Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, No. 5, April : 2023

AUTHOR PROFILING BASED ON HOMONYM AND POLYSEMY: A STYLOMETRY FEATURE

Kathuroju Srinivas, Assistant Professor, Area of Information Technology & Analytics
Institute of Public Enterprise, Shamirpet, Dist: Medchal-Malkajgiri, Hyderabad, Telangana, India
ksrinivas2000@gmail.com

Abstract
Author profiling (AP) is a significant task in various applications Forensics, Cybersecurity, Marketing and education etc. AP illustrates the author's age group, gender, native language and place. Most of the text is written by distinct users on social media applications like Whatsapp, Facebook and Twitter in the form of articles, tweets and text on a particular topic. The writing style of articles, tweets and text can vary between one author and the other. We focus on author profiling in forecasting the age group and gender based on the use of polysemy and homonym, a stylometry feature in the tweets. We have used Bag-of-Words, N-gram model for feature extraction. Besides this, SVM and Logistic Regression are used to check the accuracy of tweets containing polysemy and homonym.

Keywords: Author Profiling, Homonym, Polysemy

1. Introduction

The text in social media is increasing enormously and has become a part of our lives nowadays. The text is named as a tweet on Twitter. However, the tweets we have seen on Twitter or social media are unstructured and semi-structured. Many researchers have contributed to Author profiling to discover the demographic traits of the author like age group, gender, academic ground, born language, and place by examining their writing techniques. The writing techniques or styles of the author depends on the stylometry features.

In general, every human being maintains his writing technique, and it will exist while composing on Twitter tweets, blogs, reviews, news articles and manuscripts. Generally, the writers' writing styles differ based on the choice of issues and the writing techniques like parts-of-speech, grammar, and vocabulary richness etc. In this paper, we use Polysemy and homonym, a stylometry feature for author's age group classification and gender recognition. Polysemy means a word or phrase which has the same spelling, different meaning, and sense. A homonym means a word with different spelling and meaning. The use of polysemy and homonym stylometric features helps to capture the writing style of different authors.

2. Related Work

In [1], the use of stylometric characteristics based on morphological and syntactic attributes for recognizing writing patterns in Spanish, French and Portuguese is viable, as it points to promising results. Furthermore, it was noticed that the more textual information obtained from the authors, the better the decisions made by the classifier. Overall, the approach proved to be stable and robust to the experiments. In [2], 397 Stylometric characteristics and 350 words extracted using tools like AASF, Madamira and Weka. Also used SMO-SVM as a base classifier. Collected Arabic text from Dar Al-ifta AL Misriyyah site by OctoParse 7.0.2 tool. In [3], the prototype accepts Greek text and forecasts the gender of the author. The projections are founded solely on the text and the hypothesis that carries details about gender. In [4], established strategy on the Universal Sentence Encoder prototype to acquire low dimensional vectors of compositions and utilise them as traits to accomplish author profiling. To assess the vectors' quality accepted by USE, they operated traits for training two machine learning algorithms that typically brings promising outcomes in the author profiling. In [5], indicating the polysemy class of terms can even be advantageous for deciding the context required for obtaining models that adequately contemplate the importance of term representatives in handling text. In [6],

BERT Large seems to capture nuanced word-sense distinctions similar to human annotators and, to some degree is, capable of grouping sense interpretations by their contextualized embeddings. [7] used deep learning model employing CNN and LSTM approaches. As an effect, it is better accuracy in signifying the age and the gender of author profiling from his\her composing tweets. In [8] AA analyses, genre and topic, are critical aspects that enhance trait sample implementation. On the other side, an identical connection between the issue and author technique exists in multi-domain and multi-genre issue category schemes where the genre and author technique damage the forecast interpretation. In such circumstances, the genre and author technique details can be utilised likewise for trait extract. In [9] Suggested a method employing the LDA paradigm in concurrence with n-grams to deliver condensed dimension topical illustration of Urdu Corpus. Illustrated how the topical terms of LDA could be employed with improved sqrt-cosine distance metric to categorise test manuscripts. In [10], Enhanced algorithms such as automated authorship attribution and plagiarism detection. Help forensic specialists or grammatists in constructing profiles of writers. Aid intelligence applications to investigate antagonistic and terrifying statements. [11] The prototype was developed and utilized distinct classification algorithms for the diverse proposals. The classification algorithms are linear classifiers, SVM and MLP classifiers. Trained prototypes for the individual task, then the outcome has been integrated. In [12], A novel engineered hybrid feature vector and an N-gram characteristic vector is operated with ensemble weighting for two classifiers, the RF and LR, respectively. They offered feature vector applied emojis, female suffixes, and a set of well-selected function terms, i.e. curse terms, feelings, politics, etc. In [13], the necessary facts regarding an author are augmented with the most delinquent up-to-date details about the author from the ACM DL before queries are dispatched to Scopus. They obtained better valid author profiles using the Scopus API by acquiring more details about individual authors. However, the author profiles in the two publisher networks are not instantly connected. In [14] suggested solution (S3) has been assessed employing several real-world datasets regained from Project Gutenberg. Compared results with existing state-of-the-art authorship attribution strategies. Comprehensive empirical investigations include their practice, has topped existing state-of-the-art techniques. In [15], the experiment was conducted with the multiple regular 8000 words in this assignment. The composition vectors were described with these 8000 words. Individual vector weight is defined with the term frequency details in a manuscript. Best word frequency is not adequate to enhance the exactness of author profiles. The suggested PSTW measurement conveyed the most acceptable exactness for age and gender forecast. In [16], a distinct characteristic is the interactive linking of text and visualizations where text explicitly defines the considerable essential designs and visualization permit for contextual inquiry of journal logs. In [17], they created a corpus of 3,000 scholarly studies from three outstanding Bengali authors. Then, various investigations were conducted on the proportional dataset and achieved a state-of-the-art test accurateness of 98% utilising character bigrams (tf) as elements and Naive Bayes classifier. In [18] presented representatives, conducted by the designated linguistic modality, endeavour to mitigate the problems associated to the characteristic engineering approach in the existing authorship analysis. Experimentations on the publicly known standard datasets for the authorship verification issue, authorship identification and authorship description issue indicate

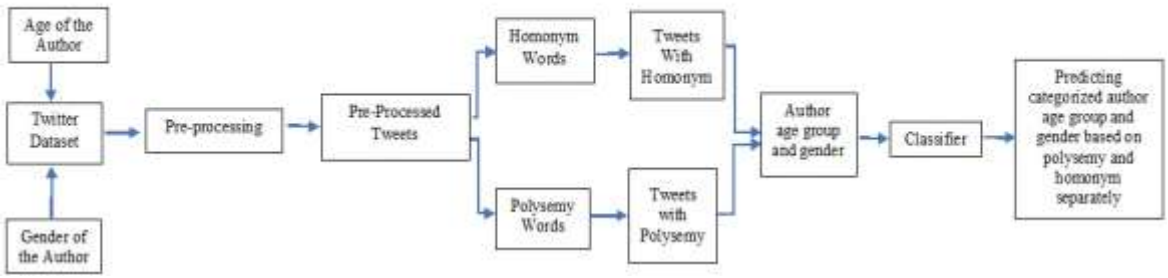


Fig. 1. Predicting category of the author's age group and gender

that presented instances are useful and powerful on various datasets and AA issues. In [19], offered a standard multilingual Facebook author profiles corpus to create and assess author profiling approaches. The offered corpus includes 479 profiles along with demographic details. In addition, this analysis also donates a manually developed bilingual glossary of 7749 admissions to decode Roman Urdu phrases into English. In [20] Preferably, authorial research methods must remain appropriate. Anticipated growth in text pieces that arise from online references. It is reckless to think that text pieces will be a specific length, an author will use identical identities across numerous manuscripts, etc. These inconsistencies confuse stylometry; deriving methods strong against these elements are required. Next, researchers must think of different points to efficiently use stylometry in useful applications. In [21] PMSVM algorithm employed in concurrence with various and complementary characteristics is a promising advancement over the state-of-art methods. Experimenters' research characteristic groups bypassed computational problems. In acquisition, the incremental matching study revealed that recent attribution practices could significantly decrease the integer of users to be studied in realistic circumstances. In [22], Prototypes are set, the evaluation is brought out. Demonstrate the suggested ways by random signatures from four datasets of handwritten signatures. Also, the problem of performing this task for human examiners is presented through an illustrated Turing Test as a baseline. Ultimately, display valuable assistance to cracking the issue since our automated techniques top the forensic and non-forensic human evaluation outcomes. [23] Pulled 93 stylometric features: 34 word-based, three syntactic, and 56 content-specific. TF to calculate the importance of individual stylometric features. In [24], new corpora have been created, wrapping numerous vocabularies for plagiarism detection, author identification and author profiling. Concurrently with the yielded evaluation outcomes, this unique help greatly determines the state of the art in the separate rooms. In [25], They executed an authorship attribution method to experiment and approximate four different Naive Bayesian classification models for probability analysis relying on the presence or absence and characteristic frequency. Probability analysis is founded on the mean and standard deviation of the characteristics. They assessed implementation on an enormous corpus of four separate datasets and analyzed the impact of controlling and normalization on the attribution method.

3. Proposed Approach

Our proposed approach is on Author profiling in finding out the age group and gender of authors found on polysemy and homonym words. In this paper, we use a Covid 19 tweets on Twitter dataset in forecasting the age group and gender of the author by using Bag-of-Words and N-gram models for feature extraction. In addition to this, we use SVM and LR to predict the age group and gender of the author from the tweets based on polysemy and homonym words.

The following are the steps for implementing our proposed approach:

Step 1: Collecting the corpus of tweets on Twitter which consists of age and gender of author.
Step 2: Perform data pre-processing to achieve pre-processed dataset of tweets.
Step 3: Collect 500 polysemy and homonym words from standard dictionaries.
Step 4: Extract the tweets which consists of polysemy and homonym words separately.
Step 5: Determine the author's age group and gender based on the extracted tweets.
Step 6: Use classifier to predict the categorized age group and gender based on tweets with polysemy and homonym separately.
Step 7: Train the classification models with the training dataset of polysemy and homonym tweets.
Step 8: Test the classification models for accuracy.

3.1 Methodology

Data collection is the first phase for author profiling. Many Researchers built the corpus with the text data from these sources like Twitter, Blogs, and social media etc. We have used the Twitter dataset of Covid 19 tweets from Kaggle.com. The dataset is a collection of over one million tweets, consists

of age and gender of the author. The data has to be pre-processed since it contains hashtags, URL's, lowercase, stop-words (use of articles), punctuation marks etc. The motive of pre-processing is to make the noise-free data and for effective feature extraction.

3.2 N-grams

N-grams are successive series of terms or characters in a composition. The 1-gram is named unigram. The 2-grams are named bigrams, and so on. We have used unigram model to split the complete tweet into independent terms separated by comma.

If N= n1,n2,…

Where N= complete set of words in a tweet

n1,n2,….= independent terms separated by comma

3.3 Bag-of-Words (BoW)

We have used BoW model inorder to collect 500 polysemy and homonym words from standard dictionaries. The pre-processed tweets are checked for the availability of our collected polysemy and homonym bag-of-words, which leads to the extraction of tweets and thus categorized as tweets with polysemy and homonym separately.

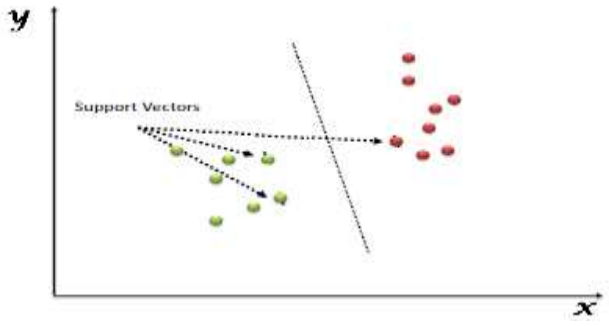
Next, in our approach the age and gender of the author associated with extracted tweets will be known. Further, Author age has been categorized into five groups like 16-25, 26-35,36-45,46-55,56-XX and gender is a binary classification of two groups Male and Female. Then finally we have used classification models SVM and LR to determine the categorized age group and gender of the author based on the use of polysemy and homonym separately in tweets.

	Age Group	**No. of Tweets with Polysemy**	**No. of Tweets with Homonym**
1	16-25	9524	12658
2	26-35	9645	12554
3	36-45	9559	12491
4	46-55	9398	12513
5	56-XX	9527	12411
	Total	**47653**	**62627**

Table 1. Author age group categorization with usage of polysemy and homonym

3.4 Support Vector Machine

(SVM) is a supervised machine learning algorithm that can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is a number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well.



Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, No. 5, April : 2023

3.5 Logistic Regression

Logistic Regression is a Machine Learning algorithm which is used for the classification problems, it is a predictive analysis algorithm and based on the concept of probability. Some of the examples of classification problems are Email spam or not spam, Online transactions Fraud or not Fraud. Logistic regression transforms its output using the logistic sigmoid function to return a probability value. We can call a Logistic Regression a Linear Regression model but the Logistic Regression uses a more complex cost function, this cost function can be defined as the 'Sigmoid function' or also known as the 'logistic function' instead of a linear function. The hypothesis of logistic regression tends it to limit the cost function between 0 and 1. Therefore linear functions fail to represent it as it can have a value greater than 1 or less than 0 which is not possible as per the hypothesis of logistic regression.

$$0 \leq h_\theta(x) \leq 1$$

Logistic regression hypothesis expectation

$$f(x) = \frac{1}{1 + e^{-(x)}}$$

Formula of a sigmoid function

4. Discussion & Results

In the collected Twitter dataset, the overall tweets with homonym words are widely used than the polysemy words. The categorization of author age group based on usage of polysemy and homonym in the tweets is shown in the Table2. The author's age group of 26-35 highly uses polysemy, and 16-25 use more homonyms in writing the tweets on Twitter. Table 2 shows the gender prediction of the author found on the use of polysemy and homonym separately. On comparing males use slightly more polysemy and homonyms words than females in writing tweets on Twitter.

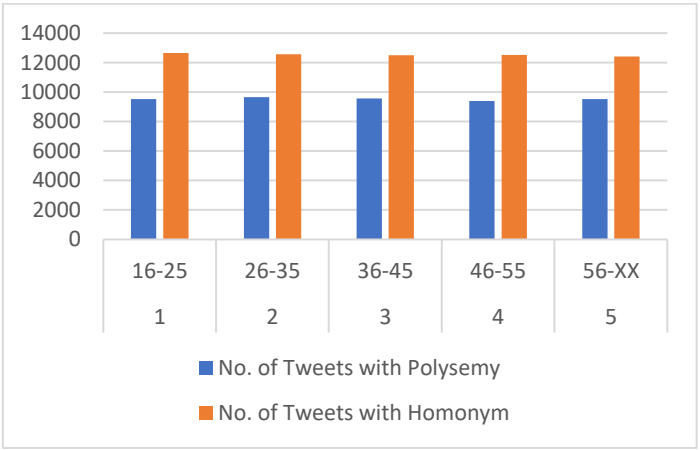


Fig. 2. Author age group categorization and usage of polysemy and homonym separately

Gender	Polysemy	Homonym
Male	23894	31333
Female	23759	31294

Table2. Gender of the author using polysemy and homonym

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, No. 5, April : 2023

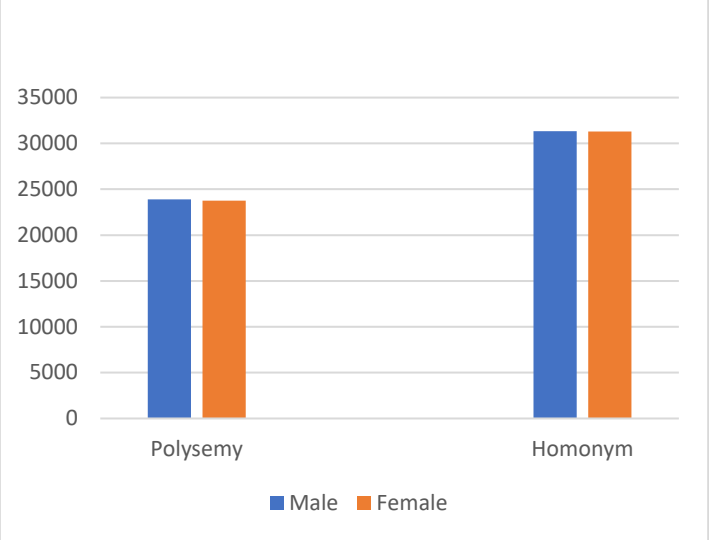


Fig. 3. Gender of the author using polysemy and homonym

Classifier	SVM	LOG
Polysemy	0.71	0.68
Homonym	0.75	0.73

Table3. Accuracy of polysemy and homonym on tweets using SVM and LR

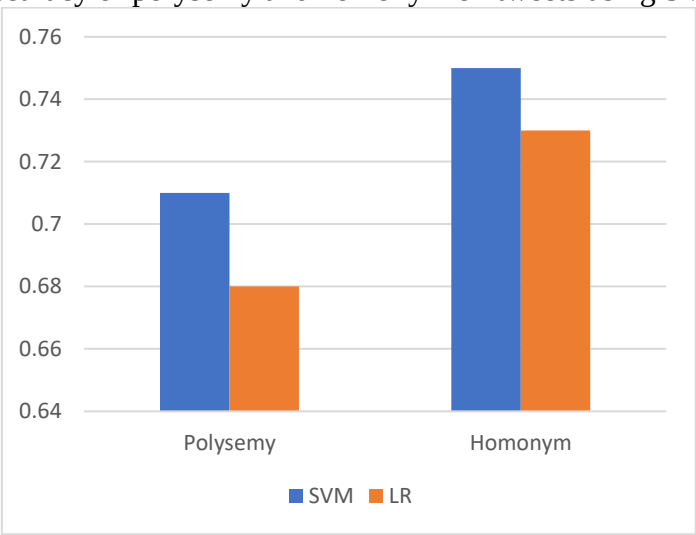


Fig. 4. Accuracy of polysemy and homonym on tweets using SVM and LR

5. Conclusion

In this paper, we have used the over one million tweets of Covid 19 dataset on Twitter. We have composed 500 polysemy and homonym words together from standard dictionaries. In comparison of accuracy the SVM is better than the Logistic Regression. The SVM gives 0.71 and 0.75 in predicting the age group and gender of the author based on polysemy and homonym, a stylometry feature, whereas Logistic regression results 0.68 and 0.73. As a future scope, a new stylometry feature has to be identified based on which demographics of the authors' can be predicted and also can be applied on cross-genre social media applications.

6. References

[1]. P. Varela, M. Albonico, E. Justino And J. Assis "Authorship Attribution In Latin Languages Using Stylometry" IEEE Transactions, (2020).
[2]. Mohammed Al-Sarem, Faisal Saeed, Abdullah Alsaeedi,Wadii Boulila and Tawfik Al- Hadhrami, "Ensemble Methods For Instance-Based Arabic Language Authorship Attribution" IEEE (2020).
[3]. Spiros Baxevanakis, Stelios Gavras, Despoina Mouratidis, Katia Lida Kermanidis "A Machine Learning Approach for Gender Identification of Greek Tweet Authors" ACM (2020).
[4]. Aquilino Francisco Sotelo , Helena Gomez-Adorno, Oscar Esquivel-Flores , and Gemma Bel-Enguix "Gender Identification in Social Media Using Transfer Learning"Springer (2020).
[5]. Aina Gari Soler, Marianna Apidianaki "Let's Play Mono-Poly: BERT Can Reveal Words'Polysemy Level and Partitionability into Senses" ACL (2020).
[6]. Janosch Haber and Massimo Poesio "Patterns of Lexical Ambiguity in Contextualised Language Models"ACL (2020).
[7]. Roobaea Alroobaea, Sali Alafif, Shomookh Alhomidi, Ahad Aldahass, Reem Hamed, Rehab Mulla, Bedour Alotaibi, "A Decision Support System for Detecting Age and Gender from Twitter Feeds based on a Comparative Experiments", IJACSA (2020).
[8]. Hayri Volkan Agun and Ozgur Yilmazel "Incorporating Topic Information in a Global Feature Selection Schema for Authorship Attribution"IEEE (2019).
[9]. Waheed Anwar, Imran Sarwar Bajwa, M. Abbas Choudhary, And Shabana Ramzan "An Empirical Study on Forensic Analysis of Urdu Text Using LDA-Based Authorship Attribution" IEEE (2019).
[10]. Farhan Ullah, Junfeng Wang, Sohail Jabbar, Fadi Al-Turjman, And Mamoun Alazab "Source Code Authorship Attribution Using Hybrid Approach of Program Dependence Graph and Deep Learning Model", IEEE (2019).
[11]. Hamada A. Naye "NAYEL@APDA: Machine Learning Approach for Author Profiling and Deception Detection in Arabic Texts" FIRE (2019).
[12]. Shereen Hussein, Mona Farouk, ElSayed Hemayed "Gender identification of egyptian dialect in twitter" Egyptian informatics Journal (2019).
[13]. Karim Alinani, Annadil Alinani, Dua Hussain Narejo, and Guojun Wang,"Aggregating Author Profiles from Multiple Publisher Networks to Build a List of Potential Collaborators" IEEE (2018).
[14]. Raheem Sarwar, Chenyun Yu , Ninad Tungare, Kanatip Chitavisutthivong, Sukrit Sriratanawilai, Yaohai Xu, Dickson Chow, Thanawin Rakthanmanon, And Sarana Nutanong "An Effective and Scalable Framework for Authorship Attribution Query Processing" IEEE (2018).
[15]. Sai Satyanarayana Reddy Seelam, Shrawan Kumar, Gopi Chand M, Raghunadha Reddy T "A New Term Weight Measure for Gender and Age Prediction of the Authors by analyzing their Written Texts" IEEE (2018).
[16]. Shahid Latif and Fabian Beck,"VIS Author Profiles: Interactive Descriptions of Publication Records Combining Text and Visualization" IEEE (2018).
[17]. Shanta Phani, Shibamouli Lahiri, Arindam Biswas, "A Supervised Learning Approach for Authorship Attribution of Bengali Literary Texts", ACM Trans (2017).
[18]. Steven H. H. Ding, Benjamin C. M. Fung , Farkhund Iqbal, and William K. Cheung,"Learning Stylometric Representations for Authorship Analysis" IEEE (2017).
[19]. Mehwish Fatima, Komal Hasan , Saba Anwar , Rao Muhammad Adeel Nawab," Multilingual author profiling on Facebook" Elsevier (2017).
[20]. Tempestt Neal, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, And Damon Woodard,"Surveying Stylometry Techniques and Applications" ACM (2017).
[21]. Anderson Rocha, Walter J. Scheirer,Christopher W. Forstall, Thiago Cavalcante, Antonio Theophilo, Bingyu Shen, Ariadne R. B. Carvalho, and Efstathios Stamatatos," AuthorshipAttribution forSocial Media Forensics", IEEE (2016).
[22]. Moises Diaz, Miguel A. Ferrer, Soodamani Ramalingam, Richard Guest,"Investigating the Common Authorship of Signatures by Off-line Automatic Signature Verification without the Use of Reference Signatures" , IEEE (2015).
[23]. Rafael T. Anchi, Francisco Assis Ricarte Neto, Rogerio Figueiredo de Sousa, and Raimundo Santos Moura,"Using Stylometric Features for Sentiment Classification", Springer (2015).
[24]. Efstathios Stamatatos, Martin Potthast, Francisco Rangel, Paolo Rosso, and Benno Stein, "Overview of the PAN/CLEF 2015 Evaluation Lab", Springer (2015).
[25]. Alaa Saleh Altheneyan , Mohamed El Bachir Menai,"Naive Bayes classifiers for authorship attribution of Arabic texts", Elsevier (2014).