J. Electrical Systems 20-5s (2024): 2834-2840

[1]Swapna M
[2]Nikitha K

Contextually Propagated Term Weight based Approach for Author Profiling: Gender, Age and Language Variety Prediction

Abstract: - Author profiling creates characterization of an author based on attributes such as age, gender, language, dialect region variety, personality and so on. In recent years it has garnered significant attention for its varies applications across forensic linguistics, marketing, cybersecurity and social media analytics. Most of the research focused on stylistic, content based and term weight measures based feature representations. We observed that context semantics is not considered in the feature representation. In this propose contextually propagated term weight measures for feature representation. We implemented SVM, Random Forest and XG Boost machine learning algorithms on those feature representations. The results demonstrated that the proposed contextually propagated term weight with inverse category frequency outperformed the existing methods..

Keywords: Machine learning, Contextual semantics, Inverse category frequency, Term weights

I. Introduction

Due to raise of social media, billions of users are interacting and sharing their daily activities on the platforms, such as Twitter and Facebook. Huge data is generated with this interaction. Most of this generated data is unstructured. The natural language processing is needed to extract the useful information from this unstructured data. It is one of the challenging tasks to extract the characteristics of the author from the data. The extracted author information can be used in several applications in forensics, security services, recommendation systems, marketing, and defense. Police can use the author profiling to identify the suspects of the crime, e-commerce companies can use author profiling for recommendation of products, personalized marketing of products. From the data it is possible to extract some attributes such as age, gender, language, personality of the user. This is known as Author Profiling. Author profiling uses various features such as style and content from the data and discovers the characteristics of the author. The punctuation, number of words and POS distribution are considered as style based features. The n-grams were used as content based feature to distinguish the author.

PAN lab organizes challenges for authorship attribution, author identification and author profiling. The challenge includes dataset in various languages. PAN 2018 dataset has been used for author profiling to predict gender from the multi-modal data. The multi-modal dataset has both text and image data. 2017 PAN dataset has only text data which is used to predict gender and language variety. Various classification based machine learning and deep learning models were used for the author profiling task. Section 2 covers the related work and proposed methodology is presented in section 3. Section 4 covers the result analysis for both PAN 2017 and 2018 dataset on text data. The conclusion of the paper is presented in the section 5.

II. Literature Survey

Author Profiling (AP) involves discerning demographic characteristics of writers such as native language, education level, gender, age, personality traits, location, occupation, and more. This is achieved by analyzing and comprehending their writing styles. PAN labs started challenges from 2013 and scholars worldwide participated and developed systems for author profiling. The traditional machine learning approaches were used hand-crafted features for classification attributes of the author. The psychological traits of the author were used and developed a tool based on word count and word categories [1]. On the chat based dataset the stop word usage, frequency of emotions and message length were used along with K-NN, Naïve Bayes classifiers [2]. Schler et al. [3] presented a multi-class real window using hyperlinks, POS tags and unigrams and predicted the gender. They could come out that male authors write about technology, politics and money where as female authors write about lifestyle. [4] examined the differences in writing styles between male and female authors within a genre-controlled corpus. Findings revealed that pronouns such as "you," "I," "her," "she," "their," "myself," "herself," and "yourself" are

[1] *Corresponding author Research Scholar Dept. of Informatics, Osmania University, Hyderabad Email: mandaswapnareddy@mail.com
[2] Dept. of Computer Science, Government, Mahatma Jyothiba Phule, Warangal, India

indicative of female authors. Conversely, determiners like "a," "that," "the," "these," and quantifiers like "one," "more," "two," and "some" were identified as indicators of male authors. [5] considered character-based, word-based, syntactic based features and analyzed the author profiles. [6] study revealed that various genders tend to use profanity associated with different hate categories. For instance, males are more likely to use language from the "disability" hate group.

Content based features were used for this task and achieved highest performance [7]-[9]. [7] proposed a model using POS n-grams and function words and the performance of the model is around 79%. Burger et al. [10] proposed word level unigram, bigram and character level 1 to 5 –g as features and it has outperformed the other models. The rank-1 team in PAN challenge of 2017 have used n-grams and character level n-grams as features and also applied Term Frequency and Inverse Document Frequency (TFIDF) as transformation to features. Other researchers used different combinations such as POS tagging, stemming, lemmatization and got mixed results [11]. The feature vector was created by different character and word n-grams with Latent Semantic Analysis (LSA) and predicted the gender using SVM [9]. Reddy et al.[22] proposed document specific term weight measure for author profiling. Kavadi et al. [12] proposed novel method called the SPW approach has been introduced. The SPW approach achieved the highest accuracies in comparison to existing methods for celebrity profiling. In the SPW approach, two Term Weight Measures (TWMs), namely IQCFTW and SUTW, are employed to calculate term weight, while the Document Weight Measure (DWM) is used to compute document weight. Adi Narayana Reddy K et al. [13] presented fusion of stylistic features and word embedding and predicted gender and age using deep learning.

The impact of neural networks in implicit feature learning is discussed in [14]-[17].The top-ranked team at PAN-2018 utilized a deep learning-based strategy for learning representations of text and images from tweets. The model based on Long Short-Term Memory (LSTM) with a fusion mechanism exhibited the most outstanding results ,The team that ranked second demonstrated that traditional machine learning-based methods continue to outperform when dealing solely with text. They employed word and character n-grams to represent tweets and utilized support vector machines (SVM) for classification purposes [9]. Utilized a bidirectional Long Short-Term Memory (LSTM) model to learn abstract-level representations of data and implemented an attention mechanism to identify the crucial features [18]. Adi Narayana Reddy K et al. [13] presented fusion of stylistic features and word embedding and predicted gender and age using deep learning. In this paper we propose contextually propagated term weights for feature representation.

III. DATASET DESCRIPTION AND METRICS

A. Dataset Description

The PAN 2015, 2016, 2017 and 2018 datasets [19,20] consisting of tweets in different languages. PAN 2015 dataset contains tweets in English, Spanish, Italian and Dutch languages. This helps in identifying Age [21-31] Gender, language variety and personality type. PAN 2016 dataset contains tweets in English, Spanish and Dutch languages. This helps in identifying Age, Gender, or language variety. PAN 2017 dataset contains tweets in English, Spanish, Portuguese and Arabic languages. This helps in identifying, Gender, or language variety. PAN 2018 dataset contains tweets in English, Spanish, and Arabic languages along with images. This helps in identifying Gender using tweets and images.

B. Evaluation Metrics

To evaluate the models researchers were used accuracy, precision, recall and f1-score. The accuracy is generally described as the ratio of correctly predicted instances to the total number of instances. It is represented in equation (1) as

Accuracy = Number of correct predictions / Total instances (1)

Precision measures the accuracy of predicted positive class and is represented in equation (2) as

Precision = True positive / (True positives + False positives) (2)

Recall is ratio of positive predictions and total actual positives and is represented in equation (3) as

Recall = True positive / (True positives + False negatives) (3)

F1-score is harmonic mean of precision and recall

F1-Score = 2 * Precision * Recall / (Precision + Recall) (4)

IV. METHODOLOGY

Authors generally use different words while writing topics in social media. The choice of words varies among different profile categories of the author. The primary objective of this proposal is to identify the terms which are contextually more significant within the profile categories. As depicted in Fig. 1 the method starts with data cleaning using preprocessing. The preprocessing removes the stop words and applies stemming. Once the unwanted data is cleaned, identify most contextually significant terms within the documents. Next compute contextually propagated term weights for each of the profile. These term weights are used to represent the document vector. Finally, the classification models such as SVM, Random Forest and XG Boost are trained on the document vector. Each profile is predicted using the trained model.

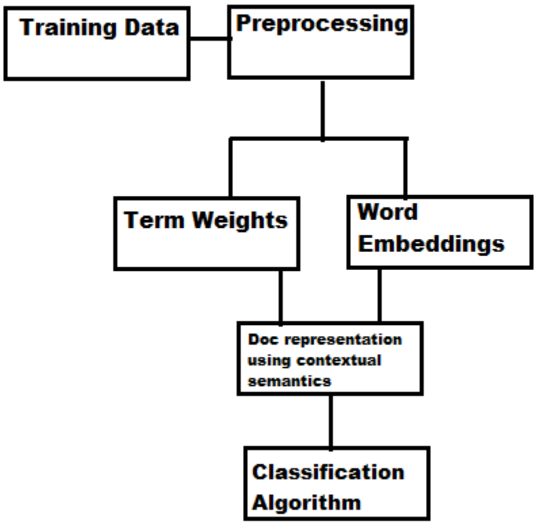

Fig. 1. Contextual Semantic model Author Profiling

A. Contextually Propagated Term Weights (CPTW)

Word embedding is a method that represents words as vectors in a continuous space. The placement of each word vector is determined by its usage and context within the dataset. This approach ensures that words with similar contexts are positioned closely together in the vector space. Contextually propagated term weights were introduced by [21]. They defined the embedded neighborhood of a term w_j in document d_i as the set of all similar terms with a cosine similarity to w_j of at least a certain threshold in the embedding space.

a. CPTW

Let N(w_j) denote the set of words present in the embedded neighborhood of the word w_j including itself. For each word $w_k \in N(w_j)$, we define contextually propagated term weight as

$$\gamma(w_k) = f(w_k, d_i) cos\,(v_j, v_k) \qquad (5)$$

Where $\gamma(w_k), f(w_k, d_i), and\ cos\,(v_j, v_k)$ are contextually propagated term weight of word w_k, frequency of word w_k in document d_i and cosine similarity of word embedding for word w_j and w_k respectively. Using contextually propagated term weight, we compute contextually propagated term weight (CPTW) of document d_i, denoted as CPTW(d_i)

$$CPTW(d_i) = \sum_{j=1}^{m} e_j (\alpha_j \sum_{w_k \in N(w_j)} \gamma(w_k)) \qquad (6)$$

Where m is the size of the vocabulary, e_j is the one-hot encoding at index j it is 1 and remaining places it is zero and α_j is the normalization constant, computed as $\alpha_j = 1/\sum_{w_k \in N(w_j)} cos\,(v_j, v_k)$.

b. $CPTW_{IDF}$

Equation (5) uses frequency of term w_k and document d_i and cosine similarity of the words w_k and w_j. The frequency of w_k in d_i increases artificially that occur frequently across the entire document collection. To handle this effect an inverse document frequency (IDF) like component is added to the contextually propagated term weight that ensures term discriminative in the data collection. It is defined as

$$\gamma_{IDF}(w_k) = \gamma(w_k) log\,(\frac{N}{df(w_k)} \alpha_j cos\,(v_j, v_k)) \qquad (7)$$

Where $\gamma(w_k)$ is computed in equation (5), N is total number of documents in the data collection, $df(w_k)$is the number of documents that contain word w_k and α_j is the normalization constant. Based on the above term weight, we define $CPTW_{IDF}$for each document as

$$CPTW_{IDF}(d_i) = \sum_{j=1}^{m} e_j(\alpha_j \sum_{w_k \in N(w_j)} \gamma_{IDF}(w_k)) \quad (8)$$

Equation (8) computes propagated IDF scores for each word in the embedded neighborhood like TF-IDF. This indicates that the IDF score of each word is determined by a weighted aggregate of the IDF scores for all the words within its embedded neighborhood.

c. $CPTW_{ICF}$

Equation (8) adds the term discrimination to collection of data. Inverse category frequency (ICF) adds discrimination power to terms across all categories.

$$\gamma_{ICF} = \gamma_{IDF}(w_k) log \left(\frac{|C|}{cf(w_k)} \alpha_j cos\,(v_j, v_k)\right) \quad (9)$$

Where γ_{IDF} is computed in equation (6), |C| is the category count in the dataset and $cf(w_k)$ is discrimination power of the term across all the categories. Based on equation (8) we define $CPTW_{ICF}$for each document as

$$CPTW_{ICF}(d_i) = \sum_{j=i}^{m} e_j(\alpha_j \sum_{w_k \in N(w_j)} \gamma_{ICF}(w_k) \quad (10)$$

Equation (10) computes propagated ICF scores for each word in the embedded neighborhood. This indicates that the ICF score of each word is determined by a weighted aggregate of the ICF scores for all the words within its embedded neighborhood.

V. EXPERIMENTS AND RESULTS

Support Vector Machines (SVM) stands as a supervised machine learning algorithm employed for classification and regression purposes. Its functionality revolves around identifying a hyperplane that effectively distinguishes data points belonging to distinct classes within a high-dimensional space. XGBoost, an abbreviation for eXtreme Gradient Boosting, is a highly potent and extensively utilized machine learning algorithm celebrated for its effectiveness and efficiency, particularly in scenarios involving structured/tabular data. Within the realm of ensemble learning, it finds its place in the gradient boosting framework. Random Forest stands out as a flexible and resilient machine learning algorithm utilized for both classification and regression assignments. Its operation involves the creation of numerous decision trees during the training phase, generating the average prediction in regression scenarios or the mode of predictions for classification tasks. The algorithm's effectiveness lies in its adoption of an ensemble learning strategy, consolidating predictions from multiple decision trees to improve overall performance and fortify resistance against overfitting.

Let consider set of N documents {D_1,D_2,…,D_N} in the dataset. The proposed CPTW computes scores for each of the document using term weights. This procedure is used for each of the task in author profiling. Initial preprocessing is applied on the dataset to extract most frequent terms. On the preprocessed data, using Word2Vec is trained extracted the word embedding for each of the word. Word embedding computes a context vector for each of the words. The CPTW, 〖CPTW〗_IDF and 〖CPTW〗_ICF measures used to determine weights of the documents.

A. Gender Prediction

We considered 1000 documents for gender sub-task. In this 500 tweets per gender is considered. Three machine learning algorithms such as Support Vector Machine (SVM), Random Forest (RF) and XGBoost (XGB) are trained on the vectors computed by CPTW, 〖CPTW〗_IDF and 〖CPTW〗_ICF. The accuracy results are presented in Table 1. The 〖CPTW〗_ICF measure outperformed the other two measures. CPTW redistributes the weight of a word to words found in similar contexts in embedding. The ICF score of each word is determined by a weighted aggregate of the ICF scores for all the words within its embedded neighborhood. The SVM algorithm demonstrated strong performance in gender prediction, outperforming both RF and XGB.

Table 1. Gender Prediction Accuracy

	SVM	RF	XGB
CPTW	90.58	89.03	90.73
$CPTW_{IDF}$	92.07	91.91	91.09
$CPTW_{ICF}$	93.76	92.46	92.62

B. Age Prediction

We considered 516 documents for Age sub-task. For prediction of age we considered three age brackets as 18-20, 23-25 and 28-61. Three machine learning algorithms such as Support Vector Machine (SVM), Random Forest (RF) and XGBoost (XGB) are trained on the vectors computed by CPTW, 〖CPTW〗_IDF and 〖CPTW〗_ICF. The accuracy results are presented in Table 2. The 〖CPTW〗_ICF measure outperformed the other two measures. The XGB algorithm demonstrated strong performance in gender prediction, outperforming both RF and SVM as it captures the non-linear relationship in the data.

Table 2. Age Prediction Accuracy

	SVM	RF	XGB
CPTW	83.9	84.05	83.39
$CPTW_{IDF}$	83.58	85.93	85.98
$CPTW_{ICF}$	87.02	86.85	87.45

C. Language Variety Prediction

We considered varieties in English language. It has six varieties from Australia, Canada, Great Britain, Ireland, New Zealand and United States. The 500 tweets for each variety. Three machine learning algorithms such as Support Vector Machine (SVM), Random Forest (RF) and XGBoost (XGB) are trained on the vectors computed by CPTW, 〖CPTW〗_IDF and 〖CPTW〗_ICF. The accuracy results are presented in Table 3. The 〖CPTW〗_ICF measure outperformed the other two measures. The RF algorithm demonstrated strong performance in gender prediction, outperforming both XGB and SVM as it is capable of capturing non-linear relationships and complex patterns in the data.

Table 3. Language Variety Prediction Accuracy

	SVM	RF	XGB
CPTW	83.79	84.38	83.90
$CPTW_{IDF}$	84.16	84.90	84.95
$CPTW_{ICF}$	84.94	85.27	85.09

VI. RESULTS ANALYSIS

In this study, the experiment was carried out using stylistic features, content based features and proposed contextually propagated term weights to predict author profiles. The Fig. 2. presents the comparative results. The accuracy of the gender sub-task stands out when compared with the accuracies for age and language variety. Content-Based Features (CBFs) demonstrate commendable accuracy levels, outshining stylistic features.

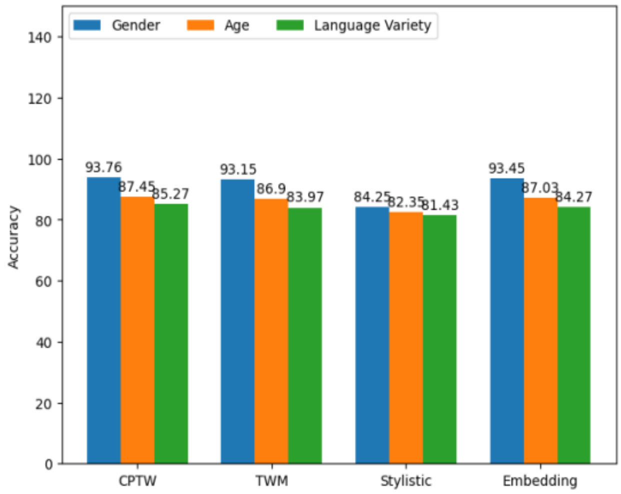

Fig. 2. Comparison of Accuracies

Stylistic features are instrumental in distinguishing the authors' writing styles, with most features hinging on the author's personality. Conversely, CBFs rely on the content found in the writings. Generally, most authors, irrespective of gender, incorporate a set of content-bearing terms in their writings, which explains the superior accuracy of CBFs over stylistic features. In the representation of document vectors, Term Frequency (TF) is employed for CBFs.

The proposed CPTW with ICF outperforms the other existing methods (Stylistic, embedding and TWM) in terms of accuracy. It identifies apt contextually propagated term weight to assign fitting weights to the terms. Unlike conventional document representations, which utilize feature weights, our novel method uses ICF scores for embedded neighborhood for representing document vectors. Moreover, the proposed method mitigates the limitations of existing techniques, establishing relationships between terms using context.

VII. CONCLUSION

Author profiling aims to discern demographic and psychological characteristics of authors from their written text. We introduced contextually propagated term weights, which redistributes the weight of a word to words found in similar context in embedding. This redistribution process serves to generalize the semantics of a text, resulting in enhanced discriminative power. It is computationally efficient. In the experiment we considered three sub-tasks such as gender, age and language variety. The proposed contextually propagated term weights with inverse category frequency outperformed the existing stylistic features, content-based and term weights.

REFERENCES

[1] F. N. Andola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size," arXiv preprint arXiv:1602.07360, 2016.

[2] R. M. Ortega-Mendoza, A. P. López-Monroy, "The winning approach for author profiling of Mexican users in Twitter at mex.a3t@ibereval-2018," in Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), CEUR WS Proceedings, 2018.

[3] R. M. Ortega-Mendoza, A. P. López-Monroy, A. Franco-Arcega, M. Montes-y Gómez, "Emphasizing personal information for author profiling: New approaches for term selection and weighting," Knowledge-Based Systems, vol. 145, pp. 169–181, 2018.

[4] L. Wendlandt, R. Mihalcea, R. L. Boyd, J. W. Pennebaker, "Multimodal analysis and prediction of latent user dimensions," in International Conference on Social Informatics, Springer, pp. 323–340, 2017.

[5] T. Mikolov, K. Chen, G. Corrado, J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.

[6] G. Farnadi, J. Tang, M. De Cock, M.-F. Moens, "User profiling through deep multimodal fusion," in Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, ACM, pp. 171–179, 2018.

[7] T. Liu, K. Cho, G. A. Broadwell, S. Shaikh, T. Strzalkowski, J. Lien, S. M. Taylor, L. Feldman, B. Yamrom, N. Webb, et al., "Automatic expansion of the MRC psycholinguistic database imageability ratings," in LREC, pp. 2800–2805, 2014.

[8] J. W. Pennebaker, M. E. Francis, R. J. Booth, "Linguistic inquiry and word count: LIWC 2001," Mahway: Lawrence Erlbaum Associates 71, 2001.

[9] F. Rangel, P. Rosso, M. Koppel, E. Stamatatos, G. Inches, "Overview of the author profiling task at PAN 2013," Notebook Papers of CLEF, pp. 23–26, 2013.

[10] F. Rangel, P. Rosso, I. Chugur, M. Potthast, M. Trenkmann, B. Stein, B. Verhoeven, W. Daelemans, "Overview of the 2nd author profiling task at PAN 2014," in Proceedings of the Conference and Labs of the Evaluation Forum (Working Notes), pp. 1–30, 2014.

[11] C. Suman, A. Naman, S. Saha, P. Bhattacharyya, "A Multimodal Author Profiling System for Tweets," in IEEE Transactions on Computational Social Systems, doi: 10.1109/TCSS.2021.3082942.

[12] D. Prasad Kavadi, et al., "A machine learning approach for celebrity profiling," International Journal of Ad Hoc and Ubiquitous Computing, vol. 38, no. 1-3, pp. 111-126, 2021.

[13] K. Adi Narayana Reddy, et al., "Fusion-Based Celebrity Profiling Using Deep Learning," in Intelligent System Design: Proceedings of INDIA 2022, Singapore: Springer Nature Singapore, 2022, pp. 107-113.

[14] S. K. Lakkaraju, D. Tech, S. Deng, "A framework for profiling prospective students in higher education," in Encyclopedia of Information Science and Technology, Fourth Edition, IGI Global, pp. 3861–3869, 2018.

[15] R. Layton, "Relative cyberattack attribution," in Automating Open Source Intelligence, Elsevier, pp. 37–60, 2016.

[16] A. Farseev, L. Nie, M. Akbari, T.-S. Chua, "Harvesting multiple sources for user profile learning: a big data study," in Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, ACM, pp. 235–242, 2015.

[17] C. P. Estruch, R. Paredes, P. Rosso, "Learning multimodal gender profile using neural networks," in Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, pp. 577–582, 2017.

[18] Bsir, M. Zrigui, "Document model with attention bidirectional recurrent network for gender identification," in Advances in Computational Intelligence, Springer, pp. 621–631, 2019.

[19] F. Rangel, P. Rosso, M. M.-Y. Gómez, M. Potthast, B. Stein, "Overview of the 6th author profiling task at PAN 2018: Multimodal gender identification in Twitter," Work. Notes Papers CLEF, vol. 2125, pp. 1–38, Feb. 2018.

[20] F. Rangel, P. Rosso, M. Potthast, B. Stein, "Overview of the 5th author profiling task at PAN 2017: Gender and language variety identification in Twitter," Working Notes Papers of the CLEF, pp. 1–26, 2017.

[21] C. Hansen, et al., "Contextually propagated term weights for document representation," in Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019.

[22] T. Raghunadha Reddy, B. Vishnu Vardhan, P. Vijayapal Reddy, "Profile specific document weighted approach using a new term weighting measure for author profiling," International Journal of Intelligent Engineering and Systems, vol. 9, no. 4, pp. 136-146, 2016.

[23] A. Mallikarjuna Reddy, V. Venkata Krishna, L. Sumalatha, "Efficient Face Recognition by Compact Symmetric Elliptical Texture Matrix (CSETM)", Jour of Adv Research in Dynamical & Control Systems, Vol. 10, 4-Regular Issue, 2018.

[24] Mallikarjuna Reddy, A., Rupa Kinnera, G., Chandrasekhara Reddy, T., Vishnu Murthy, G., et al., (2019), "Generating cancelable fingerprint template using triangular structures", Journal of Computational and Theoretical Nanoscience, Volume 16, Numbers 5-6, pp. 1951-1955(5), doi: https://doi.org/10.1166/jctn.2019.7830

[25] Mallikarjuna Reddy, A.,Venkata Krishna, V. and Sumalatha, L." Face recognition approaches: A survey" International Journal of Engineering and Technology (UAE), 4.6 Special Issue 6, volume number 7 , 117-121,2018.

[26] Swarajya Lakshmi V Papineni, Snigdha Yarlagadda, Harita Akkineni, A. Mallikarjuna Reddy. Big Data Analytics Applying the Fusion Approach of Multicriteria Decision Making with Deep Learning Algorithms International Journal of Engineering Trends and Technology, 69(1), 24-28, doi: 10.14445/22315381/IJETT-V69I1P204.

[27] Sudeepthi Govathoti, A Mallikarjuna Reddy, Deepthi Kamidi, G BalaKrishna, Sri Silpa Padmanabhuni and Pradeepini Gera, "Data Augmentation Techniques on Chilly Plants to Classify Healthy and Bacterial Blight Disease Leaves" International Journal of Advanced Computer Science and Applications(ijacsa), 13(6), 2022. http://dx.doi.org/10.14569/IJACSA.2022.0130618.

[28] Santhosh Kumar, C.N., Pavan Kumar, V., Reddy, K.S., Similarity matching of pairs of text using CACT algorithm, International Journal of Engineering and Advanced Technology, 2019, 8(6), pp. 2296–2298, DOI:10.35940/ijeat.f8685.088619

[29] Mallikarjuna A. Reddy, Sudheer K. Reddy, Santhosh C.N. Kumar, Srinivasa K. Reddy, "Leveraging bio-maximum inverse rank method for iris and palm recognition", International Journal of Biometrics, 2022 Vol.14 No.3/4, pp.421 - 438, DOI: 10.1504/IJBM.2022.10048978.

[30] Reddy, K.S., Naga Santhosh Kumar, C.H., Effective data analytics on opinion mining, International Journal of Innovative Technology and Exploring Engineering 8(10), pp. 2073-2078, 2019.

[31] A. M. Reddy, K. SubbaReddy and V. V. Krishna, "Classification of child and adulthood using GLCM based on diagonal LBP," 2015 International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), Davangere, 2015, pp. 857-861, doi: 10.1109/ICATCCT.2015.7457003.