TENCON 2021 - 2021 IEEE Region 10 Conference (TENCON)
7-10 Dec 2021. Auckland, New Zealand

Text-Based Gender Classification of Twitter Data using Naive Bayes and SVM Algorithm

Angelic Angeles
Department of Computer Science
National University, Manila
Manila, Philippines
angelesal@students.national-u.edu.ph

Maria Nikki Quintos
Department of Computer Science
National University, Manila
Manila, Philippines
quintosmh@students.national-u.edu.ph

Manolito Octaviano Jr.
Department of Computer Science
National University, Manila
Manila, Philippines
mvoctavianojr@national-u.edu.ph

Rodolofo Raga Jr.
Department of Computer Science
Jose Rizal University
Mandaluyong, Philippines
rodolfo.raga@jru.edu

TENCON 2021 - 2021 IEEE Region 10 Conference (TENCON) | 978-1-6654-9532-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/TENCON54134.2021.9707402

***Abstract*— This paper presents the development of the gender classification system on Twitter tweets. Three feature extraction techniques are explored: Bag of words and 2 variations of meta-attributes extraction. Feature sets are fed to Multinomial Naive Bayes and Support Vector Machine, and results were compared to see which algorithm can produce the best results in the classification task. Experiments show that the SVM outperformed the Naïve Bayes algorithm, obtaining a performance of 56.31%.**

Keywords: machine learning, classification, gender, social media, Twitter, tweets, extraction, meta-attributes, Naive Bayes, Multinomial Naive Bayes, Support Vector Machine (SVM).

I. Introduction

Social media are the platforms used to enable social networking or communication among other individuals, and it has taken over several individuals' lives; its influence continues to grow in society. Since we live in the digital age, an era where technology is used to disseminate information using technology has increased and continuously become an enabler to revolutionize the economy and digital transformation. The development of all these technological advancements in information and communication has led us to social evolution wherein social interactions between individuals arise and change. That is when social technologies or social networking have entered our daily lives. Twitter is one of the widely known social media platforms with 330 million monthly active users, 145 million daily active users, and 500 million tweets sent out per day as of 2020 [1]. It has become a major tool for social networking studies [2]. The user's interests and characteristics can be easily distinguished through their profiles, tweets, re-tweets, hashtags, etc., that a user share. That is why most researchers have also used Twitter as a tool to gather data and information [2].

This study aims to automate the characteristics of a user based on the users' writing style on Twitter and classify them in a demographic category, specifically in gender. In addition, the impact of different feature extraction techniques applied to several machine learning algorithms is observed.

II. Review of Related Works

Gender classification is a binary classification problem wherein two classes or categories, male or female, can be predicted based on the tweets. According to American Psychological Association, gender deals with the attitudes, feelings, and behaviors that a given culture associates with a person's biological sex. On the other hand, sex refers to a person's biological status and is typically assigned at birth (or before during ultrasound) based on the appearance of external genitalia [3]. The paper is based on the concept of gender rather than sex since we are dealing with the user's attitudes, feelings, and behaviors. Therefore, Gender Classification can be recognized as a gender detection tool of a user, whether the user is a male or female, by analyzing the content, attitude, and behavior exhibited based on the writing style of their tweets.

Existing works considered the use of psycholinguistic and sociolinguistic characteristics in determining behavior and/or gender [12; 13]. Extraction of features and creation of meta-attributes based on the characters, use of words or specific terms, writing style, and syntax used by the user is one of the techniques to discover the patterns of how the user tweet and distinguish the gender of the user [5], [11]. Some efforts exploit the advantage of the n-gram approach (word-level or character-level) [14; 15] while others focused on advancing the models, utilizing deep learning approaches [16; 17].

III. Methodology

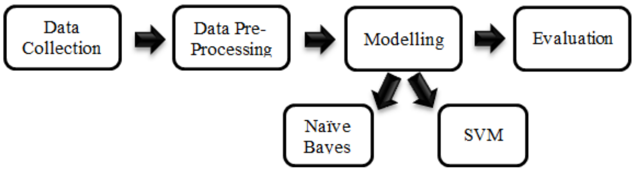

***Figure 1**. Gender Classification Development Framework*

The overview of the processes made for gender classification is shown in Figure 1. The development is divided into four stages: data collection, where the data is collected and used; data pre-processing, where the information is being cleaned and formatted; modeling, where the Naïve Bayes and SVM classifier learns the pattern of the training data; evaluation, where the test data is being tested to the generated model to check the performance of the model.

A. Data Collection

The dataset collected is from Kaggle datasets, a commonly used dataset by many researchers for gender classification. The dataset has 26 independent variables and 20,050 data. This research focuses on the text-based gender classification, which only gets the "gender" and "text" variables, wherein the "text" is consists of the user's tweets [18].

B. Data Pre- Processing

To classify the gender based on the writing style of the user's tweets; the dataset is cleaned to achieve the goals of the research. The following are the pre-processing techniques used to be able to achieve the text-based approach in classifying the gender:

- Two variables are used which are the "text" and "gender" variables from the 26 variables given on the dataset.
- Duplicate tweets are removed as this may cause a bias decision in classifying the gender.
- A sanity check is performed on the data retrieved and looks for any null values on the gender variable.
- Each tweet is converted into lowercase. Then, unnecessary characters such as hyperlinks and non-ASCII characters were removed.

Cleaned tweets were separated and grouped into two sets: male ("1") and female ("0"). To balance the train and testing data for the classifiers, each label set is split into 75:25 ratios. Then, 75% of the two sets were combined and used as training data of the classifier, while the remaining 25% as testing data. Overall, the training data consist of 9531, wherein 4611 tweets are labeled as male while 4920 are labeled female. On the other hand, the testing data consists of 3178, wherein 1525 tweets are labeled male while 1653 are female.

C. Feature Generation

Three different feature combinations are used in classifying the gender.

1) Bag-of-words Features

The simplest form of text representation is applied. Each tweet is represented as a bag of its word. This does not consider the order or semantics of the text.

2) Meta-attribute Features

For identifying the gender expression in English tweets, meta-attributes of tweets were extracted as shown in Table I. This will help define the gender linguistics cues that will represent a person's gender. Furthermore, the study adopts ten meta-attributes from the work of [11], as shown in Table II. These meta-attributes were the top attributes of the previous study that were ranked based on their significance.

From these features, two sets of feature combinations were generated: **META-1**, where the features are set of features in **Table II**; **META-2**, combinations of features in both **Table I** and **Table II**.

***Table I**. Generated Meta-Attributes*

Name	Description
Vowels	Total number of vowels in a sentence
Consonants	Total number of consonants in a sentence
Spaces	Total number of spaces in a sentence
Pronouns	Total number of pronouns in a sentence
Auxiliary Verbs	Total number of auxiliary verbs in a sentence
Conjunctions	Total number of conjunctions in a sentence
Interjections	Total number of interjections in a sentence
Prepositions	Total number of prepositions in a sentence

***Table II.** Meta-attributes features adopted in the work of Filo et. al [11]*

Characters and Syntax based Meta- Attributes	
C$_1$	Total number of characters
C$_5$	Ratio between the number of white spaces and the total number of characters
Word base Meta- Attributes	
W$_1$	Total number of words
W$_2$	Average number of characters per word
Textual Morphology based Meta- Attributes	
TM$_1$	Ratio between the total number of articles and the total number of words
TM$_2$	Ratio between the total number of pronouns and the total number of words
TM$_3$	Ratio between the total number of auxiliary verbs and the total number of words
TM$_4$	Ratio between the total number of conjunctions and the total number of words
TM$_5$	Ratio between the total number of interjections and the total number of words
TM$_6$	Ratio between the total number of prepositions and the total number of words

IV. Naïve Bayes and SVM Algorithm

The classifiers used in this work are the Naive Bayes and Support Vector Machine.

A. Naïve Bayes

Naive Bayes is a powerful probabilistic machine learning algorithm that is used for classification. Naive Bayes works by using the Bayes' Theorem which is predicting the probabilities for each class such as the probability of that given record or data point belongs to a particular class. The class or category with the highest probability is considered the most likely class [6]. Naive Bayes treats all features independent of each other. Even if these features depend on each other, all of these independently contribute to the probability of the outcome that is why it is known as 'Naive'. Naive Bayes is simple and performs well in many real-world problems such as document categorization, e-mail filtering, spam detection, etc. and it is known to outperform even highly practical classification methods. Naive Bayes works by using the Bayes' Theorem, it is calculated by getting the posterior probability P(c|x) from the product of likelihood P(x|c) and class prior probability P(c) divided by the predictor prior probability P(x) [8]. **Equation *1*** shows the formula for the Bayes theorem.

$$P(c|x) = \frac{P(x|c)\,P(c)}{P(x)}$$

***Equation 1**. Bayes Theorem*

Where:

- P(c|x) is the posterior probability of class wherein (c is the target) given predictor (x is attributes).
- P(c) is the prior probability of class.
- P(x|c) is the likelihood which is the probability of predictor wherein (x is attributes) given class (c is the target).
- P(x) is the prior probability of predictor.

This study will specifically implement Multinomial Naive Bayes; the only difference of it with Naive Bayes is that it is a multinomial distribution that requires integer feature counts, rather than some other distribution. This works well and is suitable for classification with discrete features, such as word counts in a text which is one of the features in this study [4]. Multinomial Naive Bayes is a naive bayes based algorithm that produces the frequencies of an event happening. This algorithm predicts a label based on how many times a feature has occurred. The only difference between the original formula of Naive Bayes compared to the Multinomial Naive Bayes is that C is now a sequence having a Ck wherein "k" is the index or counter to determine the times a feature has occurred [8]. Multinomial Naïve Bayes formula is shown in **Equation *2***.

$$p(C_k \mid x)\frac{p(C_k)\,p(x \mid C_k)}{p(x)}$$

***Equation 2.** Multinomial Naïve Bayes*

B. Support Vector Machine (SVM)

SVM (Support Vector Machine) - is a supervised machine learning algorithm used for classification and regression problems. However, SVM is used most likely in classification problems. SVM works by plotting each data item as a point in n-dimensional space (where n is the number of features or independent variables that the dataset has) with the value of a particular coordinate which is the value of each feature. Then, perform classification by finding the hyper-plane that differentiates the two classes very well [10]. Hyper-plane is a (n minus 1)-dimensional subspace for a n-dimensional space [10]. Look at how any hyper-plane is mathematically represented on **Equation *3***:

$$\beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \cdots + \beta_n * x_n = 0$$

***Equation 3**. Hyperplane Representation*

Where:

- β is each beta is a parameter for one of the many dimensions we have in our space.
- β0 is the intercept
- β1 is the first axis, and so on.
- X is a point X that satisfies the above equation then it means the point is on the hyperplane

The equation on **Equation *4*** is the equation of a hyperplane in a two-dimensional space, which is a line.

$$\beta_0 + \beta_1 * x_1 + \beta_2 * x_2 = 0$$

***Equation 4.** Hyperplane in a two-dimensional space*

V. Results and discussion

A. Parameter Searching

In this experiment, the tuning of parameter "C" or the amount of regularization is applied on the SVM classifier, and the tuning of parameter "Alpha" ranging from 1.0-10.0 which controls the form of the model itself is applied on the Naïve Bayes classifier with the corresponding feature extraction techniques used. The hyperparameters are experimented to find the best fit value for minimizing the type 1 and type 2 errors, and to further evaluate the highest possible performance that the SVM and Naïve Bayes can produce using the different values set on the hyperparameters. This experimentation will also help in deciding the best combination between the algorithm, and feature extraction technique used by only using the hyperparameter.

1) Bag-of-words Features

***Table III.** Parameter Searching using the Bag-of-Words Features*

SVM Algorithm				
c val / α value	**Acc**	**Pre**	**Rec**	**F1-score**
1.0	**59.09**	**57.75**	**54.95**	**56.31**
2.0	58.27	56.69	55.27	55.97
3.0	57.89	56.27	55.01	55.63
Naïve Bayes Algorithm				
1.0	60.85	62.69	45.5	52.73
2.0	59.88	63.46	38.62	48.02
3.0	59.18	63.9	34.36	44.69

Table III shows the results from the experimentation conducted on both models. The highest f1 score obtained by the SVM algorithm is when the *C* value is 1.0, obtaining an f1 score of 56.31%. This is because of the rule of regularization, which states that the higher the value of *C* the lower the regularization, and the lower the value of *C* the higher the regularization. Thus, when the *C* value is 1.0 which is the lowest value of the hyperparameter, *C* produced the highest performance score. Naïve Bayes on the other hand has also the same result when it comes to tuning the values of the hyperparameter *Alpha*. When the *Alpha* value is set to the lowest value, the model produced the highest f1 score of 52.73%. Overall, the SVM classifier obtained a higher f1 score compared to the Naïve Bayes classifier. This demonstrates that the SVM classifier when using the bag of words outperformed the Naive Bayes classifier.

2) Meta-attribute Features

The study compares the results of **META-1** and **META-2.**

***Table IV**. Comparison of the results on META-1 and META-2*

SVM Algorithm (META-1)				
c val / α value	**Acc**	**Pre**	**Rec**	**F1-score**
2.0	55.69	54.34	47.93	50.94
3.0	55.53	54.14	47.93	50.85
4.0	55.25	53.88	46.88	50.14
Naïve Bayes Algorithm (META-1)				
7.0	55.22	58.76	22.42	32.46
9.0	55.28	58.9	22.55	32.62
10.0	55.28	58.87	22.62	32.68
SVM Algorithm (META-2)				
4.0	56.16	56.3	38.62	45.81
3.0	56.13	56.24	38.68	45.84
10.0	56.19	56.33	38.75	45.92
Naïve Bayes Algorithm (META-2)				
2.0	56.35	54.71	52.45	53.56
7.0	56.38	54.75	52.52	53.61
9.0	**56.41**	**54.78**	**52.59**	**53.66**

Table IV shows the comparison of the results obtained by **META-1** and **META-2.** The SVM classifier on **META-1** obtained the highest f1 score of 50.94% when the *C* value is 2.0. Naïve Bayes on the other hand obtained the highest f1 score of 32.68% when *alpha* is 10.0. While on **META-2**, the result obtained by the SVM classifier achieved the highest f1 score of 45.92% when the *C* value is 10.0. The Naïve Bayes on the other hand obtained the highest f1 score of 53.66% when *alpha* is 9.0. It is observed that the Naïve Bayes outperformed the results obtained by other models using **META-2**. Overall, the Naïve Bayes classifier on **META-2** obtained the highest f1 score among the other models. Since multinomial naïve bayes requires integer feature counts, it works well and is suitable for classification with discrete features which is used in **META 2** data. It also predicts a label based on how many times a feature has occurred, thus this supports the result that Multinomial Naïve Bayes outperformed the SVM algorithm when features are used.

B. Feature Scaling

The study also explores the application of various feature scaling to see whether it will improve the model. **Table V** shows the result of feature scaling applied to the **META-1** and **META-2** features for the SVM and Naïve Bayes algorithm, respectively. The models used in feature scaling experiment are the best models obtained by each **META-1** and **META-2** experiments in **Table IV**.

***Table V**. Parameter searching on SVM and Naive Bayes Classifier*

SVM Algorithm (META-1)				
Feature Scaling	**Acc**	**Pre**	**Rec**	**F1-score**
Standard Scalar	55.28	53.79	48.39	50.94
MaxAbs Scalar	**55.66**	**53.94**	**52**	**52.95**
Naïve Bayes Algorithm (META-1)				
Standard Scalar	56.16	54.81	49.31	51.91
MaxAbs Scalar	53.14	62.32	0.59	10.8
SVM Algorithm (META-2)				
Standard Scalar	56.01	56.28	37.31	44.87
MaxAbs Scalar	56.01	56.12	38.16	45.43
Naïve Bayes Algorithm (META-2)				
Standard Scalar	54.27	52.46	50.22	51.32
MaxAbs Scalar	54.02	55.07	22.75	32.2

Feature scaling is applied to the models to improve the accuracy. Based on the results shown in **Table V**, the SVM classifier using **META-1** obtained the highest f1 score of 52.95% using the MaxAbs scalar. While Naïve Bayes, obtained the highest f1 score of 51.91% using standard scalar. On the other hand, the SVM classifier using **META-2** achieved the highest f1 score of 45.43% using the MaxAbs scalar. While Naïve Bayes achieved the highest f1 score of 51.32% using standard scalar.

Based on the experimentation results, the SVM classifier on both **META-1** and **META-2** obtained the highest f1 score using MaxAbs scalar, this is because *max abs* scale each feature by the maximum absolute value and translates each feature individually thus, it does not shift or center the data and it does not destroy the sparsity. While Naïve Bayes on both **META-1** and **META-2** obtained the highest f1 score using standard scaler this is because it transforms the data such that its distribution will have a mean value of 0 and standard deviation of 1, in other words, it makes the data independent to each column of the data. Overall, the highest f1 score is obtained by the SVM classifier on **META-1** using the MaxAbs scalar.

VI. CONCLUSION

This paper aims to analyze two different machine learning algorithms which are Naive Bayes specifically Multinomial Naive Bayes and Support Vector Machine (SVM) specifically Support Vector Classifier (SVC) to automate the characteristics of a user based on the users' writing style in social media specifically on Twitter and classify them in a demographic category specifically in gender. In addition, the impact of extracting the features using the techniques used on the application of the algorithm will be observed. While the experimentation will determine the best combination between the techniques used and the algorithm applied.

The experiments conducted show different results. On the bag of words and parameter searching technique, the SVM classifier outperformed the Naïve Bayes classifier. While on the meta-attribute features, wherein **META-1** and **META-2** are compared, the Naïve Bayes classifier outperformed the SVM classifier. Overall, SVM classifier outperformed the Naïve Bayes using the bag of words.

As for future works, we intend to explore other machine learning algorithms other than Naive Bayes and Support Vector Machine (SVM), using datasets with more user-profiles and tweets, as well as develop or create new meta-attributes based on psycholinguistics and sociolinguistic characteristics of the user such as dictionaries of negative, positive, and neutral words, part of speech (POS), semantics or meaning of the word which can improve the performance and results shown in the study.

REFERENCES

[1] Lin, Y. (2020, September 28). 10 Twitter Statistics Every Marketer Should Know in 2020 [Infographic]. Retrieved October 07, 2020, from https://www.oberlo.com/blog/twitter-statistics.

[2] Twitter gender classification using user unstructured information - IEEE Conference Publication. (2015, November 30). Retrieved October 12, 2020, from https://ieeexplore.ieee.org/document/7338102

[3] American Psychological Association. (n.d.). Definitions related to sexual orientation and gender diversity in APA documents. https://www.apa.org/pi/lgbt/resources/sexuality-definitions.pd

[4] Sklearn.naive_bayes.MultinomialNB¶. (n.d.). Retrieved October 13, 2020, from https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html

[5] Lopes Filho, José Ahirton & Pasti, Rodrigo & De Castro, Leandro. (2016). Gender Classification of Twitter Data Based on Textual Meta-Attributes Extraction. 10.1007/978-3-319-31232-3_97.

[6] How the Naive Bayes Classifier works in Machine Learning. (2017, February 19). Retrieved October 12, 2020, from https://dataaspirant.com/naive-bayes-classifier-machine-learning/

[7] Sunil RayI am a Business Analytics and Intelligence professional with deep experience in the Indian Insurance industry. I have worked for various multi-national Insurance companies in last 7 years. (2020, April 01). Learn Naive Bayes Algorithm: Naive Bayes Classifier Examples. Retrieved October 14, 2020, from https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

[8] JonathanRadotski. (n.d.). JonathanRadotski/multinomial_naivebayes. Retrieved October 14, 2020, from https://github.com/JonathanRadotski/multinomial_naivebayes

[9] Sunil RayI am a Business Analytics and Intelligence professional with deep experience in the Indian Insurance industry. I have worked for various multi-national Insurance companies in last 7 years. (2020, April 15). SVM: Support Vector Machine Algorithm in Machine Learning. Retrieved October 12, 2020, from https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/

[10] Sirohi, K. (2020, March 13). Support Vector Machine (Detailed Explanation). Retrieved October 14, 2020, from https://towardsdatascience.com/support-vector-machine-support-vector-classifier-maximal-margin-classifier-22648a38ad9c

[11] Filho, J. L., Pasti, R., & De Castro, L. (2016, March). Gender Classification of Twitter Data Based on Textual Meta-Attributes Extraction. Retrieved October 16, 2020, from https://www.researchgate.net/publication/293794120_Gender_Classification_of_Twitter_Data_Based_on_Textual_Meta-Attributes_Extraction

[12] Cheng, Na, Xiaoling Chen, Rajarathnam Chandramouli, and K. P. Subbalakshmi. "Gender identification from e-mails." In *2009 IEEE Symposium on Computational Intelligence and Data Mining*, pp. 154-158. IEEE, 2009.

[13] dos Santos, W. R., Ramos, R. M., & Paraboni, I. (2019). Computational personality recognition from facebook text: psycholinguistic features, words and facets. New Review of Hypermedia and Multimedia, 25(4), 268-287.

[14] Miller, Z., Dickinson, B., & Hu, W. (2012). Gender prediction on twitter using stream algorithms with n-gram character features.

[15] Nieuwenhuis, M., & Wilkens, J. (2018, September). Twitter text and image gender classification with a logistic regression n-gram model. In Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018).

[16] ElSayed, S., & Farouk, M. (2020). Gender identification for egyptian Arabic dialect in Twitter using deep learning models. Egyptian Informatics Journal, 21(3), 159-167.

[17] Schaetti, N. (2018, September). Character-based convolutional neural network and resnet18 for twitter authorprofiling. In Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018), Avignon, France (pp. 10-14).

[18] F.Eight, "Twitter user gender classification" *Kaggle*, 21-Nov-2016. [Online]. Available: https://www.kaggle.com/crowdflower/twitter-user-gender-classification. [Accessed: 10-Sep-2021].