JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 8, ISSUE 01, 2021

Gender Classification using Twitter Text Data

Mrs.M.Jhansi Lakshmi,Assistant Professor, Department of Information Technology,
CMR Engineering College, Hyderabad, Telangana,
E-Mail-id mettu.jhansilakshmi@cmrec.ac.in

V.Rohitha ,178R1A1254@cmrec.ac.in,
G.Akhil Reddy ,178R1A1218@cmrec.ac.in,
P.Navya,178R1A1243@cmrec.ac.in

Abstract - Social networking and other content sharing services are becoming increasingly popular throughout the world. The ability to identify a person's gender based on these brief communications is an intriguing research topic that has applications in forensics, marketing analysis, advertising, and recommendation. The usage of tweets and Natural Language Processing (NLP) methods in a gender classification system will be investigated in this study. This study will look into how a system for classifying people by their gender uses tweets and Natural Language Processing (NLP) techniques.

For the sake of this study, a brand-new dataset that includes the user gender and related tweets has been created. This dataset was created since there was no publicly accessible standard dataset with the volume needed to carry out this inquiry. According to the findings, the conventional Bag of Words model did not produce any noteworthy categorization outcomes. However, using a variety of machine learning techniques, word embedding models have greatly outperformed them. As a result, it has been demonstrated that word embedding models are the best method for identifying gender from text data from Twitter.

Keywords — Natural Language Processing, Gender Classification, Machine Learning, Twitter, Word Embedding.

Introduction – One of the most well-known social media platforms is Twitter, which enables users to read and post 280-character messages on its website [1]. Tweets are the common name for these messages. One of the reasons Twitter has become so well-liked among social media users is the message size restriction. Global internet users have an equal opportunity to engage with other users on Twitter, including top celebrities, politicians, famous people, etc., and follow them on a regular basis by reading their tweets. Social media's rising popularity has produced a special chance to study human civilization broadly [2]. One of the well-known marketing websites, Omnicare agency, claims that more than 330 million individuals use Twitter on a regular basis, that there are more than 500 million tweets sent every day, and that 139 million people use it everyday. Additionally, it has been found that more than 66% of Twitter users are male and only 34% are female [3],[4].

Twitter has a default setting that makes tweets visible to everyone, but users can restrict messages so that they are only visible to selected followers. Public sharing of tweets is at the heart of Twitter's philosophy. Over 90% of Twitter users, according to Kvamme, prefer to make their exchanges public [5].

Marketing firms and the government are becoming increasingly interested in gender classification as social media usage rises.

For a number of different organizations, such as those involved in projecting movie box office results, election prediction, forensics, cybercrime, and making recommendations [5]–[7], extracting knowledge from this hidden content could be very helpful. This practice of many social media users, known as anonymous profiles, involves concealing their real identities, including user names, ages, and gender. This necessitates the development of sophisticated methods for determining the gender of these profiles. Experts in forensics are typically necessary to track cybercriminals [2].

The purpose of gender classification, for example, is to divide the object into two categories based on specific qualities. This is known as a binary classification problem [8]. The researcher has plenty of opportunity to uncover trends in this data because such user text data is publicly accessible. Various text pre-processing and Natural language processing (NLP) techniques can be used to normalize such unstructured/structured text input. NLP, a branch of artificial intelligence (AI), gives robots the capacity to comprehend spoken and written human language. Different approaches to analyzing and comprehending this data as well as automatic interpretation of human language are included in NLP [9].

In order to train the classifier and automatically determine a twitter user's gender, this study will use a variety of Machine Learning (ML) techniques.

utilizing several NLP methods on a tweet text. In general, classifying a token sequence for a document or sentence is what most NLP tasks, including sentiment analysis, entail. To extract features from the tweet text using TF-IDF (Term Frequency - Inverse Document Frequency), vectorize the text, and then use ML classification to train and predict the output label. The effectiveness of the model will then be evaluated using this. As no such annotated twitter data at the needed scale is currently available, this study will initially involve generating the dataset, which will be made available to the public for additional research.

Research Objectives - The primary goal of this study is to create the dataset. A tagged dataset of around 20,000 users is available via Kaggle [10]. To produce a labeled dataset for this study, a bootstrapped version of this dataset will be utilized to retrieve tweets for each of those people.

The primary goal of this study is to identify gender by using various ML techniques and NLP strategies based on text attributes to distinguish between gender differences. This kind of research has led to the creation of a number of applications in the field of applied research, including marketing and psychological analysis [11]. By examining their previous tweets, it is important to identify the linguistic profile of the user if an anonymous Twitter user interacts as part of some questionable activity. The results of this study may reveal a trend and aid in analyzing the user's history and gender (whether it be male or female) [12]. Several businesses utilize Twitter to distribute digital advertisements for their target gender in terms of marketing and recommendations in order to grab their attention. This has shown to be quite successful in terms of increasing revenue and reach, particularly for the e-commerce sector. Additionally, social networking platforms like Twitter are renowned for having a significant number of bots that disseminate false information to the general public, which has the potential to influence elections and major campaigns. Therefore, it is crucial to be able to determine whether these Tweets are being written by a human or a robot, and to comprehend who is writing them [13].

Gender Classification in Text - For the two classes in the example given, male or female, gender classification can be approached as a two-class or binary classification problem in practice. Without knowing anything about the user, the anonymous text or message is to be assigned to one of the classes in this classification [14]. Text analysis is a highly difficult work for computers, but it is extremely simple for people. It is frequently simple and quick for a human to identify the gender class through visual inspection. These systems accept text messages or language sentences as input. The most common NLP machine learning task, or function approximating, is classifying a series of tokens, such as those in a text or sentence:

$$f1 \rightarrow (1,0)$$

Sentiment, among other factors, may decide f1 in this example domain. The objective is to assign the data points a male or female category using the numbers 1 (male) and 0 (female) (female). The sole distinction in this case is that it works with various text pairs in various assignments.

During this categorization procedure, additional characteristics of gender classification including sex and gender identification must be taken into consideration. The American Psychological Association defines sex as a biologically determined human state that is typically categorized as male, female, or intersex. Gender expression refers to how a person interacts and communicates in a certain culture, which can be inferred from the way that individual's interest in dress or communication style [15]. Currently, sex research has been used to study gender classification rather than gender expression. For the sake of uniformity, the two conceivable genders of male and female are chosen. Thus, determining whether a Twitter user is male or female is the goal of gender classification research. Analyzing their profile content elements allows for this. as well as the gender expression on exhibit. Additionally, some study [2] has concentrated on examining and analyzing meta-factors or very relevant attributes for gender identity.

Related Work - There hasn't been any extensive research done on the problem of gender detection using Twitter data. Although comparable NLP studies, such as those by M. Koppel, 2002 & 2009, and S. Goswami, 2009 & 2012 [16] – [19], have been done to automatically identify the gender of a text's author. The development of research in this area of social media has benefited from improvements in ML and text classification algorithms. In order to categorize gender, prior research projects have employed a variety of ML algorithms, including as Modified Balance Winnow, SVM, and Nave Bayes, which will be examined in this study.

Rao [20] introduced one of the earliest experiments on detecting gender classification using latent user variables. Gender, age, political orientation, and nationality were the main variables under study. The author set out to investigate the value of linguistic content processing using status updates, communication behavior, and social network structure. The goal was to identify latent user traits so that some latent features in text might be automatically detected. Rao conducted thorough study to determine the most crucial attributes and which ones perform best by using them to solve a supervised ML problem [20]. SVM classification methodology was used to bring the analysis to a close. According to the study, Twitter offered many socio-linguistic indices in comparison to other social media platforms. For instance, emoticons like <3 and question signs (?), which are strong signals, are more frequently used by female users. Whereas a girl used the LOL laughing indicator and most men used the LMAO laughing indicator. OMG was discovered to be utilized by females in addition to excitement markings. The possessive word characteristics, such as "my gf," "my wife," "my husband," and "my bf," were another potent indicator of men tweeting [20]. When Clay used tweets from Nigeria to determine the gender of Twitter users, he discovered a similar result. Using solely unigram features from tweets, which were more reliable indications of user gender, this study deployed a supervised ML technique. According to Clay's research, text data from Twitter "tweets" is strongly predictive of gender [21].

In a recent study, Marco Vicente proposed leveraging unstructured text information taken from Twitter user profiles to automatically identify the gender class. The author proposed using manually specified features to draw out relevant data from user profile characteristics like screen name. This research provided name-related qualities that incorporate user experience, which were then assessed using a dataset of 242,000 users. Several supervised machine learning methods, such as SVM, LR, Naive Bayes, K-means, and fuzzy c-Means clustering, were assessed. In order to identify the gender of Twitter users, a supervised ML classifier had a high accuracy of 97%. With the unsupervised method based on fuzzy c-Means, approximately 96% accuracy was likewise achieved for this problem.

The author noted that utilizing a labelled training set had a benefit and that there was a good chance that employing a larger dataset would result in greater accuracy [32].

Methodology - The research's approach included several processes, such as data collection (creating the data set), data preprocessing, data transformation, data mining, and the use of machine learning models.

A. Data Pre-Processing

The pre-processing of the data is a highly important stage because better results can only be obtained with high-quality data. Steps like feature engineering, data visualizations, and ML classifiers are used to do this. Social media data needs to be thoroughly cleaned because it is unstructured, or in other words, it is raw and very noisy. The primary goal of this stage is to eliminate erratic and noisy tweet data. It is necessary to eliminate tweets that have very little meaning in a text context, such as extra blank space, special characters, hashtags, emoticons, and smileys.

The twitter pre-processor library has been used to sanitize tweets in order to guarantee their high quality. The following text can be removed from tweets using this library: URLs, Hashtags, Mentions, Reserved Words, Emojis, and Smileys. The next crucial stage is to make sure that each user's tweet length is reasonable, therefore additional analysis is done and the length is divided into 5–50 words. This filtering will make sure that suitable length is taken into account for subsequent actions. Last but not least, the dataset for the study exclusively includes English tweets. was further filtered, leaving just tweets that were written in "English." The "Spacy" library proved extremely useful in this endeavor. and the process of extracting the English tweets takes a while.

B. Data Transformation

The next step is to convert the word sequences into numerical characteristics after the cleaning process. Additionally, because ML classifiers are not intelligent enough to analyze text in its raw form, this is a crucial duty. The principle of instance optimization, math, statistics, etc. serve as the foundation for categorization. Additionally, such data is unstructured, loud, and dispersed, particularly when it is in natural language text format or has high dimensionality added to it. The majority of the time, feature extraction is crucial to achieving the research goal of analyzing tweets from a specified dataset and increasing the precision of gender classification in unstructured text [34].

C. Data Mining

The next stage is to classify the revised data using machine learning models after multidimension data transformation has been finished. In this study, supervised machine learning (ML) is considered to accurately predict the output label class in accordance with prior studies [29], [35], and [36]. In this study, the classifiers LR, MLP, SVM, Naive Bayes, RF, and XGBoost were considered. These binary classification techniques are straightforward but efficient. A straightforward pipeline that consists of the TF-IDF and LR model is used to build a baseline. To compare the results with the benchmark, additional algorithms were incorporated with various feature engineering techniques (Word Embeddings - W2Vec and GloVe).

D. Utilizing Machine Learning Models

In order to forecast the gender class on a featured Twitter dataset, this research used six different ML classification algorithms (LR, MLP, SVM, Naive Bayes, RF, and XGBoost). SVM, Naive Bayes, and LR have all shown to be quite effective and In the past, binary categorization issues were extensively exploited. Regardless of the volume of data, these strategies are also well known for having strong predictive performance [21], [26], [29], [31], [32]. Naive Bayes is renowned for being straightforward and straightforward to use. In terms of accuracy, simplicity, and efficiency of categorization, this technique has also surpassed alternatives in practical applications [37]. The best model performance and execution speed are both found in XG boost. In this research, additional techniques like RF and MLP are also examined for comparison purposes to analyze the differences in results obtained when compared with other well-known classifiers.

Twitter Dataset - The goal of this study is to create a labelled Twitter dataset that can be used to infer gender from tweet content. There isn't a public dataset available yet that includes gender and tweets; So, it has been agreed that a dataset will be created as part of this research that includes the gender and its associated tweet data. A twitter dataset with 20,000 rows was utilized to start this approach [10]. This dataset sought to identify Twitter "user ids" and the genders that went with them.

The purpose of the study is to perform gender classification on text data, therefore the only fields needed for further processing are the user id and gender columns, leading to a two-column data filter. Also, since unisex and brand names cannot be classed as male or female, the focus of this research is on the gender categories of male or female. Two filters are applied to the "gender: confidence" column to make sure of that. Then, the "gender: confidence" column for both groups was filtered with the value "1," which denotes a complete gender identification. Second, the "man" and "female" options were added to the "gender: confidence" column filter. As a result, the dataset was decreased from 20,000 users to 10,020, including 4,653 men and 5,367 women.

It is necessary to eliminate a lot of the noise data from the Tweets dataset, such as URLs, hashtags, mentions, emojis, reserved terms (RT, FAV), smileys, and numbers. For cleaning tweets, the "Twitter Pre-Processor" library was utilized. Cleaning tweets was one of the most crucial phases because the existence of these characters lowers the quality of the corpus created. Keeping English tweets for analysis was a crucial part of the data pre-processing process, therefore the language detection technique was used with the spacy library. Lastly, the stopwords are eliminated at runtime using the NLTK package. The data is now suitable for additional feature engineering, such as ML classifications, after the cleaning process [33].

When noisy data was cleaned, the number of "more than 20" tweets was reduced from 346,000 to 41,932. The fact that there were 19,387 fewer male tweets than female tweets—nearly half of the 41,000—shows that the sample is balanced. The number of tweets was further lowered after language identification. There were roughly 21,021 tweets from both men and women in total, which suggests that nearly 50% of the tweet data was in another language.

Results - The main dataset is initially divided into train and test sets in an 80:20 ratio. While the testing set only has 4,205 tweets, the training set contains 16,816 tweets. To achieve the assessment metrics with the least amount of bias, training data is utilized to fit the model to it and then evaluated using completely unrelated data. The text input is then converted to numeric data using the standard TF-IDF feature extraction method before being supplied to the ML classifier to determine the gender class. In order to train ML models with more semantically rich representations, word embedding features are used. The non-negative values produced by word embedding pre-trained models cannot be used with Naive Bayes and MLP models, hence these models were eliminated from the results.

The performance metrics for the bulk of word embedding models have improved, and the LR model with W2Vec vectors produced substantially better results. Also, it demonstrates the ability of word embedding techniques to identify the semantic meaning of various words, an analogy that would have aided in raising performance metrics.

Last but not least, this study's findings differ from those of earlier, related studies since our dataset is distinct, random, objective, diversified, and a decent depiction of social media. The dataset chosen for this study was truly random and impartial in that it was not chosen from any particular group, such as politicians or celebrities. Millions of data points in the used datasets should influence how accurate the trained models are. It was noted that Miller's [25] research on gender prediction involves manually classifying data as male or female tweets in order to achieve high accuracy with less tweets.

References –

[1] K. Gligorić, A. Anderson, and R. West, 'How Constraints Affect Content: The Case of Twitter's Switch from 140 to 280 Characters', ArXiv180402318 Cs, Apr. 2018.

[2] J. A. Lopes Filho, R. Pasti, and L. De Castro, 'Gender Classification of Twitter Data Based on Textual Meta-Attributes Extraction', 2016, pp. 1025–1034.

[3] S. Aslam, 'Twitter by the Numbers (2020): Stats and Demographics', 05-Jan-2020. [Online]. Available: https://www.omnicoreagency.com/twitter-statistics/. [Accessed: 08- Mar-2020].

[4] H. Bagheri and M. J. Islam, 'Sentiment analysis of twitter data', ArXiv171110377 Cs, Dec. 2017.

[5] H. Kvamme, 'Gender prediction on Norwegian Twitter accounts', 119, 2015.

[6] V. A. K and S., 'Sentiment Analysis of Twitter Data:A Survey of Techniques', Int. J. Comput. Appl., vol. 139, no. 11, pp. 5–15, 2016.

[7] A. &. I. Farzindar, Natural Language Processing for Social Media, 2nd Revised. San Rafael, United States: Morgan & Claypool Publishers, 2017.

[8] S. Aleksandr and T. L. D. G. R. R. I. M, 'Machine Learning Models of Text Categorization by Author Gender Using Topic-Independent Features', in 5th International Young Scientist Conference on Computational Science, 2016, vol. 101, pp. 135–142.

[9] S. Harispe and R. S. J. S. a, Semantic Similarity from Natural Language and Ontology Analysis. s.l.:Morgan & Claypool, 2015.

[10] Kaggle, Twitter User Gender Classification. 2016.

[11] S. &. B. P. Mukherjee, 'Gender classification of microblog text based on authorial style', Inf. Syst. E-Bus. Manag., vol. 15, no. 1, pp. 117–138, 2016.

[12] F. Rangel, P. Rosso, M. Potthast, and B. Stein, 'Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter', p. 26.

[13] A. Bacciu, M. La Morgia, E. Nemmi, V. Neri, A. Mei, and J. Stefa, Bot and Gender Detection of Twitter Accounts Using Distortion and LSA Notebook for PAN at CLEF 2019. 2019.

[14] C. N. and C. R. S. K. P, 'Author Gender Identification from Text', Digit. Investig. Digit. Investig., vol. 8, no. 1, pp. 78–88, 2011.

[15] 'American Psychologist', Am. Psychiatr. Assoc., vol. 67, no. 1, pp. 10–42, 2012.

[16] M. Koppel and S. A. A. R. S, 'Automatically categorizing written texts by author gender', Lit. Linguist. Comput., vol. 17, no. 4, pp. 401–412, 2002.

[17] M. Koppel and J. S. S. A, 'Computational methods in authorship attribution', J. Am. Soc. Inf. Sci. Technol., vol. 60, no. 1, pp. 9–26, 2009.

[18] S. Goswami and S. S. M. R, Stylometric analysis of bloggers age and gender. San Jose, California, USA: Third International AAAI Conference on Weblogs and Social Media, 2009.

[19] S. Goswami and M. S, Fuzzy based approach to stylometric analysis of blogger's age and gender. Pune, India: IEEE, 2012.

[20] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta, Classifying latent user attributes in twitter. Toronto, ON, Canada, ACM, 2010.

[21] C. Fink, J. Kopecky, and M. Morawskib, 'Inferring Gender from the Content of Tweets: A Region Specific Example', in Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media, Dublin, Ireland, 2013.

[22] J. D. Burger, J. Henderson, G. Kim, and G. Zarrella, Discriminating gender on twitter. Stroudsburg, PA, USA, ACL, 2011.

[23] F. A. Zamal, W. Liu, and D. Ruths, Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors. Dublin, Ireland, AAAI, 2012.

[24] W. Deitrick, 'Gender Identification on Twitter Using the Modified Balanced Winnow', Commun. Netw., vol. 4, no. 3, pp. 189–195, 2012.

[25] Z. Miller, B. Dickinson, and W. Hu, 'Gender Prediction on Twitter Using Stream Algorithms with N-Gram Character Features', Int. J. Intell. Sci., vol. 2, pp. 143–148, 2012.

[26] W. Liu and D. Ruths, What's in a Name? Using First Names as Features for Gender Inference in Twitter. California, USA: AAAI, 2013.

[27] T. Schnoebelen and V. Kuperman, 'Using Amazon Mechanical Turk for linguistic research', Psihologija, vol. 43, no. 4, pp. 441–46, 2010.

[28] M. Buhrmester, T. Kwang, and S. D. Gosling, 'Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data?',

Perspect. Psychol. Sci., vol. 6, no. 1, pp. 3–5, 2011.

[29] D. Bamman, J. Eisenstein, and T. Schnoebelen, ‘Gender identity and lexical variation in social media’, J. Socioling., vol. 18, no. 2, pp. 135–160, 2014.

[30] P. Ludu, Inferring gender of a twitter user using celebrities it follows, ArXiv preprint. 2014.

[31] M. Merler, L. Cao, and J. R. Smith, ‘You are what you tweet... pic! gender prediction based on semantic analysis of social media images’, in Torino, Italy, Multimedia and Expo (ICME), 2015 IEEE International Conference, 2015.

[32] M. Vicente, F. Batista, and J. P. Carvalho, Twitter gender classification using user unstructured information. Istanbul, Turkey, IEEE, 2015.

[33] Z. Jianqiang, Pre-processing Boosting Twitter Sentiment Analysis? Chengdu, China: IEEE, 2015.

[34] S. Thomas Oshiobughie Ugheoke Regina, Detecting the Gender of a Tweet Sender. Regina, CA: University of Regina, 2014.

[35] K. Mouthami, K. N. Devi, and V. M. Bhaskaran, Sentiment analysis and classification based on textual reviews. Chennai, India: IEEE, 2013.

[36] Ankita and N. Saleenaa, ‘An Ensemble Classification System for Twitter Sentiment Analysis’, Procedia Comput. Sci., vol. 132, pp. 937–946, 2018.

[37] B. Y. Pratama and R. Sarno, Personality classification based on Twitter text using Naive Bayes, KNN and SVM. Yogyakarta, Indonesia: IEEE, 2015.

[38] M. Farahmand, Pre-trained Word Embeddings or Embedding Layer? — A Dilemma. 2019.

[39] RaRe-Technologies, How to Develop Word Embeddings in Python with Gensim. 2018.