ISSN 2277-2685
IJESR/Jan-Mar. 2025/ Vol-15/Issue-1/174-180
Jarugu Supriya Lakshmi *et. al.*, / International Journal of Engineering & Science Research

Celebrity Insight Generation Using Word Embedding

[1]**Jarugu Supriya Lakshmi**, suppu2707@gmail.com
[2]**M Naresh**, nareshmtech08@gmail.com
[1&2]Newton's Institute of Engineering, Guntur, Andhra Pradesh

Abstract
Celebrity profiling is a specialized branch of author profiling focused on identifying attributes like gender, birth year, fame, and occupation through textual analysis. Social media has become a platform for celebrities to share interests and engage with fans, but it has also led to impersonation issues. To address this, researchers are developing methods to verify whether texts are genuinely authored by celebrities and determine their profiling characteristics. In 2019, the PAN competition introduced a celebrity profiling task, challenging participants to predict celebrity attributes based on written texts. Researchers employed various stylistic features and machine learning techniques for this task. Our approach leverages word embedding techniques like Word2Vec and FastText to represent words as vectors, capturing semantic relationships. These word vectors were aggregated to create document-level representations, which were then classified using Naïve Bayes Multinomial, Support Vector Machine (SVM), and Random Forest algorithms. The results highlighted that combining Word2Vec with Random Forest achieved the highest accuracy for predicting fame and occupation, showcasing the effectiveness of advanced word embeddings and robust machine learning in celebrity profiling.

Keywords: Text Analysis, Word Embeddings, Word2Vec, FastText, Naïve Bayes Multinomial, Support Vector Machine (SVM), Random Forest

1. INTRODUCTION

Celebrities actively use platforms like Twitter and Instagram to share opinions and personal updates, sparking public interest in their demographics. **Celebrity Profiling** is a fascinating research area focused on analyzing celebrity text to predict attributes like gender, fame level, occupation, and birth year. Introduced in the 2019 PAN competition, Celebrity Profiling tasked researchers with predicting: **Gender**: Male, Female, Non-binary, **Birth Year**: 1940–2011, **Fame Level**: Rising, Star, Superstar, **Occupation**: Creator, Performer, Sports Professional, Science Professional, Manager, Religious Figure, Politician. The dataset contained 48,335 multilingual tweets (33,836 for training and the rest for testing), offering a robust foundation for profiling methods. Celebrity Profiling stems from **Author Profiling**, introduced in PAN 2013, which analyzes text to predict attributes like age, gender, location, and language.

Profiling relies on document classification, where models are trained on labeled data to categorize documents based on learned features. In Celebrity Profiling, the labels focus on fame, gender, birth year, and occupation.

2. LITERATURE SURVEY

Author profiling has gained prominence as an essential task in computational linguistics and machine learning, particularly in analyzing text to infer demographic, stylistic, and behavioral attributes. Rangel et al. (2013) [1] presented a comprehensive overview of the PAN 2013 Author Profiling task, emphasizing advancements in

profiling techniques. De-Arteaga et al. (2013) [2] contributed to this domain by leveraging corpus statistics, lexicons, and stylistic features for author characterization. Moreno-Sandoval et al. (2019) [3] introduced sociolinguistic features for celebrity profiling on Twitter during the PAN 2019 workshop. Similarly, Petrik and Chuda (2019) [4] employed TF-IDF-based profiling techniques, demonstrating the effectiveness of simple term frequency measures. Asif et al. (2019) [5] explored a word-distance approach to enhance celebrity profiling accuracy, while Martinc et al. (2019) [6] delved into the social trends of celebrities through Twitter analysis. Further innovations were proposed by Radivchev et al. (2019) [7], who applied machine learning models like Logistic Regression and Support Vector Machines (SVMs) for profiling. Pelzer (2019) [8] explored transfer learning methodologies to improve profiling accuracy. The foundational work by Mikolov et al. (2013) [9] on distributed word representations laid the groundwork for many modern NLP applications, including profiling. Bojanowski et al. (2017) [10] extended this research by incorporating subword information into word vectors, enriching linguistic feature extraction. Raghunadha Reddy et al. (2016) [11] proposed a novel term-weighting measure to enhance profile-specific document analysis, while Ali et al. (2012) [12] and Breiman (2001) [13] highlighted the efficacy of random forests and decision trees in machine learning tasks. Vapnik (2013) [14] provided a comprehensive exposition of statistical learning theory, which underpins many profiling algorithms. Kavuri and Kavitha (2020) [16] proposed a stylistic feature-based author profiling method, focusing on feature engineering. Later, Kavuri and Kavitha (2022) [17] developed a term-weight measure to improve author profiling accuracy. Surya et al. (2022) [18] emphasized language variety prediction using word embeddings and machine learning algorithms. These studies highlight the importance of feature extraction and model selection in author profiling tasks. The PAN 2019 website [15] serves as a valuable resource for the latest advancements in celebrity profiling, offering insights into ongoing research efforts. These references collectively underscore the evolution of author profiling methodologies and their applications in diverse contexts.

3. MATERIALS AND METHODS

Our approach leverages content-based features with a focus on informative words in text. Traditional methods often create document vectors directly from these words, neglecting their contextual importance. To address this, we utilized word embedding techniques, specifically **Word2Vec** and **FastText**, to represent words as contextualized vectors. Document vectors were created by aggregating the word vectors for each document. These vectors were then trained using **Naïve Bayes Multinomial**, **Support Vector Machine (SVM)**, and **Random Forest** to predict fame and occupation of celebrity authors. By capturing semantic relationships between words, this method significantly improves the accuracy of celebrity profiling.

Step 1: Pre-processing

- Remove URLs, punctuation, invalid characters, and multiple whitespaces.
- Convert text to lowercase.
- Remove stop words.
- Apply stemming to standardize word forms.

Step 2: Bag of Words (BoW)

BoW transforms text into vectors by tokenizing words and creating a unique wordlist. Example:

Corpus:

S1: "I love rain" → [1, 1, 1, 0, 0]

S2: "rain rain go away" → [0, 0, 2, 1, 1]

Each word's presence or count in the sentence is represented in the vector.

Step 3: Word2Vec Model

Word2Vec generates word embeddings by learning word relationships within context windows.

- **CBOW (Continuous Bag of Words):** Predicts a target word using the average representation of surrounding words.
- **Skip-Gram:** Predicts surrounding words using the target word.

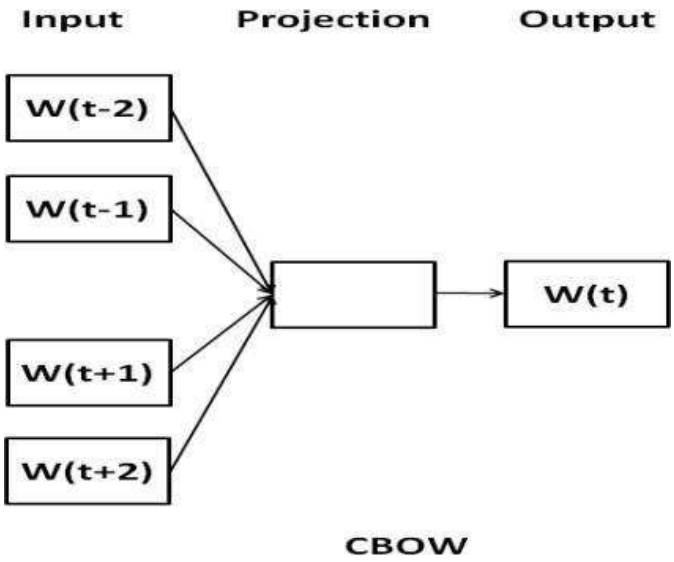

Fig. 1. CBOW Model

Step 4: Fast Text

FastText improves Word2Vec by embedding rare and unseen words using character-level n-grams. Example:

For the word "embedding" with trigrams: <em, emb, mbe, bed, ddi, din, ing, ng> FastText combines these sub-word embeddings to represent unseen words more effectively.

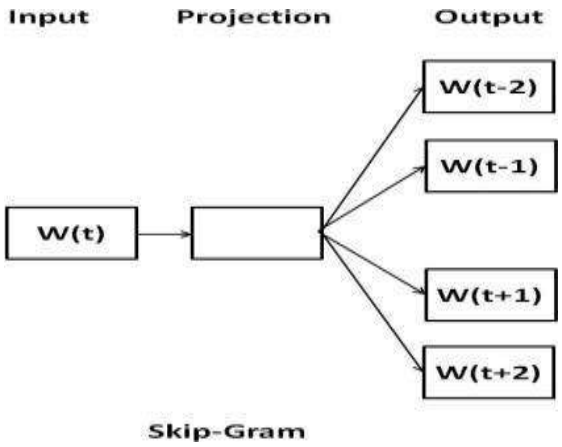

Fig.2.Skip-Gram Model

3.1. Proposed Approach for Celebrity Profiling

The proposed system predicts celebrity fame and occupation using word embeddings and machine learning.

1. **Preprocessing:** Clean text by removing punctuation, stopwords, and irrelevant data.
2. **Word Embeddings:** Generate vectors using Word2Vec, GloVe, or FastText to capture semantic relationships.
3. **Document Representation:** Aggregate word vectors to form document vectors.
4. **Model Training:** Train machine learning algorithms (e.g., SVM, Random Forest) on document vectors to build classification models.
5. **Prediction & Evaluation:** Use the models to predict fame and occupation and evaluate accuracy.

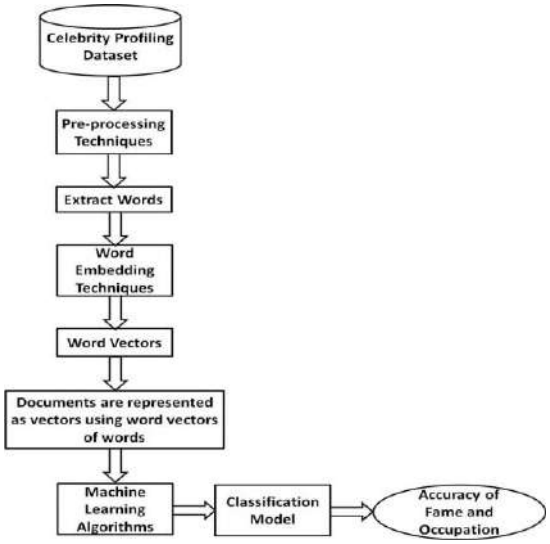

Fig.3.The Proposed model

4. EXPERIMENTAL RESULTS

In this experiment, the document vectors are represented with the word vectors generated by the word embedding techniques of Word2Vec and Fast Text. Three classifiers such as NBM, SVM and RF are used to train the model by using these vectors. The Table 5.3.1.1 presents the accuracies of the proposed approach for fame prediction.

Table 1. The accuracies of proposed approach for Fame prediction

Machine Learning Algorithms / Word Embedding Techniques	NBM	SVM	RF
Word2Vec	47.08	60.83	88.33
FastText	42.9	56.66	88.33

In Table 1, the random forest classifier attained best accuracies for fame prediction when compared with the accuracies of other two classifiers such as NBM and SVM. The Word2Vec obtained good accuracies for fame prediction when compared with the accuracies of Fast Text technique. The random forest classifier withWord2Vec attained best accuracy of 88.33 for fame prediction.

Table 2. presents the accuracies of the proposed approach for occupation prediction.

Machine Learning Algorithms/ Word Embedding Techniques	NBM	SVM	RF
Word2Vec	90.4	90.8	99.1
FastText	79.58	85.4	97.5

In Table 2, the random forest classifier attained best accuracies for occupation prediction when compared with the accuracies of other two classifiers such as NBM and SVM. The Word2Vec obtained good accuracies for occupation prediction when compared with the accuracies of Fast Text technique. The random forest classifier with Word2Vec attained best accuracy of 99.1 for occupation prediction.

5. CONCLUSION

Celebrity profiling, a subset of author profiling, involves predicting demographic traits such as gender, fame, birth year, and occupation by analyzing textual data. This study focuses on the 2019 PAN Competition task for celebrity profiling. We proposed a word embedding-based approach using Word2Vec and FastText to represent text as document vectors. These vectors were evaluated using three machine learning algorithms: Naïve Bayes

Multinomial, SVM, and Random Forest. The Random Forest classifier with Word2Vec achieved the highest accuracies, 88.33% for fame prediction and 99.1% for occupation prediction, demonstrating the effectiveness of the approach.

REFERENCES

[1] Rangel Pardo, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G.: Overview of the Author Profiling Task at PAN 2013. In: Forner, P., Navigli, R., Tufis, D. (eds.) CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain. CEUR-WS.org (Sep 2013).

[2] Maria De-Arteaga, Sergio Jimenez, George Duenas, Sergio Mancera, and Julia Baquero. Author Profiling Using Corpus Statistics, Lexicons and Stylistic Features—Notebook for PAN at CLEF 2013.

[3] Luis Gabriel Moreno-Sandoval, Edwin Puertas, Flor Miriam Plaza-del-Arco, Alexandra Pomares-Quimbaya, Jorge Andres Alvarado-Valencia, and L.Alfonso Ureña-L\`opez. Celebrity Profiling on Twitter using Sociolinguistic Features—Notebook for PAN at CLEF 2019. In Linda Cappellato, Nicola Ferro, David E. Losada, and Henning Müller, editors, CLEF 2019 Labs and Workshops, Notebook Papers, September 2019. CEUR- WS.org.

[4] Juraj Petrik and Daniela Chuda. Twitter feeds profiling with TF-IDF—Notebook for PAN at CLEF 2019. In Linda Cappellato, Nicola Ferro, David E. Losada, and Henning Müller, editors, CLEF 2019 Labs and Workshops, Notebook Papers, September 2019. CEUR- WS.org.

[5] Muhammad Usman Asif, Naeem Shahzad, Zeeshan Ramzan, and Fahad Najib. Word Distance Approach for Celebrity profiling—Notebook for PAN at CLEF 2019. In Linda Cappellato, Nicola Ferro, David E. Losada, and Henning Müller, editors, CLEF 2019 Labs and Workshops, NotebookPapers, September 2019. CEUR-WS.org.

[6] Matej Martinc, Bla\vz \vSkrlj, and Senja Pollak. Who is hot and who is not? Profiling celebs on Twitter—Notebook for PAN at CLEF 2019. In Linda Cappellato, Nicola Ferro, David E. Losada, and Henning Müller, editors, CLEF 2019 Labs and Workshops, Notebook Papers, September 2019. CEUR-WS.org.

[7] Victor Radivchev, Alex Nikolov, and Alexandrina Lambova. Celebrity Profiling using TF-IDF, Logistic Regression, and SVM—Notebook for PAN at CLEF 2019. In Linda Cappellato, Nicola Ferro, David E. Losada, and Henning Müller, editors, CLEF 2019 Labs and Workshops, NotebookPapers, September 2019. CEUR-WS.org.

[8] Björn Pelzer. Celebrity Profiling with Transfer Learning—Notebook for PAN at CLEF 2019. In Linda Cappellato, Nicola Ferro, David E. Losada, and Henning Müller, editors, CLEF 2019 Labs and Workshops, Notebook Papers, September 2019. CEUR-WS.org.

[9] T. MIKOLOV, I. SUTSKEVER, K. CHEN, G. S. CORRADO, AND J. DEAN, Distributed representations of words and phrases and their compositionality, in Advances in neural informationprocessing systems, 2013, pp. 3111–3119.

[10] Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T. (2017). Enriching word vectors with subword information, Transactions of the association for computational linguistics 5: 135–146.

[11] Raghunadha Reddy T, Vishnu Vardhan B, Vijayapal Reddy P, “Profile specific Document Weighted

approach using a New Term Weighting Measure for Author Profiling ", International Journal of Intelligent Engineering and Systems, 9 (4), pp. 136- 146, Nov 2016.

[12] J. Ali, R. Khan, N. Ahmad, and I. Maqsood. Random forests and decision trees. International Journalof Computer Science Issues (IJCSI), 9(5):272, 2012.

[13] L. Breiman. Random forests. Machine Learning, 45(1):5–32, 2001.

[14] V. Vapnik. The nature of statistical learning theory. Springer science & business media, 2013.

[15] https://pan.webis.de/clef19/pan19-web/celebrity-profiling.html

[16] Karunakar Kavuri, Kavitha, M. (2020). "A Stylistic Features Based Approach for Author Profiling". In: Sharma, H., Pundir, A., Yadav, N., Sharma, A., Das, S. (eds) Recent Trends in Communication and Intelligent Systems. Algorithms for Intelligent Systems. Springer, Singapore. https://doi.org/10.1007/978-981-15-0426-6_20.

[17] Karunakar. Kavuri and M. Kavitha, "A Term Weight Measure based Approach for Author Profiling," 2022 International Conference on Electronic Systems and Intelligent Computing (ICESIC), 2022, pp. 275-280, doi: 10.1109/ICESIC53714.2022.9783526.

[18] Chennam Chandrika Surya, Karunakar K, Murali Mohan T, R Prasanthi Kumari, "Language Variety Prediction using Word Embeddings and Machine Leaning Algorithms", Journal For Research in Applied Science and Engineering Technology, https://doi.org/10.22214/ijraset.2022.48280.