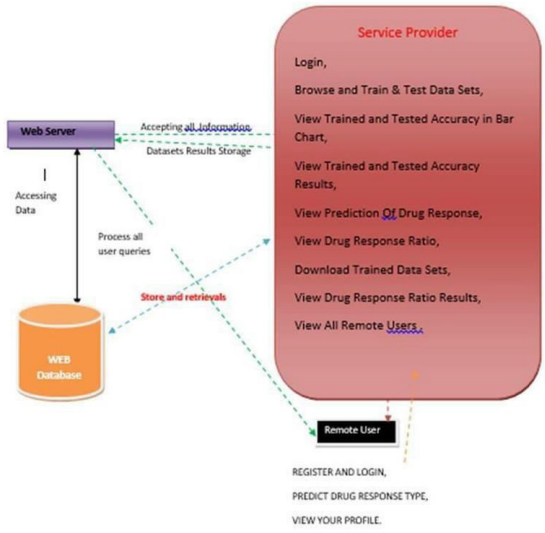
Class imbalance is a sizeable issue in medication review datasets, on account that the quantity of evaluations conveying sturdy feelings may be disproportionately limited in assessment to neutral ones. Chawla et al. [21] delivered the synthetic Minority Over-sampling technique (SMOTE) to mitigate this problem by developing synthetic samples of minority classes, thereby enhancing classifier efficacy on unbalanced datasets. Powers [22] underscored the necessity of making use of whole measures in model assessment, extending past conventional accuracy, recall, and F1-score. He presented measures like informedness, markedness, and ROC evaluation, enhancing comprehension of type model performance and efficacy.

Wang et al. [23] enhanced data resampling techniques by combining SMOTE with Tomek linkages to expand the SMOTETomek set of rules. This hybrid technique gets rid of overlapping statistics points and improves the representation of the minority magnificence, yielding cleaner and more distinguishable statistics for training system learning models. This technique has verified efficacy in areas consisting of personality detection and indicates ability for enhancing sentiment categorization in pharmaceutical evaluation datasets.

The amalgamation of these strategies establishes the idea for cutting-edge drug recommendation systems propelled by using sentiment analysis. Using improvements in NLP, machine learning, and data visualization, researchers and developers can also create systems that examine sizeable drug-related textual information, figure sentiment tendencies, and bring tailor-made pointers. The tested literature highlights the significance of multidisciplinary techniques, integrating linguistic, statistical, and computational viewpoints to address the intricacies and form of healthcare statistics. With the growing quantity of user-generated remedy reviews, these technology are essential for extracting massive insights that facilitate more informed and efficient decision-making inside the pharmaceutical sector.

3. MATERIALS AND METHODS

The proposed system seeks to establish a sentiment-driven drug recommendation framework utilising machine learning algorithms, which includes decision Tree, Random forest, Gradient Boosting, okay-Nearest neighbors (KNN), Logistic Regression, and support Vector machine (SVM), to analyze huge patient reviews. Utilising sentiment analysis methodologies and function engineering, the device identifies significant styles from user comments to suggest tailor-made drugs in line with particular situations. This method mitigates the deficiency of healthcare practitioners, specifically in remote regions, and diminishes medical inaccuracies stemming from prescription errors. It facilitates informed drug selection by discovering efficacious treatments with less unwanted results, hence boosting patient protection [6]. Employing patient-generated records now not simplest equips purchasers with clean medication insights however also aids healthcare practitioners in making data-driven judgments. The system is flexible to new information and scalable for full-size use, rendering it an invaluable asset in modern telemedicine and digital healthcare settings.



"Fig.1 Architecture"

This discern delineates a mechanism for forecasting remedy response. A web server gets data and retains datasets and results in a web database. It manages person inquiries and allows data access for storage and retrieval. The service issuer consists of functions including login, browsing and training drug response datasets, visualizing accuracy, examining prediction effects and ratios, downloading statistics, and coping with distant users. Remote users can register, log in, forecast drug reaction types, and check their profiles, presumably engaging with the service provider through the web server to make use of the drug response prediction functionalities.

i) Dataset Collection:

This research use the Drug review Dataset (drugs.com), received from the UCI machine learning Repository [12]. The dataset consists of 215,063 entries and encompasses six attributes: medication name, patient review, medical condition, beneficial count (denoting the number of users who considered the review useful), overview access date, and a "10-star rating" representing ordinary patient satisfaction. This dataset is critical for sentiment evaluation and comparing remedy effectiveness, supplying empirical insights from patients. The entire framework enables machine learning packages in personalised medication recommendation systems and aids in identifying side effects and evaluating treatment efficacy.

ii) Data Cleaning & Visualization:

This studies used many data cleaning methods to assure quality and relevance. Initially, all extraneous text and rows with missing or null values—specially the 1,200 null entries in the "condition" column—have been eliminated. Duplicate data have been eradicated by using confirming the originality of the review IDs. Entries with inconsequential or

pointless values have been removed, ensuing in a final dataset of 212,141 rows. Visualization tools had been applied to study critical components, which includes the distribution of null values, the identification of the 20 maximum established illnesses with reachable medicinal drugs, and the analysis of "the 10-star rating system". The rating distribution exhibited a sentiment polarity, with the majority of ratings clustered at values of 10, 9, 1, and 8, signifying stated user reviews.

iii) Feature Selection:

Both automatic and manual techniques have been utilized for feature selection to improve model performance. Word2Vec generated semantic-rich word embeddings by transforming words into excessive-dimensional vectors that encapsulate contextual similarities. In assessment to TF and TF-IDF, Word2Vec maintains semantic links by positioning related words in proximity inside vector space. Along automatic vectorization, human feature engineering was utilized to enhance forecast accuracy. Fifteen characteristics have been retrieved, comprising numerical values consisting of usable count, encoded categorical data from the condition column, and temporal variables (day, month, 12 months) derived from the review date. TextBlob was employed to calculate sentiment polarity ratings from both sanitized and unsanitized critiques as supplementary capabilities.

iv) Train & Test:

Four separate datasets had been generated for model training and evaluation using Bag of words (BoW), TF-IDF, Word2Vec, and manual feature extraction methodologies. The dataset was partitioned into education and testing subsets, adhering to a 75% to 25% ratio, respectively. a set random state was applied during the splitting process to guarantee consistency and repeatability across all datasets. This facilitated the era of a same collection of random values, so assuring the consistency of training and testing samples for each feature set. This consistency facilitated an equitable assessment of model efficacy across the diverse feature extraction methodologies hired in the look at.

v) Algorithms:

A decision tree is a supervised learning technique that partitions data into subsets according to feature values, creating a tree-like structure in which each node signifies a choice based on a certain characteristic. It is extensively utilized for classification and regression assignments, as it streamlines intricate decision-making procedures. The objective of a decision tree is to construct a model capable of making predictions based on identified patterns within the data. Nonetheless, it may overfit to noisy data, necessitating trimming to enhance generalization.

Gradient boosting is an ensemble learning technique that integrates several weak learners, usually decision trees, to formulate a more robust prediction model by concentrating on the faults of preceding learners. It is typically utilized for regression and classification problems. Gradient boosting aims to reduce errors by iterative enhancement, where each model rectifies the defects of its predecessor. This approach enhances predictive accuracy but may incur high computing costs and is susceptible to overfitting without adequate regularization.

K-Nearest Neighbors (KNN) is a non-parametric, instance-based technique employed for classification and regression tasks. The method operates by identifying the 'k' nearest training samples to a test instance and generating predictions based on the predominant class (for classification) or the mean (for regression) of these neighbors [10]. KNN is straightforward and efficient, particularly

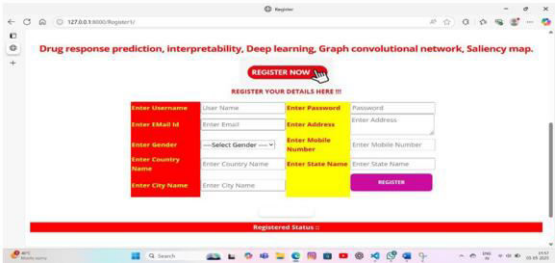
for smaller datasets; nevertheless, it gets computationally intensive as the dataset size escalates. The main objective is to categorize data points according to their closeness to other data points inside the feature space.

Logistic Regression is a statistical technique for binary classification that estimates the likelihood of a class label based on input characteristics using a logistic function [11]. It is extensively utilized for binary classification tasks, such as forecasting the presence or absence of a disease based on patient data. Logistic regression aims to quantify the association between input factors and a binary outcome, providing a straightforward yet robust method for modeling linear decision limits in classification tasks.

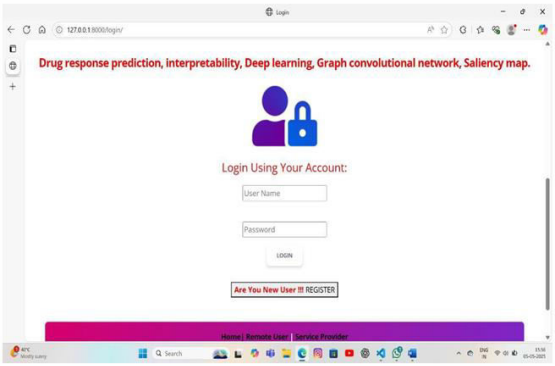
Random Forest is an ensemble technique that constructs several decision trees and amalgamates their predictions via voting (for classification) or averaging (for regression) to enhance accuracy and mitigate overfitting [12]. It is frequently employed for classification and regression problems. The main objective of random forest is to enhance model resilience by the incorporation of randomization in tree building, hence mitigating the effects of overfitting linked to an individual decision tree. It is adaptable and has high performance across many data formats.

Support Vector Machine (SVM) is a supervised technique used for classification and regression purpose. Support Vector Machine (SVM) works to identify the hyperplane that best divides the data points in diverse classes in a high dimensional feature space [13]. It is the goal of SVM to maximize the margin between classes therefore ensuring that members of data sets are correctly classified. It is cost effective in high dimension spaces and for jobs which have non linear decision boundaries, while computationally demanding when recording extensive documents.

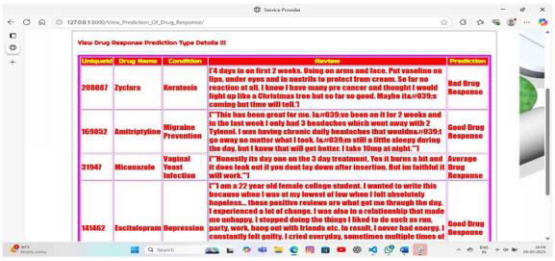
4. RESULTS AND DISCUSSION



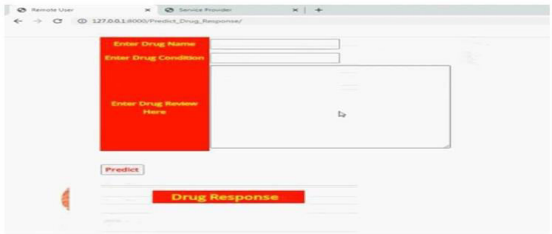
"Fig.2 Registration Page"



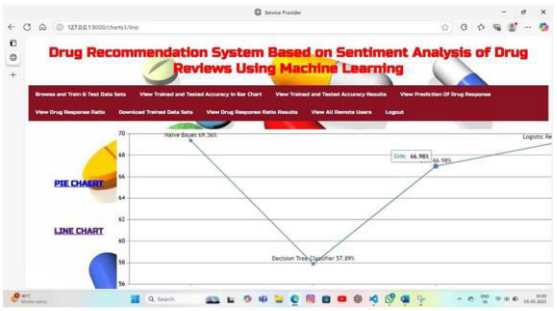
"Fig.3 Login Page"



"Fig.4 Dataset of Drug Response"



"Fig.5 Input"



"Fig.6 Accuracy of Models"

"Fig.7 Ratio of Responses"

5. CONCLUSION

Reviews have come to be an important issue of our regular lives; whether purchasing, making online purchases, or eating at a restaurant, we first consult critiques to make informed selections. This research investigates sentiment analysis of drug critiques to expand a recommender system using various system learning classifiers, together with "Logistic Regression, Perceptron, Multinomial Naive Bayes, Ridge classifier, Stochastic Gradient Descent, and LinearSVC, carried out to Bag of words (BoW) and term Frequency-Inverse document Frequency (TF-IDF)". Additionally, classifiers including decision Tree, Random forest, LightGBM, and CatBoost were employed the usage of Word2Vec and manual feature extraction strategies. We assessed them using 5 distinct metrics: "precision, recall, F1 score, accuracy, and AUC score", which indicate that the "Linear SVC on TF-IDF surpasses all other models with an accuracy of 93%". Conversely, the "decision Tree classifier with Word2Vec had the worst performance, achieving just 78% accuracy". We incorporated the most effective predicted emotion values from every method: "Perceptron on Bag of words (91%), LinearSVC on TF-IDF (93%), LGBM on Word2Vec (91%), and Random forest on manual features (88%)", and elevated them by the normalized beneficial count to derive the overall rating of the drug by using situation for the reason of building a recommender system. Future endeavors will entail the assessment of various oversampling techniques, the usage of various n-gram values, and the development of algorithms to enhance the efficacy of the recommender system.

REFERENCES

[1] Telemedicine, https://www.mohfw.gov.in/pdf/Telemedicine.pdf

[2] Wittich CM, Burkle CM, Lanier WL. Medication errors: an overview for clinicians. Mayo Clin Proc. 2014 Aug;89(8):1116-25.

[3] CHEN, M. R., & WANG, H. F. (2013). The reason and prevention of hospital medication errors. Practical Journal of Clinical Medicine, 4.

[4]DrugReviewDataset,https://archive.ics.uci.edu/ml/datasets/Drug%2BReview%2BDataset%2B%2528Drugs.com%2529#

[5] Fox, Susannah, and Maeve Duggan. "Health online 2013. 2013." URL: http://pewinternet.org/Reports/2013/Health-online.aspx

[6] Bartlett JG, Dowell SF, Mandell LA, File TM Jr, Musher DM, Fine MJ. Practice guidelines for the management of community-acquired pneumonia in adults. Infectious Diseases Society of America. Clin Infect Dis. 2000 Aug;31(2):347-82. doi: 10.1086/313954. Epub 2000 Sep 7. PMID: 10987697; PMCID: PMC7109923.

[7] Fox, Susannah & Duggan, Maeve. (2012). Health Online 2013. Pew Research Internet Project Report.

[8] T. N. Tekade and M. Emmanuel, ”Probabilistic aspect mining approach for interpretation and evaluation of drug reviews,” 2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES), Paralakhemundi, 2016, pp. 1471- 1476, doi:10.1109/SCOPES.2016.7955684.

[9] Doulaverakis, C., Nikolaidis, G., Kleontas, A. et al. GalenOWL: Ontology-based drug recommendations discovery. J Biomed Semant 3,14 (2012). https://doi.org/10.1186/2041-1480-3- 14

[10] Leilei Sun, Chuanren Liu, Chonghui Guo, Hui Xiong, and YanmingXie. 2016. Data-driven Automatic Treatment Regimen Development and Recommendation. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’16). Association for Computing Machinery, New York, NY, USA, 1865–1874.
DOI:https://doi.org/10.1145/2939672.2939866

[11] V. Goel, A. K. Gupta and N. Kumar, ”Sentiment Analysis of Multilingual Twitter Data using Natural Language Processing,” 2018 8th International Conference on Communication Systems and Network Technologies (CSNT), Bhopal, India, 2018, pp. 208-212, doi:10.1109/CSNT.2018.8820254.

[12] Y. Bao and X. Jiang, ”An intelligent medicine recommender system framework,” 2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA), Hefei, 2016, pp. 1383-1388, doi:10.1109/ICIEA.2016.7603801.

[13] Zhang, Yin & Zhang, Dafang & Hassan, Mohammad & Alamri, Atif & Peng, Limei. (2014). CADRE: Cloud-Assisted Drug Recommendation Service for Online Pharmacies. Mobile Networks and Applications. 20.348-355. 10.1007/s11036-014-0537-4.

[14] J. Li, H. Xu, X. He, J. Deng and X. Sun, ”Tweet modeling with LSTM recurrent neural networks for hashtag recommendation,” 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, 2016, pp. 1570-1577, doi: 10.1109/IJCNN.2016.7727385.

[15] Zhang, Yin & Jin, Rong & Zhou, Zhi-Hua. (2010). Understanding bag-of-words model: A statistical framework. International Journal of Machine Learning and Cybernetics. 1. 43-52. 10.1007/s13042-010-0001-0.

[16] J. Ramos et al., “Using tf-idf to determine word relevance in document queries,” in Proceedings of the first instructional conference on machine learning, vol. 242, pp. 133–142, Piscataway, NJ, 2003.

[17] Yoav Goldberg and Omer Levy. word2vec Explained: deriving Mikolov et al.’s negative-sampling word-embedding method, 2014;arXiv:1402.3722.

[18] Danushka Bollegala, Takanori Maehara and Kenichi Kawarabayashi. Unsupervised Cross-Domain Word Representation Learning, 2015;arXiv:1505.07184.

[19] Textblob, https://textblob.readthedocs.io/en/dev/.

[20] van der Maaten, Laurens & Hinton, Geoffrey. (2008). Viualizing data using t-SNE. Journal of Machine Learning Research. 9. 2579-2605.

[21] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique, 2011, Journal Of Artificial Intelligence Research, Volume 16, pages 321-357, 2002; arXiv:1106.1813. DOI: 10.1613/jair.953.

[22] Powers, David & Ailab,. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. J. Mach. Learn. Technol. 2. 2229-3981. 10.9735/2229- 3981

[23] Z. Wang, C. Wu, K. Zheng, X. Niu and X. Wang, ”SMOTETomek- Based Resampling for Personality Recognition,” in IEEE Access, vol. 7, pp. 129678-129689, 2019, doi: 10.1109/ACCESS.2019.2940061