IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

PREDICTION OF THE PERSONAL ATTRIBUTE OF TWITTER USER USING MACHINE LEARNING AND NATURAL LANGUAGE PROCESSING

[1]Shubhada Pandurang Marwadkar, [2] Vikas N. Honmane

[1] M. Tech Student, Dept. of Computer Science and Engineering, Walchand College of Engineering Sangli, India,
[2]Department of Computer Science and Engineering, Walchand College of Engineering, Sangli, India,
[1]Computer Science and Engineering,
[1]Walchand College of Engineering, Sangli, India

Abstract: **In modern society a large number of communication and interaction take place on various social media platforms such as Twitter and Facebook. For privacy reasons personally identifiable information like gender, age and occupation class of people is not available publicly. But precise prediction of such information is relevant in the area of advertising, forensics and business intelligence. In recent years user generated text and content has grown exponentially mainly in the form of blogs, tweets, comments and social networking messages. The rise in textual data has sparked interest in inducing user attribute such as age, gender, occupation using various machine learning algorithms and pre-processing of natural language processing.**

***Index Terms - Personal attribute, Machine Learning, Natural Language Processing, Deep Learning*.**

I. INTRODUCTION

For security reasons, information about users such as people's age and gender are not publicly available. In the field of advertising, forensics and business intelligence, however accurate prediction of this information has significant application. It could be extremely valuable for the evaluation of suspects to understand the user's profile based on the available text. Likewise, from a marketing point of view, marketers may be interested in knowing what kind of people like or hate their products based on the study of blogs and online product review. Twitter has a large number of different age, gender and professions users. Individuals in advertising and marketing are trying to find specific user features to exploit. It is therefore of great importance to consider the characteristics(attributes) of Twitter users. Compared with other social networking sites including Twitter and etc. Twitter has minimal user information which makes the task difficult. This work aims to predict the age, gender and occupation of Twitter users automatically. Existing approaches for this issue centered on studying classifications using user-related features like sociolinguistic, content-based, and network-based structure. The effect of these attributes on how other users interact with a user was not considered by most of the current systems. Here, together with the tweet content, we try to exploit mention tweets and list Twitter user membership details to infer age, gender and occupation. Understanding the user-generated text content with respect to age group, gender and profession this information would be useful for companies and even governments for recommendation, targeted advertising and policy formulation. Previous researcher takes user profile into account, which provides multiple features for classification such as considered the first name as an important feature in gender inference. Some researchers used color-based features, (e.g. sidebar color, background color) for pre-study and gender prediction. Description of users indicate the user gender through gender-based word (e.g. Man, woman, boy, girl) but resulting prediction low accuracy.

These individual characteristics aren't always available in full online social networks (e.g. Facebook, Twitter, LinkedIn). Firstly, customers can provide the easy-to-fill basic statistics that include name, gender, but rarely introduce their pastimes and various detailed information. Second, the maximum social media sites limit access to certain personal information due to privacy concerns. Based on our preliminary statistics on the accrued Google+ dataset consisting of 19,624 famous customers, third nearly 90 percent of personal gender statistics are provided, whereas only 12.36 percent of the person's birthday and 22.48 percent of user relationships are obtainable. [1] Previous studies have examined the relationships between personal attributes and behaviors. For example, how people communicate and write are recognized to be related to numerous non-public attributes, including educational heritage and increase environment. However, previous research has been limited via the supply of data. In the age of social media, people spontaneously submit and proportion linguistic expressions on-line, and these records may be used to infer personal attributes.

II. RELATED WORK

Recently, there is an eager interest in the automatic extraction of the user's personal attribute prediction. This information is useful to improve the recommendation system considerably to present better-suited content such as websites, brands, services, products. Different author uses differences between male and female bloggers in writing style and content as well as between users of different ages to assess an unknown user age and gender. As content-based features, N-gram words were used and POS n-grams were used as style-based features. N-gram derived applications top-word modelling used by different people groups. But in different contexts, many times the same words are used.

In recent years, Twitter has become a major tool for sharing events, expressing opinions and communicating with friends. Twitter has a large number of users with varied ages, gender, and professions. Because of this recent rise in the popularity and size of social media, there is a growing need for a system that can extract useful user's information from social media. People in advertising and marketing, try to find certain characteristics of user to target. Thus, finding characteristics (attribute) of Twitter users is of high importance.

The author in [2] objective was to predict latent user attributes, including: gender, age, regional origin and ideology, which is why five hundred users of each gender are annotated manually. The gender identification features are divided into four groups such as network structure, interaction activity sociolinguistic features, and user posting content. Both the characteristics of the network structure and the characteristics of contact activity had a common gender distribution. Using sociolinguistic features, they reported an accuracy of 71.8%, using n-grams they only achieved an accuracy of 67.7%. They reached 72.3% accuracy when using the stacked classification model based on Support Vector Machine (SVM) to combine n-gram features with sociolinguistic features. The study suggests that sociolinguistic features of Twitter are successful in the detection of gender. They achieved an accuracy of 72.3 percent when using the stacked classification model based on Support Vector Machine (SVM) to combine n-gram features with sociolinguistic features. The study suggests that sociolinguistic features of Twitter are successful in the identification of gender.

Some studies have recently suggested other features to infer gender. Reference [3] used a survey of 14,000 English Twitter users with almost 9 million tweets to look at the relation between gender, language style, and social networks. Usage of lexical functions and all user tweets, they achieved 88 per cent accuracy. Reference [3] suggests a tool for extracting user attributes from Twitter images. They created a database of users called 10 K with visual information containing tweets. They reached an accuracy of 76% by using visual graders with object quality semiconductor. The accuracy increased from 85 to 88 percent in addition to their textual recognition with visual information features.

The age prediction input options range from linguistic factors, network (e.g. contact variety, Fans Of Friends ratio) and user-related profile data (e.g. background picture, text color).[4] Some recorded works use linguistic features, likely due to difficulties in real-time network data storage, profile information unreliability along with advertising.

One of the most important work was performed in this regard as part of the World Well Being Project (WWBP)[5] where an interactive vocabulary analysis system was implemented, connecting a collection of individual terms, phrases, and topics from the open text context. Authors in [4] explicitly distinguished between four age ranges (age 13– 18, 19–22, 23–29, and 30–65) Authors in [5] analyzed the blog post and found that there is no change in the sharing of photos, but there is a slight increase in the use of postal URLs in terms of age apart from the unexplained age of 24, the most frequently reported sharing of links with users over the age of 35. Also supporting this result on URLs is [6] they continued to blog.

Authors in [6] analyzed the blog post and found that there is no improvement in photo sharing, although there is a small rise in the age-related usage of postal URLs apart from the unexplained age of 24, the most commonly recorded sharing of links with users over the age of 35 years. Also supporting this result on URLs is [7]they continued to blog research and found that age rises in the exchange of links. These discrepancies in the use of URLs and hashtags by age groups were exploited by age-related computer research but did not take into account the content associated with them, which also reflects age-related use. The published blog dataset of ICWSM 2009 improves the accuracy [8] of the Naive Bayes model learned with regard to content, slang words and stylistic features. Increase the accuracy of the baseline bag-of-words model by using coevolutionary neural network. Improve these capabilities with support for URL and hashtag. The challenges of automating recognition of occupations using Twitter knowledge are complex and their interactions make it more difficult to establish explicit rules for recognizing occupations within Twitter users profile description area. [9]In recent years, the prediction of personal attributes based totally on social facts has to turn out to be a major place of studies. Previous studies have examined personal attribute can be predicted simplest from the texts in social media posts. This suggests that by adding other features such as images or URLs embedded in tweets, much better precision can be achieved[1].

III. METHODOLOGY

A. DATA

To identify the attribute of the user such as gender, age group and occupation you need a twitter user id dataset, specific tweet count and uploaded images. All these live information's of particular user collect through the using of the twitter API.

B. DATA PRE-PROCESSING

For gender, classification removes URL, hashtag, punctuation sign, stop word and repeated letters. Classification of the age group increase in the use of URLs in posts in terms of age [6] so for age classification do not remove URL and hashtag the same as for occupation class classification.

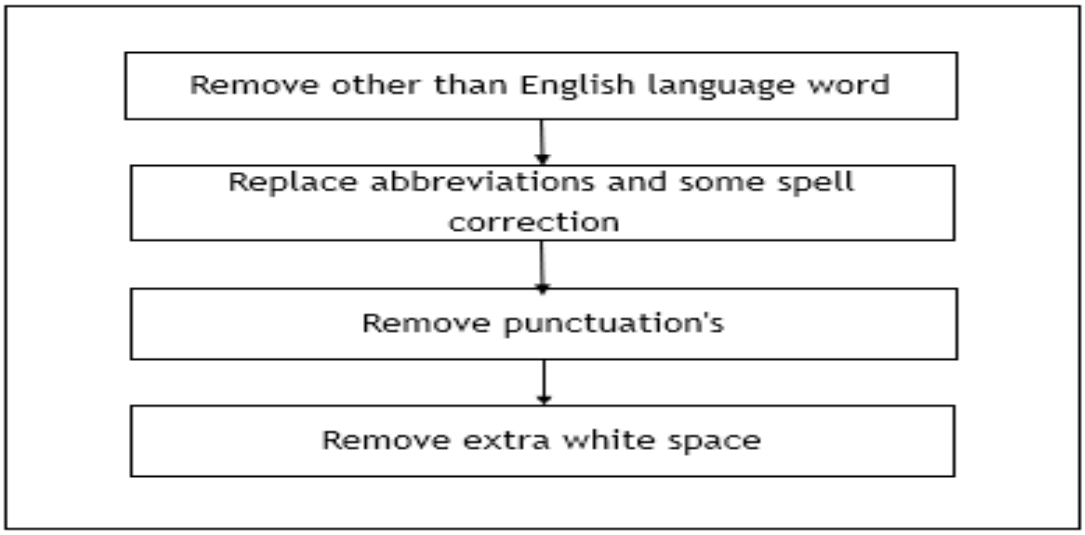

Figure 1: Data pre-processing Flow

C. FEATURE EXTRACTION

Attribute can be inferred through tweet content and uploaded images. Start by pre-processing the text to derive textual characteristics from tweets. Retweets are ignored and pre-processed texts are used for extracting word-based unigrams, bigrams, and trigrams. The second approach is according to the previous result identification of the gender-based on the image use the VGG16 and ResNet50 from ImageNet and after that combine the result of the second approach and tweet content result to identify the gender. Based on the content data identify the age and the occupation class. For content-based feature extraction Count vector, tfidf out of these two tfidf result is best as compare to the count vector.

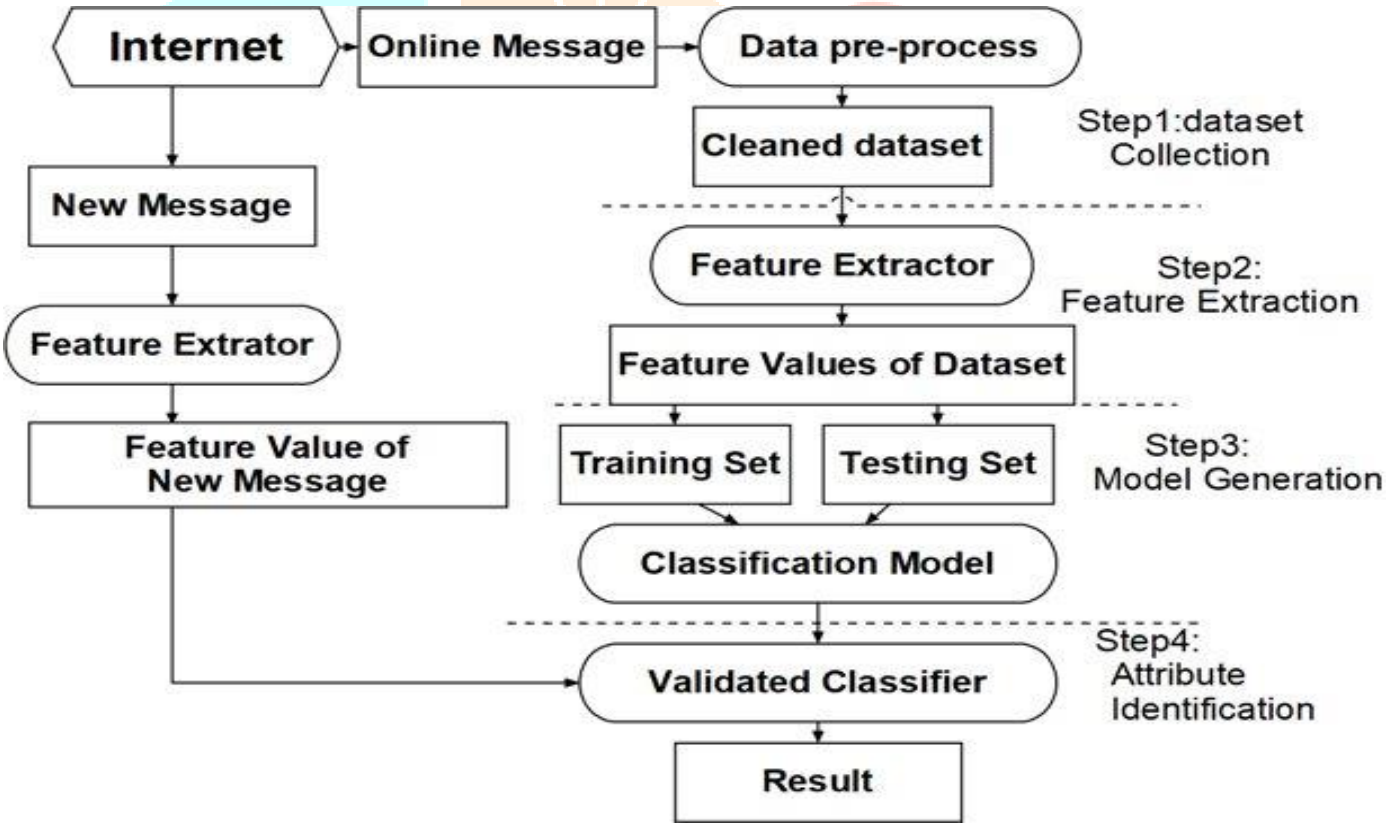

Figure 2: System Architecture using Machine Learning

a) In Count Vectorizer we only count the number of times a word occurs in the document resulting in the most commonly used words being biased. this ends up in ignoring rare words which could have helped is in processing our data more efficiently.

b) To overcome this, use TfidfVectorizer.

c) In TfidfVectorizer consider overall document weightage of a word. It helps us in dealing with most frequent words. Using it we can penalize them. TfidfVectorizer weights the word counts by a measure of how often they appear in the documents.

D. SUPPORT VECTOR MACHINE

The SVM algorithm's Is to find a hyperplane in the N-dimensional space (N-number of features) which separately classifies the attribute. To find the plane with the highest margin the total distance between both class data points.

Linear Model

$w.x-b=0$

$w.\ x_i-b>=1$ if $y_i=1$

$w.\ x_i-b<=-1$ if $y_i=-1$

so put this in $y_i(w.\ x_i-b)>=1$ one equation then multiply linear function with class label and this should be greater than or equal to one so this is the condition that want to satisfies and you want to come up with w weight and b bias and for this you use cost function and then apply gradient descent. cost function in this case use hinge loss and this is defined as

$$l=\max(0,1-y_i(w.\ x_i-b)$$

$$l=\begin{cases} 0 & \text{if } y.f(x)>=1 \\ 1-y.f(x) & \text{otherwise} \end{cases}$$

Add Regularization: Regularization balance between margin maximization and loss.

$$j=\lambda||w||^2+1/n\sum_{i=1}^{n}\max(0,1-y_i(w.\ x_i-b))$$

if $y_i.f(x)>=1$:

$$J_i=\lambda||w||^2$$

Else

$$J_i=\lambda||w||^2+1-y_i(w.\ x_i-b)$$

Gradients: Optimize objective function using gradient descent

If $y_i.f(x)>=1$:

$$\frac{dJ_i}{dw_k}=2\lambda w_k$$

$$\frac{dJ_i}{dw_k}=0$$

else:

$$\frac{dJ_i}{dw_k}=2\lambda w_k - y_i x_i$$

$$\frac{dJ_i}{dw_k}=y_i$$

Update rule:

$w=w-\alpha dw$

$b=b-\alpha db$

E. Naive Bayes:

This is a method of classification based on the theorem of Bayes with a predictor independence assumption. Bayes theorem provides a means of calculating the probability of P (C — X) posterior to P(C), P(X) and P(X — C). Check at below equation:

$$P(C|X) = P(C)P(X|C)\ P(X)$$

where: P(C|X)=Posterior Probability P(C)= Class Prior Probability P(X|C)=Likelihood P(X)=Predictor Prior Probability

F. K-Nearest Neighbours:

K-nearest neighbours uses the local neighbourhood to obtain a prediction A distance function is needed to compare the examples similarity.

Euclidean distance $(d(x_j, x_k) = \sqrt{\sum_i (x_{j,i} - x_{k,i})^2})$

Manhattan distance $(d(x_j, x_k) = \sum_i |x_{j,i} - x_{k,i}|)$

IV. RESULT:

Following results are achieved in the experimentation using Python on Anaconda platform. Experimental results for various algorithm are revealed in Figure 3, 4 and

	PRECISION	RECALL	F1-SCORE
MALE	0.71	0.80	0.75
FEMALE	0.77	0.67	0.71

Table 1: Performance of Gender Classification using SVM

	PRECISION	RECALL	F1-SCORE
MALE	0.75	0.75	0.75
FEMALE	0.80	0.80	0.80

Table 2: Performance of Gender Classification using SVM (Without library)

	PRECISION	RECALL	F1-SCORE
MALE	0.67	0.67	0.67
FEMALE	0.67	0.67	0.67

Table 3: Performance of Gender Classification using Naïve Bayes

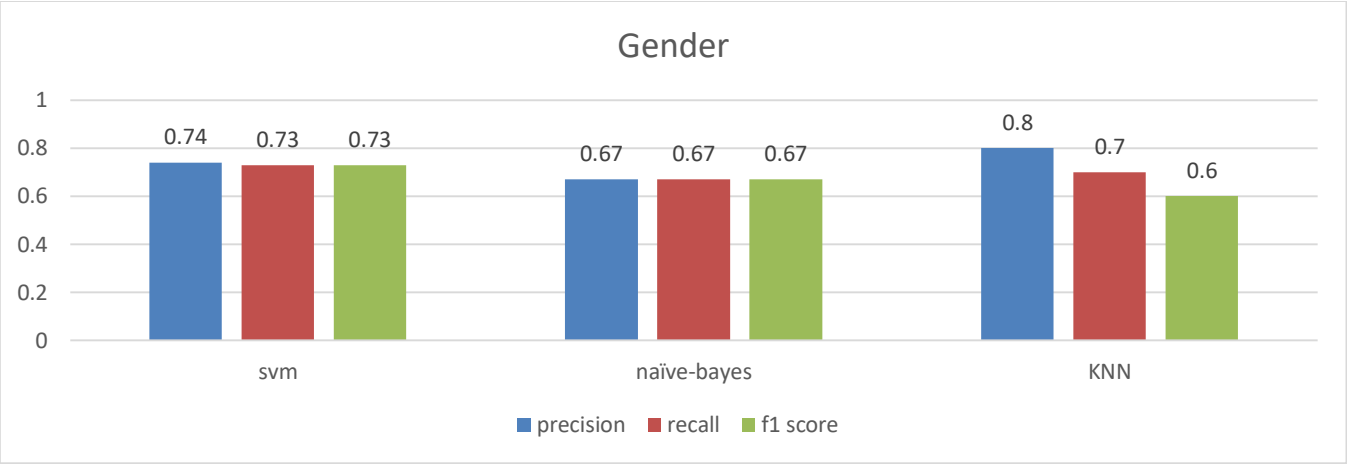

Figure 3: Performance evaluation for different algorithms for Gender classification

	PRECISION	RECALL	F1-SCORE
DIGITAL NATIVES	0.50	0.40	0.44
DIGITAL IMMIGRANTS	0.79	0.85	0.81

Table 4: Performance of Age Classification using SVM

	PRECISION	RECALL	F1-SCORE
DIGITAL NATIVES	1.00	0.14	0.25
DIGITAL IMMIGRANTS	0.57	1.00	0.73

Table 5: Performance of Age Classification using Naïve Bayes

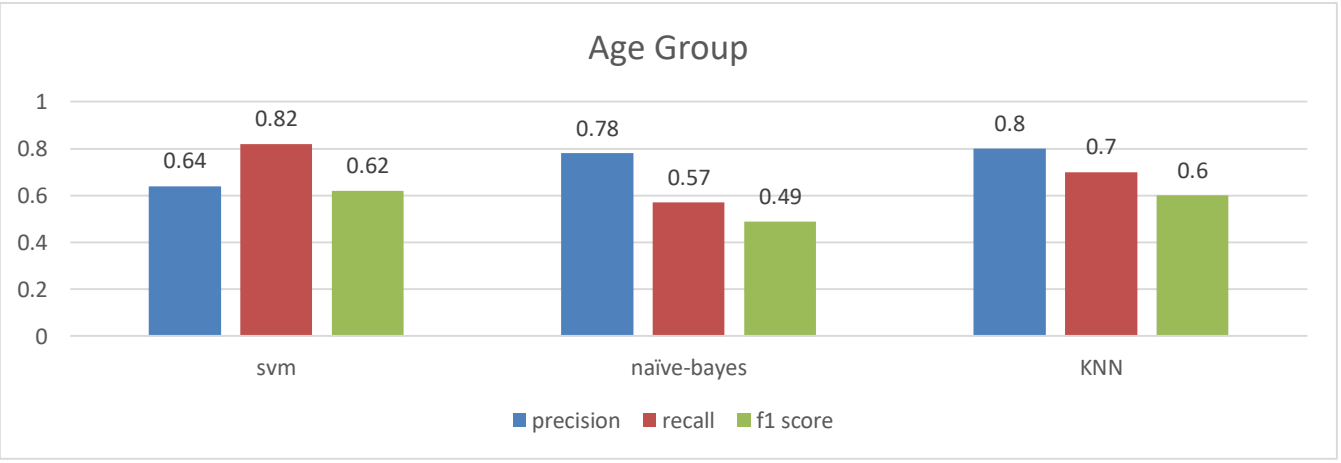

Figure 3: Performance evaluation for different algorithms for Age class classification

	PRECISION	RECALL	F1-SCORE
SPORTS	0.71	1.00	0.83
ACTOR	0.50	0.33	0.40
SINGER	0.33	1.00	0.50
PHOTOGRAPHER	1.00	0.67	0.80
POLITICS	1.00	0.67	0.80

Table 6: Performance of Occupations Classification using SVM

	PRECISION	RECALL	F1-SCORE
SPORTS	0.50	1.00	0.67
ACTOR	0.50	1.00	0.67
SINGER	1.00	0.50	0.67
PHOTOGRAPHER	1.00	0.43	0.60
POLITICS	1.00	0.56	0.71

Table 7: Performance of Occupations Classification using Naïve Bayes

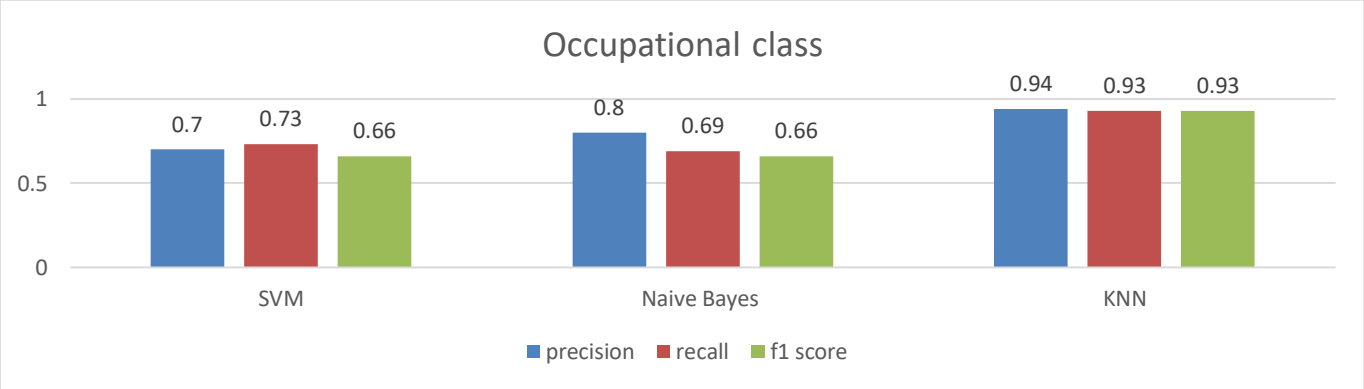

Figure 3: Performance evaluation for different algorithms for Occupation class classification

V. CONCLUSION:

Using machine learning algorithms and natural language processing we understand the inference of personal attributes from the text of tweets. We applied different word embedding method and supervised classification algorithms. The resulting word embedding were used as input for training models. To enhance prediction accuracy of those models by adding other features such as image, and embedded tweets.

REFERENCES

[1] Q. Fang, J. Sang, C. Xu, and M. S. Hossain, "Relational user attribute inference in social media," *IEEE Trans. Multimed.*, vol. 17, no. 7, pp. 1031–1044, 2015, doi: 10.1109/TMM.2015.2430819.

[2] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta, "Classifying latent user attributes in Twitter," *Int. Conf. Inf. Knowl. Manag. Proc.*, pp. 37–44, 2010, doi: 10.1145/1871985.1871993.

[3] 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, doi: 10.1111/josl.12080.

[4] M. Merler, L. Cao, and J. R. Smith, "You are what you tweet⋯pic! gender prediction based on semantic analysis of social media images," *Proc. - IEEE Int. Conf. Multimed. Expo*, vol. 2015-Augus, 2015, doi: 10.1109/ICME.2015.7177499.

[5] H. A. Schwartz *et al.*, "Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach," *PLoS One*, vol. 8, no. 9, 2013, doi: 10.1371/journal.pone.0073791.

[6] J. D. Burger and J. C. Henderson, "An exploration of observable features related to blogger age," *AAAI Spring Symp. - Tech. Rep.*, vol. SS-06-03, pp. 15–20, 2006.

[7] V. Bongard, A. Y. McDermott, G. E. Dallal, and E. J. Schaefer, "Effects of age and gender on physical performance," *Age (Omaha).*, vol. 29, no. 2–3, pp. 77–85, 2007, doi: 10.1007/s11357-007-9034-z.

[8] M. Rustagi, R. R. Prasath, S. Goswami, and S. Sarkar, "Learning age and gender of blogger from stylistic variation," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 5909 LNCS, pp. 205–212, 2009, doi: 10.1007/978-3-642-11164-8_33.

[9] Z. Jianqiang, G. Xiaolin, and Z. Xuejun, "Deep Convolution Neural Networks for Twitter Sentiment Analysis," *IEEE Access*, vol. 6, pp. 23253–23260, 2018, doi: 10.1109/ACCESS.2017.2776930.

[10] K. Santosh, A. Joshi, M. Gupta, and V. Varma, "Exploiting wikipedia categorization for predicting age and gender of blog authors," *CEUR Workshop Proc.*, vol. 1181, no. 1, pp. 33–36, 2014.

[11] M. Pennacchiotti and A. Popescu, "A Machine Learning Approach to Twitter User Classification," *Proc. Fifth Int. AAAI Conf. Weblogs Soc. Media A*, pp. 281–288, 2011.

[12] R. G. Guimarães, R. L. Rosa, D. De Gaetano, D. Z. Rodríguez, and G. Bressan, "Age Groups Classification in Social Network Using Deep Learning," *IEEE Access*, vol. 5, no. c, pp. 10805–10816, 2017, doi: 10.1109/ACCESS.2017.2706674.

[13] G. Laboreiro, L. Sarmento, J. Teixeira, and E. Oliveira, "Tokenizing micro-blogging messages using a text classification approach," *Int. Conf. Inf. Knowl. Manag. Proc.*, no. December 2016, pp. 81–87, 2010, doi: 10.1145/1871840.1871853.

[14] D. Preot'iuc-Pietro, V. Lampos, and N. Aletras, "An analysis of the user occupational class through Twitter content," *ACL-IJCNLP 2015 - 53rd Annu. Meet. Assoc. Comput. Linguist. 7th Int. Jt. Conf. Nat. Lang. Process. Asian Fed. Nat. Lang. Process. Proc. Conf.*, vol. 1, pp. 1754–1764, 2015, doi: 10.3115/v1/p15-1169.

[15] M. Vicente, F. Batista, and J. P. Carvalho, *Gender detection of twitter users based on multiple information sources*, vol. 794. Springer International Publishing, 2019.

[16] M. M., T. J., and V. K.R., "Techniques of Sentiment Classification, Emotion Detection, Feature Extraction and Sentiment Analysis A Comprehensive Review," *Int. J. Comput. Sci. Eng.*, vol. 6, no. 1, pp. 244–261, 2018, doi: 10.26438/ijcse/v6i1.244261.