Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN: 2581 - 611X)
VOLUME:8, ISSUE:2, AUGUST 2025, PAPER NUMBER:43

GENDER IDENTIFICATION OF AUTHOR FROM TEXT USING NLP

N. Shravani S[1], S. Chandana A[2], B. Sai Raman A[3], M Ms. M. kamala [4],

[1,2,3] Students, and [4] Faculty
Dept. of Computer Science Engineering,
CMR College of Engineering & Technology,
Hyderabad, India.
neelamshravani2002@gmail.com

Abstract: *Natural language processing, or NLP, has advanced significantly in recent years, enabling new uses outside text analysis. The ability to determine an author's gender from their written work is one such developing application. The method of determining an author's gender from their written work is known as "gender identification of author from text using NLP." To determine the gender of the author, language patterns, word choices, and stylistic elements must be examined. The goal of this talk is to highlight the developments in NLP techniques and algorithms that allow for precise gender assignment in text.*
Keywords: *Natural language processing, text Analysis, identification of author's gender, linguistic patterns, word choices and stylistic features.*

1. INTRODUCTION:

Authors that use NLP to identify their gender do so because they recognize that people often display different linguistic styles according to their gender, which are influenced by communication norms, personal experiences, and sociocultural variables. These variations show up in a number of language-related domains, such as discourse markers, narrative themes, word selection, grammatical structures, and sentiment expression. Researchers and practitioners are become more and more interested in the possibility of using computational tools to find these small but powerful indications as NLP continues to develop selection of words, grammatical constructions, discourse markers, expression of emotion, and even story ideas. Researchers and practitioners are become more and more interested in the possibility of using computational tools to find these small but powerful indications as NLP continues to develop.

Writers who use Natural Language Processing (NLP) to identify gender recognize the subtle linguistic patterns associated with various genders, influenced by social norms, individual experiences, and communication standards. These distinctive approaches manifest in various language-related contexts, including discourse markers, narrative themes, word selections, grammatical constructions, and sentiment expressions. The ongoing evolution of natural language processing (NLP) in examining word choice, grammatical constructions, discourse markers, emotional expression, and even the conceptualization of narrative ideas is reflected in the growing interest of researchers and practitioners in using computational tools to

identify these subtle yet powerful cues.

The possibility of automating the identification of gender specific linguistic patterns is becoming more alluring as NLP develops. Our investigation into this field made a workable solution necessary. Even though the JSON file includes author names, cultural influences make it difficult to determine gender based only on names. Using human annotators to evaluate names and provide the proper sex labels was one possible way to solve this problem. We decided against this strategy, nevertheless, for two main reasons. Considering the influence of cultural origins, gender determination from names can be complex. For example, a proficient annotator of Arabic and its dialects may find difficulty to categorize unknown names according to gender. Second, there were logistical difficulties in gathering a sizable number of participants for this assessment.

Unlike authorship attribution, gender identification is at a higher level of abstraction. The potential group of writers is not immediately accessible. Internet text messages are typically shorter than traditional text documents, such books, about which authorship attribution is primarily researched. Internet messages frequently contain unique language elements like emoticons, which are not present in traditional text publications. Texts on the Internet might have several formats or structures. owing to real-time constraints like Internet chat, instant messaging, etc., among various people and circumstances.

2. LITERATURE SURVEY:

A. EXISTING WORK

One of the most well-known social networking services, Twitter, lets users read and publish messages up to 280 characters on its platform. One of the most well-known social networking services, Twitter, lets users read and publish messages up to 280 characters on its platform Generally speaking, these communications are called tweets. The message size limit on Twitter is one of the factors contributing to its popularity among social media users. Global internet users have an equal chance to engage with other users on Twitter, including politicians, celebrities, and other well-known individuals, and to follow them on a regular basis by reading their tweets. Social media's rising ubiquity has produced a rare chance to gain broad knowledge about human culture. One of the most well-known marketing websites, Omnicore Agency, claims that over 330 million individuals use Twitter monthly.

The question of gender classification is becoming more and more relevant as social media users share more content. governmental organizations and marketing companies. Many businesses could benefit greatly from understanding this concealed content for a range of uses, such as forensics, election forecasting, cybercrime forecasting, and movie box office forecasting. We can ask human annotators to examine the names and assign the appropriate sex label as the received JSON file actually contains the Author Name. But we decided against moving on in this way for two reasons: First of all, because names are influenced by cultural background, it is not always easy to infer someone's sex from their name. For instance, an annotator who is proficient in the majority of Arabic dialects and speaks the language fluently would not be able to identify the gender attached to names that he is unfamiliar with. Recruiting enough volunteers for the assessment would not have been feasible.

This work will use various natural language processing (NLP) approaches on a text tweet using

a tagged Twitter dataset to train the classifier to automatically determine a user's gender using multiple Machine Learning (M L) methodologies. Classifying a sentence's or document's token (Inverse Document Frequency) is often necessary for most natural language processing (NLP) activities, including sentiment analysis. ML classification is then applied to train and forecast the output label.

B. RESEARCH OBJECTIVE

The primary goal of this study is to build the dataset. An approximately 20,000 user tagged dataset is offered by Kaggle . In order to create a tagged dataset for this study, We'll use this dataset as a bootstrap to get tweets for every one of those users. This study's main objective is to determine gender through the application of several machine learning (ML) and natural language processing (NLP) techniques. approaches that utilize text features to distinguish between distinct gender attributes. Numerous applications, such as psychological analysis and marketing, have been developed within the applied research field as a result of this sort of study.

Examining a user's past tweets may aid in understanding their linguistic style, particularly if they are interacting on Twitter while acting strangely. This analysis may facilitate comprehension of the user's background and gender (male or female), and it may also disclose patterns . With reference to referrals and promotion, In an attempt to reach the target gender, several organizations send digital adverts via Twitter. This has shown to be highly successful in terms of reaching a wider audience and generating more income, particularly for the e-commerce sector.

Moreover, a large number of bots on social media sites like Twitter are known to disseminate false information and fake news to the public, which may have an effect on elections and important campaigns. Because of this, it's critical to be able to distinguish between a human and a bot while writing these Tweets .

In practice, gender categorization for the two classes—male and female—given in the example can be viewed as a binary or two-class problem. Assigning an anonymous text or message to a class without knowing anything about the user is the goal of this categorization . Text analysis is a highly difficult Problem for machines, yet it is very simple for humans. It is frequently simple and quick for a human to discern the gender class by visual inspection. Text messages or language sentences serve as these systems' input. Generally speaking, the standard NLP machine learning task is to categorize a series of tokens, such as a document or phrase; that is, to approximate the function $f1 \rightarrow (1,0)$.

In this case, f1 could be determined by the domain, emotion, etc. The objective is to assign a number to each data point, representing a male or female category: 1 for males and 0 for females.

The sole distinction, though, is that it handles various text pairs in various tasks.

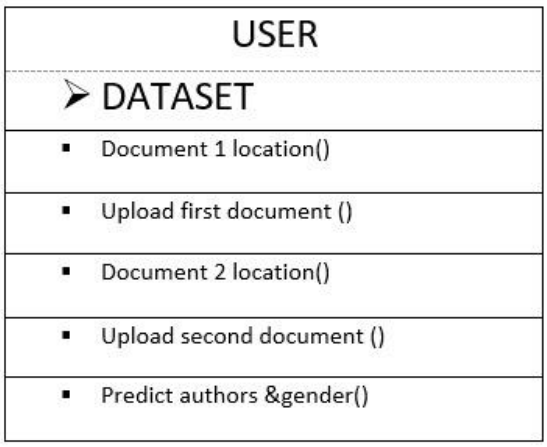
3. METHODOLOGY:

A.PROPOSED SYSTEM

In response to your request, we have used author profiling and gender text data to build Random Forest Machine Learning algorithms for this project. This trained model can be used to forecast the weather in any written content. Keep a record What gender is the author, and is it the same as before? Diverse linguistic motifs and styles are widely employed by writers of different genders in their writing. Word choices, sentence structures, sentiment expression, discourse markers, and word usage are some examples of these variations. Understanding and quantifying these differences is critical to the problem at hand. Gender-specific writing styles may be influenced by complex interactions between linguistic, psychological, and sociological factors. Given that these trends might be further modified by demographic factors such as age, education level, and cultural background, the problem is inherently complex. one area of research called Authorship Analysis (AA) looks for writers in the texts that they write. Authors Profiling (AP), a version of AA, focuses on identifying the characteristics of the authors. AP is quite significant since it allows for a fine-grained examination that often exposes discrepancies between the various author profiles. Authors who are similar to one another in terms of sex, for example, will exhibit distinguishing characteristics and stylometric markers in their writing that differentiate them from the other groups, according to the basic idea of author profiling. One of the author profiling topics that has been thoroughly studied is gender identity (GI)

Diverse gender writers frequently use diverse language motifs and styles in their works. These variances include word selections, sentence patterns, sentiment expression, discourse markers, and word usage. Comprehending and measuring these variations is essential to the issue at hand. There may be a complicated interaction between linguistic, psychological, and societal elements that affects gender-specific writing styles. The situation is made more complex by the fact that demographic factors like age, education level, and cultural background can further modify these patterns.

It was therefore necessary for us to devise a workable solution for this issue. We can ask human annotators to examine the names and assign the appropriate sex label as the received



JSON file actually contains the Author Name. But we decided against moving on in this way for two reasons: First of all, because names are influenced by cultural background, it is not always easy to infer someone's sex from their name. For instance, an annotator who is proficient in the majority of Arabic dialects and speaks the language fluently would not be able to identify

the gender attached to names that he is unfamiliar with. Recruiting enough volunteers for the assessment would not have been feasible.

B. SYSTEM ARCHITECTURE

The author gender identification can be treated as a binary classification. Given two classes {male, female}, assign an anonymous text message into 1 and 0 to one of these classes:

Class1 if the author of 1 is male

Class2 if the author of 0 is female

In order to create a hypothesis test, we must create a collection of characteristics that hold true for a sizable Number of messages authored by writers of the same gender. A given message, either 1 or 0, can be represented by a d-dimensional vector, where d is the total number of features, once the feature space has been constructed. A model (or classifier) is constructed based on a set of pre-classified messages that are known to exist, and it can be utilised to ascertain the category of a given message.

Pre-processing: of the dataset among the many forms of Internet text messages, newsgroup postings employ neutral, descriptive language, but more intimate, personal emails more accurately capture the author's actual essence. As a result, these two extreme dataset categories are used in the classifier design. Sentence-level writing style is captured by syntactic elements. Regular punctuation (commas, colons, etc.) and multiple question/exclamation marks (???,!!!) are examples of syntactic features. Writers frequently employ multiple question marks and exclamation marks to convey an attitude or mood in highly casual settings. Because men and women employ punctuation differently, syntactic features have the ability to discriminate between them. For instance, women tend to use more question marks than men do (Mulac, 1998).

A decision tree is a type of tree structure that resembles a flowchart and is created by looking at a measure of information acquisition. Each attribute (or feature) in a decision tree is represented as an internal node, each test's result as a branch, and the class label as a terminal node.class prediction is achieved by tracing a tree path from the root to a terminal node given a collection of attribute values. In Generally speaking, decision trees are a widely used categorization technique with a wide range of applications (Safavian and Landgrebe, 1991). Overfitting, however, could result from the data's large variation. The purpose of the ensemble learning technique is to enhance the classification. precision (Weiss and Damerau, 1998).

4. EXPERIMENTAL RESULT:

We used SVM, one feature set at a time, to the sub-dataset whose messages contain at least 100 words in order to examine the importance of the suggested feature sets.Table displays the categorization accuracy. It is evident that each of the five subgroups adds to the gender identification. It is demonstrated that a set of word-based characteristics and function terms are significant gender discriminators.

Accuracy comparison by using one feature subset at a time.	
Feature Subset	Accuracy (%)
Word based features	59.08
Character based features	73.48
Syntactic features	65.37
Structural features	61.26
Function words	74.81
All features	85.13

5. CONCLUSION:

Although the quality of the results was good, we attempted to increase the accuracy by adding more features to the text vectors, including word counts and average word lengths, but this had no beneficial effect on the classifier's overall performance. A technique based on deep learning may be investigated in the future. To evaluate the model's ability to handle Gen-der Identification from generic social media, such as Arabic textual content, it can also be applied to texts gathered from other social networks.

6. REFERENCES:

Mosteller F, Wallace DL. Applied bayesian and classical inference: the case of the federalist papers, ser. In: Springer series in statistics. Springer; 1984. Mulac A. The genderlinked language effect: do language differences really make a difference?; 1998. Mulac A, Lundell TL. Effects of genderlinked language differences in adults' written discourse: multivariate tests of language effects. Language and Communication 1994;14(3). Mulac A, Studley LB, Blau S. The gender-linked language effect in primary and secondary students'impromptu essays. Sex Roles 1990;23(9e10). Newman ML, Pennebaker JW, Berry DS, Richards JM. Lying words: predicting deception from linguistic styles. Personality and Social Psychology Bulletin 2003;29:665C675. Reuters corpora [Online]. Available, http://trec.nist.gov/data/ reuters/reuters.html; 2000. Bibliography of gender and language [Online]. Available, http://ccat.sas.upenn.edu/wharoldfs/popcult/bibliogs/gender/ genbib.htm; 2002, July. Linguistic inquiry and word count [Online]. Available, http:// www.liwc.net/; 2007, Jun. Enron e-mail dataset [Online]. Available, http://www-2.cs.cmu. edu/wenron/; 2005, April. Peng F, Schuurmans D, Keselj V, Wang S. Automated authorship attribution with character level language models. In: Proceedings of the 10th conference of the European chapter of the association for computational linguistics; 2003. Pennebaker J. Emotion, disclosure, and health; 1995. Pennebaker JW, Chung CK, Ireland M, Gonzales A, Booth RJ. The development and psychometric properties of LIWC2007. Austin, Texas: LIWC Inc; 2007. Rosenberg SD, Tucker GJ. Verbal behavior and schizophrenia: the semantic dimension. Archives of General Psychiatry 1978;36: 1331e7. Safavian SR, Landgrebe D. A survey of decision tree classifier methodology; May 1991. no. 3660e674. Talbot MM. Language and gender: an introduction. WileyBlackwell; 1998. Tweedie FJ, Singh S, Holmes DI. Neural network applications in stylometry: the federalist papers. The statistical study of literary vocabulary. Cambridge University Press; 1944. Zheng R, Li J, Chen H, Huang Z. A framework for authorship identification of online messages: writing-style features and classification techniques. Journal of the American Society for Information Science and Technology 2006;57(3): 378e93Kholoud Alsmearat, Mahmoud AlAyyoub, Riyad Al-Shalabi, and Ghassan Kanaan. Author gender identification from arabic text. Journal of Information Security and Applications. Emad AlSukhni and Qasem Alequr. Investigat- ing the use of machine learning algorithms in detecting gender of the arabic tweet author. International Journal of Advanced Computer Science and Applications, 7(7):319–328, 2016. Damian Radcliffe and Payton Bruni. State of social media, middle. Francisco Rangel, Paolo Rosso, Manuel Montes- y Gómez, Martin Potthast, and Benno Stein. Overview of the 6th author profiling task at pan 2018: multimodal gender identification in twitter. Boser B, Guyon I, Vapnik V. A training algorithm for optimalmargin classifiers. In: Proceedings of the 5th annual ACM workshop on computational learning theory. ACM Press; 1992. p. 144e52. Burrows J.Word patterns and story shapes: the statistical analysis

of narrative style. Literary and Linguistic Computing 1987;2:61e7. Chen H. Applying authorship analysis to extremist-group web forum messages. IEEE Intelligent Systems: Special Issue on AI for Homeland Security 2005;5(20):67e75. Cheng N, Cheng X, Chandramouli R, Subbalakshmi K.P. "Gender identification from e-mails," in IEEE Symposium on computational intelligence and data mining proceedings, 2009, pp. 154e158. Cortes C, Vapnik V. Support-vector networks. In: Machine learning; 1995. p. 273e97. Crawford M. Talking difference: on gender and language. London: Sage; 1995. Damerau Apte F, Weiss S. Text mining with decision trees and decision rules; 1998. Diederich J, Kindermann J, Leopold E, Paass G. Authorship attribution with support vector machines. Applied Intelligence 2000;19:109e23. Efron R, Thisted B. Estimating the number of unseen species: how many words did shakespeare know? Biometrika 1976;63(3): 435e47. Freund Y, Schapire RE. A decision-theoretic generalization of online learning and an application to boosting; 1995. Genkin Alexander, Lewis David, Madigan D, David. Large-scale bayesian logistic regression for text categorization. Technometrics 2007;49(3):291e304 [Online] Available, http:// dx.doi.org/10.1198/004017007000000245. Gottschalk LA, Gleser GC. The measurement of psychological states through the content analysis of verbal behavior. Berkeley: University of California Press; 1969. Holmes D. A stylometric analysis of mormon scripture and related texts, vol. 155. Royal Statistical Society; 1992. pp. 91e120. Holmes DI, Forsyth R. The federalist revisited: new directions in authorship attribution. Literary amnd Linguistic Computing 1995;10(2):111e27