Indian Institute Of Information Technology

Allahabad

5th Semester Mini Project

Unintentional Personal Information Leakage on Twitter

Submitted To:
Dr. Bibhas Ghoshal

Submitted By:
Shruti Reddy (IIT2016019)
Garima Chadha (IIT2016020)
Vaibhav Srivastava (IIT2016034)
Niharika Shrivastava
(IIT2016501)

CERTIFICATE FROM SUPERVISOR

I hereby recommend that the mini project report prepared under my supervision, titled **"Unintentional Personal Information Leakage on Twitter"**, be accepted as it fulfills the requirements of the mini-project completion of fifth semester of Bachelor of Technology in Information Technology.

Date:
Place: Allahabad

Supervisor :
Dr. Bibhas Ghoshal
. . . .

Abstract

Due to security issues of user information on the internet, and various threats involved in exploiting user data for applications in the field of targeting advertisements to desired customers, enhancing business logistics and/or data usage in forensics, netizens have restricted providing their personal details to a bare minimum.

But, predicting and/or determining certain social attributes of users on the internet has many positive implications too. Advertisements use the concept of data re-targeting and re-marketing to provide users relevant recommendations based on their interests and choices. Enhancement of search engine results for queries like “neighboring eateries” or “fun events” is made possible because of predicting trivial user location or age.

In this project, we construct methods to predict various attributes that have not been publicly mentioned in user profiles such as gender, interests/hobbies, and occupation of users using Twitter as our online social network (OSN). This will in turn help us to establish various use cases for the new user profile generated.

The main agenda of this work is to predict the Gender, Interest/Hobbies and Occupation of Twitter Users on the basis of their Tweets using Natural Language Processing Techniques.

Contents

1 Introduction

Presently, information available on Online Social Networks is a huge source of data. This data can be used to predict hidden attributes of social media users and therefore acts as a point of attraction for various groups of people. From view of business logistics, this data can be used to target users for advertisement purposes. Similarly it can also be used in enhancement of search engines by filtering results according to user preferences.

In this work, we use Twitter (Online Social Network) as the source for collection of user information in the form of Tweets. This is done by making repeated calls to the Twitter API using Tweepy to communicate with Twitter platform. These Tweets are processed using Natural Language Processing to predict Gender, Interest/Hobbies and Occupation of Twitter users.

A dataset labelled with gender of Twitter users is procured from Data For Everyone Library on Crowdflower. This is labeled with a list of users along with some other attributes. For training and testing purposes, we required a list of twitter handle with tweets by a certain user and a label to feed into the classifier. We extracted tweets corresponding to every Twitter Handle and stored this information in a csv file. With the help of TF-IDF and POS (Part of Speech) N-GRAM, the dataset was trained. The dataset was trained using Naive-Bayes Classifier.

POS N-GRAM - This is a method which generates continuous phrase on n-words out of a given sentence . Thus vectors are created that contain unigram (one word) , bi-gram (two words) and tri-gram (three word) of the tweets in the dataset.

TF - IDF - This stands for "Term Frequency - Inverse Document Frequency". This is used as a weighing factor attached to various features produced after creation of N-Gram words. Term Frequency refers to a labeled hashing of N-Gram phrases, i.e TF is the frequency of the features in the Tweet text. A problem which persists is that Stopwords like "the", "and", have a greater frequency than other words which are significant to the Tweet text. Inverse Document Frequency counters this issue. It refers to logarithmic ally scaled inverse fraction of Tweets which contain certain features. TF - IDF is the product of the TF and IDF values calculated separately.

2 Motivation

With the growing use of internet and in turn, social networking sites, communication flow between users has increased extensively in the form of blogs, wall posts, messages, connections, broadcasts, etc. Naive OSN (Online Social Network) users are unaware of the hidden patterns in their public information that might reveal unintentional or involuntary data.

In this project, we devise a method to reveal certain user attributes such as gender, interests/hobbies, and occupation. This data can be utilized in various fields of interests like improving search results, marketing purposes, and sentiment analysis. It also provides an idea of how vulnerable user data can be on the internet, and how easily privacy can be compromised.

3 Literature review

There have been number of research efforts to improve the accuracy of our classifier model, i.e., by cleaning data appropriately.

When it comes to OSNs, information exposure is usually a plus or even a must for users to join a new community [1].

An OSN usually encourages users to expose personal information because self-information disclosure can build trust, strengthen the ties between people, and bind romantic relationships or friendships[2],[3].

The usage of OSNs usually raises serious concern about the overall privacy, and it is essential to protect personal identifiable information (PII) [4], which alone or combined with other public information, can be used to distinguish or trace an individual's identity [5].

The present paper focuses on the task of gender classification by using 60 textual meta-attributes, commonly used on text attribution tasks, for the extraction of gender expression linguistic cues in tweets written in Portuguese. Therefore, taking into account characters, syntax, words, structure and morphology of short length, multi-genre, content free texts posted on Twitter to classify author's gender via three different machine-learning algorithms as well as evaluate the influence of the proposed meta-attributes in this process.[6]

Graph embedding place Twitter users in a vector space where similar users are likely to be close to each other. The user graph embedding are treated as features to train linear and non-linear supervised models for predicting income and occupational class.[7]

4 Problem Definition

To predict gender , interest and occupation of a user using tweets available on Twitter which can be further used for enhanced user experience on the web.

Input :
Input required is a list of twitter handles labeled with the corresponding value of gender or interest or occupation (labels).

user_handle - Handle of an existing Twitter user.

Data :
Data having Tweets of users is cleaned by removing punctuation marks, stopwords and URLs. TF-IDF matrix is created using the transformer available in sklearn module of python. This is fed into Naive Bayes Classifier for training purpose.

tweet_list - List of Tweets of all users.
count_vectorizer[ ] - Creating N-GRAMS from the tweets.
tf_idf_matrix[ ] - Generating TF-IDF matrix.

Output :
Prediction of following attributes is done by the trained classifier.

Gender- predicted gender of Twitter user.
Interest - predicted interest of Twitter user.
Occupation - predicted occupation class of Twitter user.

5 Software Requirements

5.1 Python

Various Python modules such as scikit-learn, nltk, tweepy, csv, pandas, matplotlib, collections etc. have been used for prediction of required attributes.

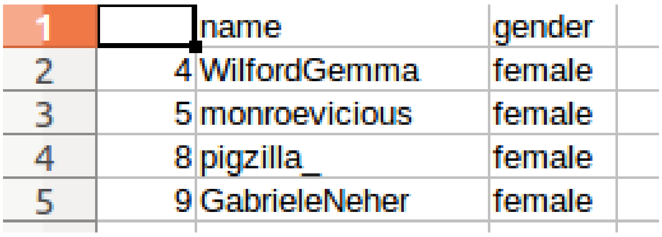
5.2 Twitter API

This has been used to mine Tweets for the purpose of data collection.

6 Dataset

6.1 Labeled Dataset

To create the dataset used for training, we must have a labeled table of users. The kind of labeled data is available in Data For Everyone Library on CrowdFlower for about 20000 Twitter Users. We extracted name (Twitter Handle) along with the label of gender. This is shown in the figure 1.

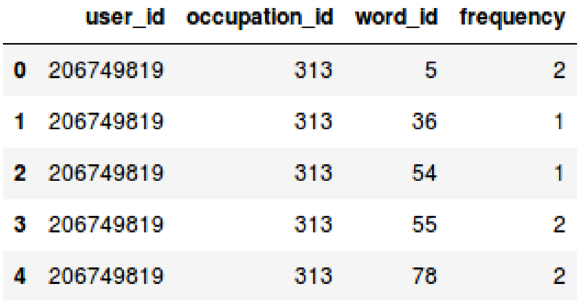


1		name	gender
2	4	WilfordGemma	female
3	5	monroevicious	female
4	8	pigzilla_	female
5	9	GabrieleNeher	female

Figure 1: Required Labeled Data

The data collection for interest is performed by creating about 12 classes like gaming, music, etc. Data is collected by mining Tweets from famous Twitter Handles that fall in these classes. Interest of a Twitter User is predicted as one of these classes.

The Standard Occupational Classification (SOC) is a UK government system developed by the Office of National Statistics for classifying occupations. It has 900+ classes in total, each with a unique id. A mapping between Twitter user and his/her occupation class is made, along with the tf-idf of their tweets.

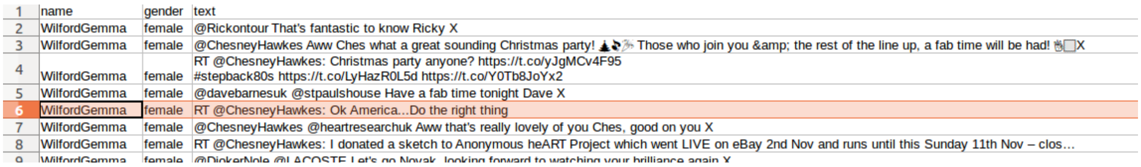


	user_id	occupation_id	word_id	frequency
0	206749819	313	5	2
1	206749819	313	36	1
2	206749819	313	54	1
3	206749819	313	55	2
4	206749819	313	78	2

Figure 2: Required Labeled Data

6.2 Tweet extraction using Twitter API

Next we use the Twitter Handle of the user to extract tweets from his/her twitter account using twitter API. A CSV file contains twitter handle along with the user tweets. A figure 2. depicting tweets extracted along with user name is shown below.



1	name	gender	text
2	WilfordGemma	female	@Rickontour That's fantastic to know Ricky X
3	WilfordGemma	female	@ChesneyHawkes Aww Ches what a great sounding Christmas party! 🎄🎶✨ Those who join you &amp; the rest of the line up, a fab time will be had! 👏□X
4	WilfordGemma	female	RT @ChesneyHawkes: Christmas party anyone? https://t.co/yJgMCv4F95 #stepback80s https://t.co/LyHazR0L5d https://t.co/Y0Tb8JoYx2
5	WilfordGemma	female	@davebarnesuk @stpaulshouse Have a fab time tonight Dave X
6	WilfordGemma	female	RT @ChesneyHawkes: Ok America...Do the right thing
7	WilfordGemma	female	@ChesneyHawkes @heartresearchuk Aww that's really lovely of you Ches, good on you X
8	WilfordGemma	female	RT @ChesneyHawkes: I donated a sketch to Anonymous heART Project which went LIVE on eBay 2nd Nov and runs until this Sunday 11th Nov – clos...
9	WilfordGemma	female	@DjokerNole @LACOSTE Let's go Novak, looking forward to watching your brilliance again X

Figure 3: Tweets of user

6.3 Cleaning of Data

Next step is to clean the data for proper training purpose. Cleaning of data refers to removing URL's from the tweets. After removal of URL's, we use tokenizer to count the frequency of different unigram , bigram and trigram phrases. So we get the mapping of different n-grams along with frequency with which they are present. Stopwords like "the","you" ,"and",etc are also removed so that they may not affect training.

6.4 Generating TF-IDF Matrix

Finally, the tf-idf matrix using tf-idf transformer (available in sklearn library) is created over the mapping. We are now left with tf-idf value of every unigram , bigram and trigram values along with the required label like male or female (in case of gender), occupational class code(in case of occupation) and interest group (in case of hobbies) to train the machine.TF-IDF can be expressed mathematically as:

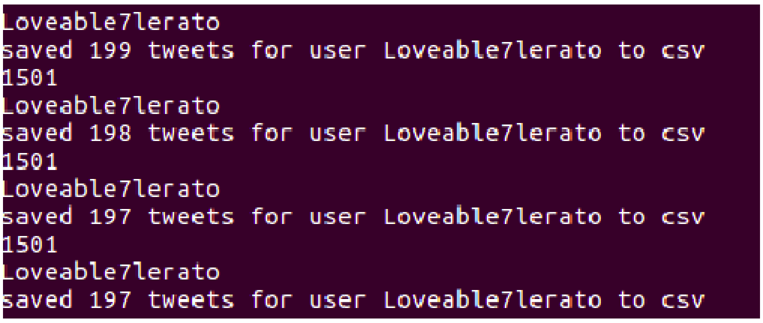
$$\mathbf{tf_{i,j}} = \frac{n_{\mathrm{i,j}}}{\sum\limits_{k} x_{\mathrm{i,j}}}$$

$$\mathbf{idf(w)} = \log \frac{N}{df_{\mathrm{t}}}$$

$$\mathbf{tf\text{-}idf(w)} = tf_{\mathrm{i,j}} \times idf(w)$$

i = word
j = document
$\mathbf{tf_{i,j}}$ = number of occurrences of i in j
$\mathbf{df_i}$ = number of documents containing i
N = total number of documents

Figure 3 below shows the data-set being collected for a user.



```
Loveable7lerato
saved 199 tweets for user Loveable7lerato to csv
1501
Loveable7lerato
saved 198 tweets for user Loveable7lerato to csv
1501
Loveable7lerato
saved 197 tweets for user Loveable7lerato to csv
1501
Loveable7lerato
saved 197 tweets for user Loveable7lerato to csv
```

Figure 4: Data Extraction in process

7 Methodology And Implementation

After the generation of a cleaned dataset consisting of Twitter Handle, Tweets of a user and a label of gender or interest or occupation of the user, we use the tf_idf_matrix for training the machine. We have used Multinomial Naive Bayes Classifier for training purpose. Next, testing is performed and the confusion matrix is generated. Precision and Recall values are also calculated based on the confusion matrix.

7.1 Methodology

7.1.1 Multinomial Naive Bayes Classifier

Multinomial Naive Bayes Classifier, a probabilistic learning method is the most suitable classifier for training purpose with discrete features. This model is used for classification of data in which features are based on word counts in text.

$$p(x/C_{\mathrm{k}}) = \frac{(\sum_i x_{\mathrm{i}})!}{\prod_i x_{\mathrm{i}}!} \prod_i {p_{\mathrm{k_i}}}^{x_{\mathrm{i}}}$$

$\mathbf{k}$ = number of output classes , $\mathbf{C_k}$ = Classes or K Possible Outcomes
$\mathbf{i}$ = event , $\mathbf{x_i}$ = number of times event i occurs
$\mathbf{p_i}$ = probability of occurring of event i

7.1.2 Confusion Matrix

This matrix summarizes the results produced by the classifier. For each classification class, it depicts the number of correct and incorrect predictions made. This depicts the accuracy of the model and how much confused it is while making predictions.

7.1.3 Recall

This is the number of true positives divided by the sum of number of true positives and false negatives. It depicts the recognition rate of users belonging to a particular class. A higher value for this ratio is desired, which indicates that classification classes are correctly recognized.

$$\textbf{Recall} = \frac{TruePositives}{TruePositives + FalseNegatives}$$

7.1.4 Precision

This is the number of true positives divided by the sum of true positives and false positives. It defines the level of correctness of a predicted label.

$$\textbf{Precision} = \frac{TruePositives}{TruePositives + FalsePositives}$$

7.1.5 F1 score

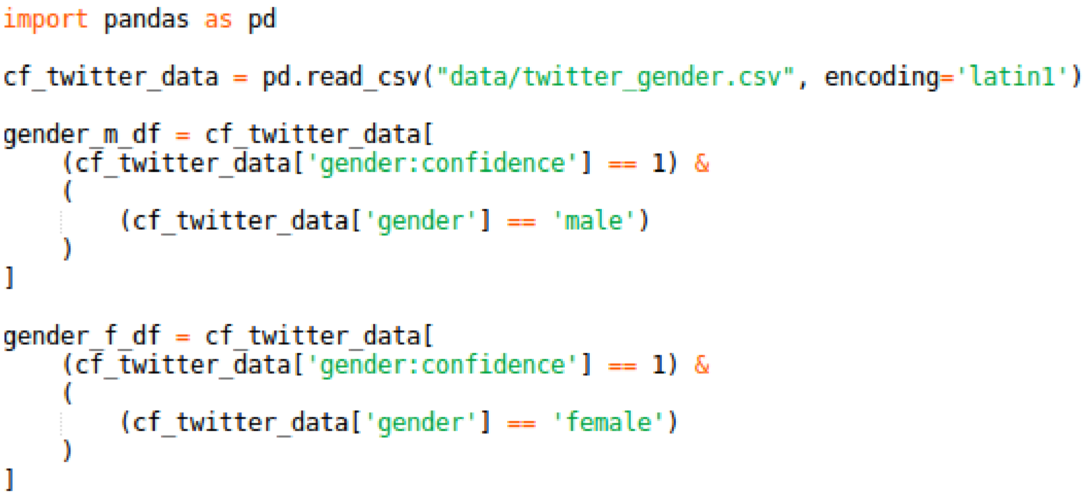
This is defined as the Harmonic Mean of Precision and Recall, hence it takes both the metrics into account.

$$\textbf{F1 Score} = 2 * \frac{Precision * Recall}{Precision + Recall}$$

7.2 Implementation

7.2.1 Gender

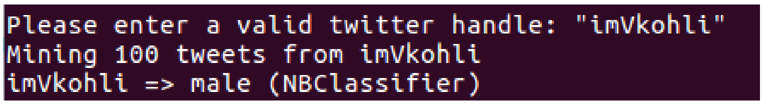
We extracted user name along with their label tag of gender in a .csv file using pandas. We collected only those user's which have gender confidence = 1. The following code snippet performs the above function. Then we take Consumer Token, Consumer Secret, Access Token Key and Access Token Secret to mine tweets using Tweepy. After extraction of tweets, frequency of n-grams are calculated . Then, we use tf-idf transformer to get the matrix. Feature Set may include n-gram like "go



```
import pandas as pd

cf_twitter_data = pd.read_csv("data/twitter_gender.csv", encoding='latin1')

gender_m_df = cf_twitter_data[
    (cf_twitter_data['gender:confidence'] == 1) &
    (
        (cf_twitter_data['gender'] == 'male')
    )
]

gender_f_df = cf_twitter_data[
    (cf_twitter_data['gender:confidence'] == 1) &
    (
        (cf_twitter_data['gender'] == 'female')
    )
]
```

Figure 5: Code Snippet

to play" , "cricket" with label as "male" along with the frequency, which is used for training purpose.
In Figure 5, twitter handle of Virat Kohli (@iamVkohli) is given as input, gender predicted is "male".

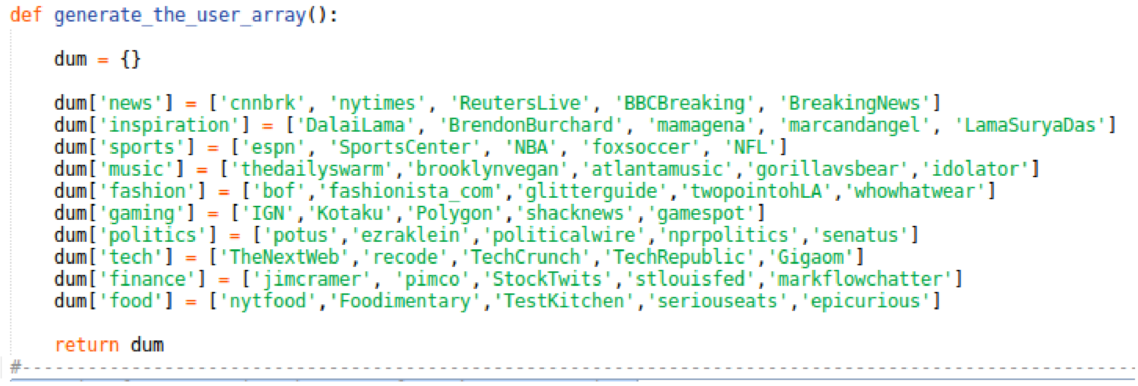


```
Please enter a valid twitter handle: "imVkohli"
Mining 100 tweets from imVkohli
imVkohli => male (NBClassifier)
```

Figure 6: Example of Gender Prediction

7.2.2 Interest

We required a labeled data-set which was not easily available. Therefore, we made some specific broad classes which depicted interests. We used 10 different classes (interests) in our project. Few of them are music, sports, politics, inspiration etc. Every class consist of some famous twitter handle of every category (interest). Since the number of user were less, so we increased the number of tweets to 4000-5000 per handle in order to increases the size of training data-set. The Figure 6 below shows the dictionary comprising of all interest classes.

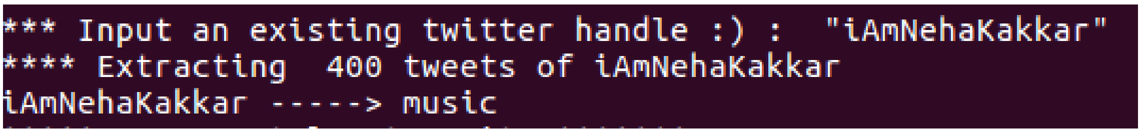


```
def generate_the_user_array():

    dum = {}

    dum['news'] = ['cnnbrk', 'nytimes', 'ReutersLive', 'BBCBreaking', 'BreakingNews']
    dum['inspiration'] = ['DalaiLama', 'BrendonBurchard', 'mamagena', 'marcandangel', 'LamaSuryaDas']
    dum['sports'] = ['espn', 'SportsCenter', 'NBA', 'foxsoccer', 'NFL']
    dum['music'] = ['thedailyswarm','brooklynvegan','atlantamusic','gorillavsbear','idolator']
    dum['fashion'] = ['bof','fashionista_com','glitterguide','twopointohLA','whowhatwear']
    dum['gaming'] = ['IGN','Kotaku','Polygon','shacknews','gamespot']
    dum['politics'] = ['potus','ezraklein','politicalwire','nprpolitics','senatus']
    dum['tech'] = ['TheNextWeb','recode','TechCrunch','TechRepublic','Gigaom']
    dum['finance'] = ['jimcramer', 'pimco','StockTwits','stlouisfed','markflowchatter']
    dum['food'] = ['nytfood','Foodimentary','TestKitchen','seriouseats','epicurious']

    return dum
#----------------------------------------------------------------------------------------------------
```

Figure 7: Matrix for different classes

Process similar to that of gender prediction is carried out and tf_idf matrix is formed. This time the labels are the names of the different classes (of interest). Then it was trained by Multinomial Naive Bayes classifier.

In Figure 7, twitter handle of Neha Kakkar (@iamNehaKakkar) is given as input, interest predicted is "music".

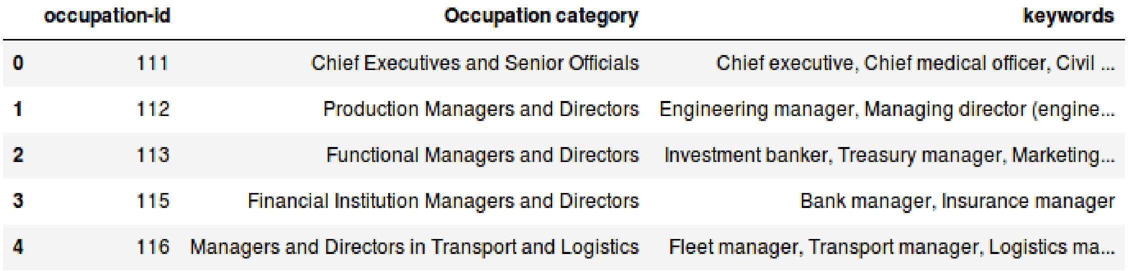


```
*** Input an existing twitter handle :) :  "iAmNehaKakkar"
**** Extracting  400 tweets of iAmNehaKakkar
iAmNehaKakkar -----> music
```

Figure 8: Example of Interest Prediction

7.2.3 Occupation

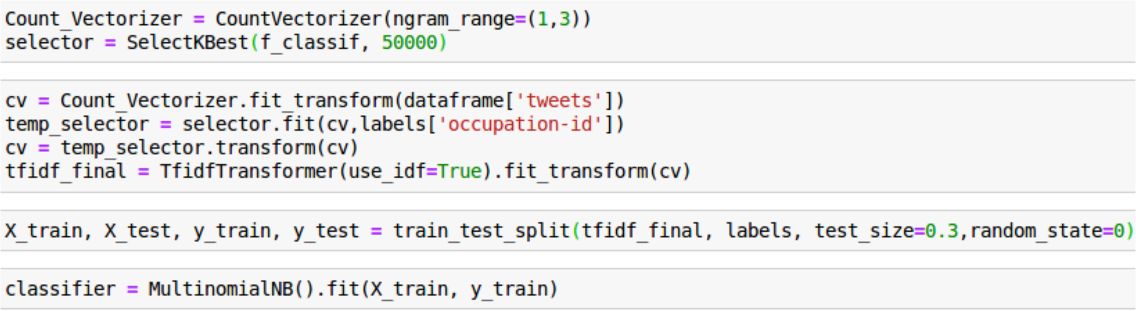
A similar approach was adopted for predicting the occupation class of the user. User handles, along with their cleaned tweets, clubbed with the occupational class code(label), was mined. We collected only those users' tweets which weren't deleted or user accounts deactivated. Also, tweets satisfying the constraint of having language as English were considered only. These are some of the occupational class divisions. After extraction of tweets, frequency of n-grams(upto n=3) are calculated



	occupation-id	Occupation category	keywords
0	111	Chief Executives and Senior Officials	Chief executive, Chief medical officer, Civil ...
1	112	Production Managers and Directors	Engineering manager, Managing director (engine...
2	113	Functional Managers and Directors	Investment banker, Treasury manager, Marketing...
3	115	Financial Institution Managers and Directors	Bank manager, Insurance manager
4	116	Managers and Directors in Transport and Logistics	Fleet manager, Transport manager, Logistics ma...

Figure 9: Standard Occupation Class (SOC)

. Then, we use tf-idf transformer to get the matrix.



```
Count_Vectorizer = CountVectorizer(ngram_range=(1,3))
selector = SelectKBest(f_classif, 50000)

cv = Count_Vectorizer.fit_transform(dataframe['tweets'])
temp_selector = selector.fit(cv,labels['occupation-id'])
cv = temp_selector.transform(cv)
tfidf_final = TfidfTransformer(use_idf=True).fit_transform(cv)

X_train, X_test, y_train, y_test = train_test_split(tfidf_final, labels, test_size=0.3,random_state=0)

classifier = MultinomialNB().fit(X_train, y_train)
```

Figure 10: Computing tf-idf

8 Results and Analytical Comparison

8.1 Gender

Confusion Matrix has been generated for a huge number of different twitter users. Also, the Precision and Recall values are calculated. Using these values, the F1-Score is computed. It is plotted against the true labels and the predicted labels of male and female gender. The matrix shows that the **male** users are predicted with a higher accuracy and **female** users are predicted with a lesser accuracy. The confusion matrix is shown in the Figure 8.

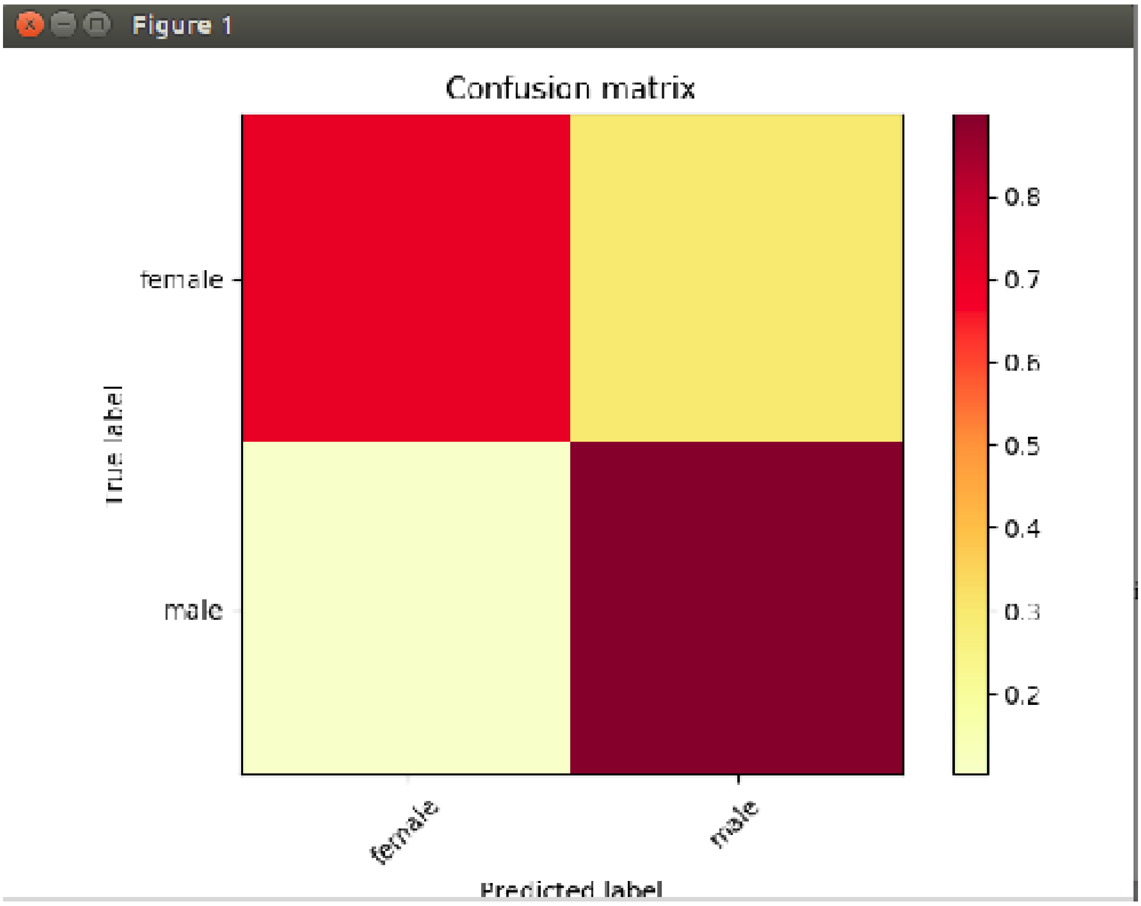


Figure 11: Confusion Matrix for Gender

The Precision , Recall and F1 - Score is calculated and is as shown in the Table 1.

	Precision	Recall	F1-Score
male	0.79	0.9	0.84
female	0.85	0.7	0.77

Table 1: Precision,Recall,F1-Score for Gender

8.2 Interest

Confusion matrix has been generated against a number of different users. Since there were 10 different classes used, so a **10 * 10 matrix** was generated which depicted the count of how many labels were correctly predicted.

Using this matrix, the True Positives, False Negatives, False Positives, True Negatives can also be analyzed for all the interest classes. Figure 9 shows the confusion matrix for interest.

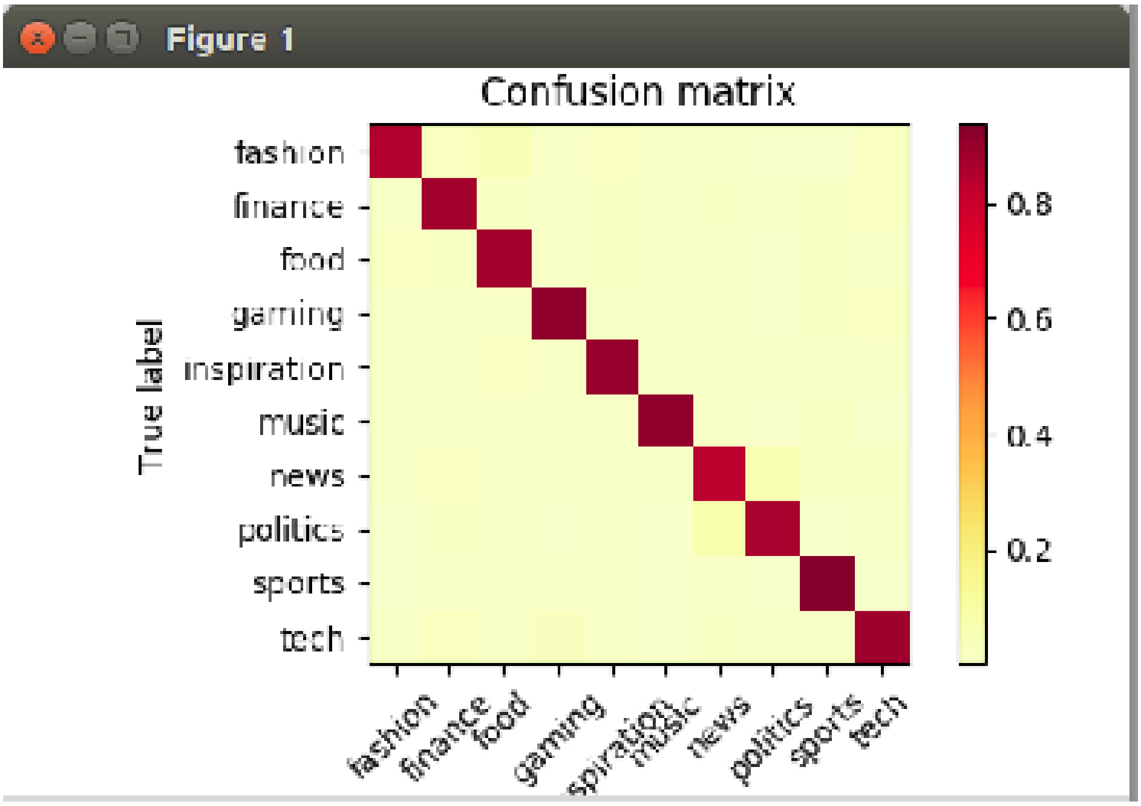


Figure 12: Confusion Matrix for Interest

The Precision , Recall and F1 - Score is calculated and is shown in the Table 2.

	Precision	Recall	F1-Score
fashion	0.90	0.85	0.87
finance	0.87	0.88	0.87
food	0.86	0.88	0.87
gaming	0.91	0.91	0.91
inspiration	0.88	0.90	0.89
music	0.93	0.91	0.92
news	0.85	0.84	0.84
politics	0.88	0.87	0.87
sport	0.89	0.94	0.91
tech	0.89	0.88	0.88

Table 2: Precision,Recall,F1-Score for Interest

8.3 Occupation

Confusion matrix has been generated against a number of different users vs various classes. Using this matrix, the True Positives, False Negatives, False Positives, True Negatives can also be analyzed for all the occupational classes. Figure 13. shows the confusion matrix for interest. Since there is a huge number of classes available

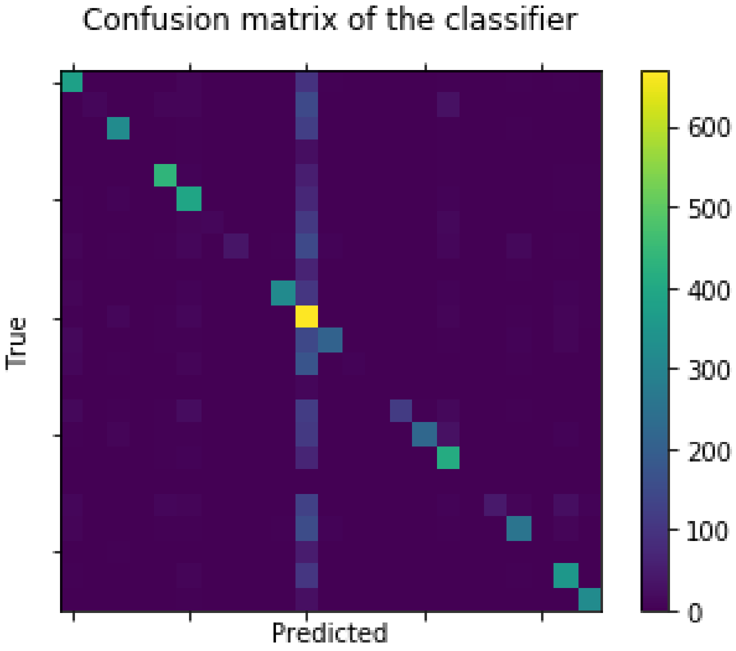
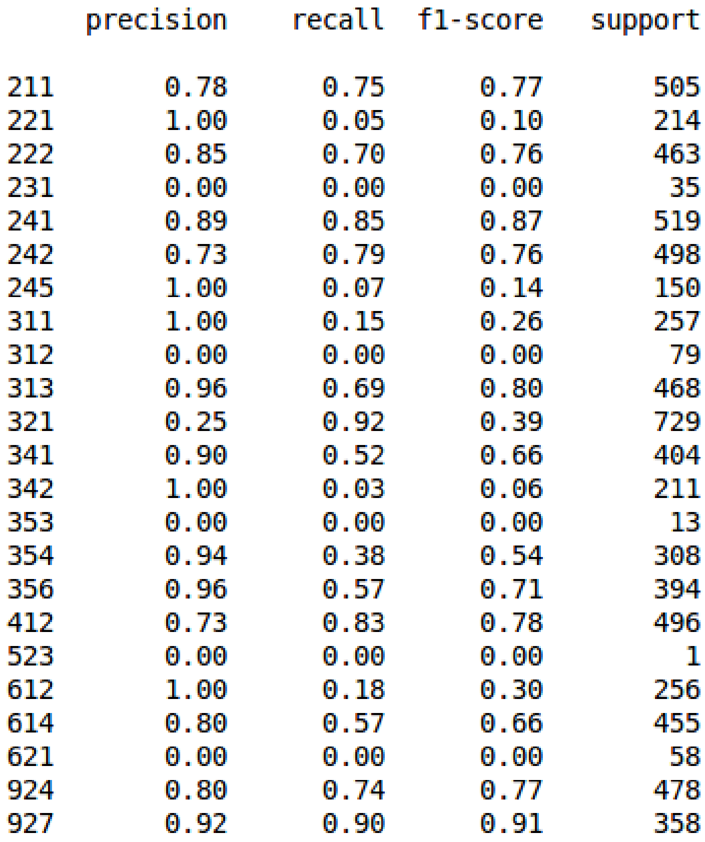


Figure 13: Confusion Matrix for Occupation

for occupation, even predicting a class nearly same as the true class will result in a wrong prediction. Hence, the accuracy is affected so.

The Precision , Recall and F1 - Score is calculated and is shown in the Figure 14.



```
          precision    recall  f1-score   support

 211           0.78      0.75      0.77       505
 221           1.00      0.05      0.10       214
 222           0.85      0.70      0.76       463
 231           0.00      0.00      0.00        35
 241           0.89      0.85      0.87       519
 242           0.73      0.79      0.76       498
 245           1.00      0.07      0.14       150
 311           1.00      0.15      0.26       257
 312           0.00      0.00      0.00        79
 313           0.96      0.69      0.80       468
 321           0.25      0.92      0.39       729
 341           0.90      0.52      0.66       404
 342           1.00      0.03      0.06       211
 353           0.00      0.00      0.00        13
 354           0.94      0.38      0.54       308
 356           0.96      0.57      0.71       394
 412           0.73      0.83      0.78       496
 523           0.00      0.00      0.00         1
 612           1.00      0.18      0.30       256
 614           0.80      0.57      0.66       455
 621           0.00      0.00      0.00        58
 924           0.80      0.74      0.77       478
 927           0.92      0.90      0.91       358
```

Figure 14: Confusion Matrix for Interest

9 Conclusion

We tried and tested our models on various data-sets and got varying accuracies. It was observed that accuracy of prediction of gender or interest or occupation increases as we increase the size of the data-set used to train the classifier. so, our technique can dynamically adapt to the changes in data popularity.

Confusion Matrices are plotted and various other metrics like precision, recall and f1-score are used for analysis purpose. For a Twitter user, Gender is predicted with an accuracy of 70%, Interest is predicted with an accuracy of 82%, and Occupation is predicted with an accuracy of 62%.

10 Future Work

To increase the accuracy of our models, we can predict one attribute by feeding the other two predicted attributes into the classifier. For eg: Accuracy of Gender Prediction may be increased by using the predicted values of interest and occupation of the Twitter Users along with their Tweets for training classifier.

Also supervised learning models like support vector machines or statistical models like logistic regression can be used to training purposes.

11 Progress Report on Board's Suggestion

Q: Please try more advanced conditional probability for Naive Bayes.

A: Multinomial Naive Bayes Classifier fits perfectly for our database. A detailed description in under **section 7** of this report.

Q: Data collection study should be elaborated.

A: A complete description of the data-set collection process is mentioned under **section 6** of this report.

Q: Comparative results will be explained in the end semester.

A: Complete analysis and comparative results are mentioned in **section 8** of this report.

References

[1] F. Lam, K. T. Chen and L. J. Chen . "Involuntary Information Leakage in Social Network Services," Third International Workshop on Security (IWSEC), Nov. 2008.

[2] A. N. Joinson and C. B. Paine . "Self-disclosure," privacy and the Internet, In the Oxford Handbook of Internet Psychology, pp. 237-252, 2007.

[3] K. R. Goldner . "Self Disclosure on Social Networking Websites and Relationship Quality in Late Adolescence," ETD Collection for Pace University, Jan. 2008.

[4] E. McCollister, T. Grance and K. A. Scarfone . "Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)," In: NIST SP - 800-122. pp 58, Apr. 2010.

[5] B. Krishnomurthy and C. E. Wills . "On the Leakage of Personally Identifiable Information via Online Social Networks" in Proc. of the 2nd ACM Workshop on Online Social Networks (WOSN), Aug. 2009.

[6] José Ahirton Batista Lopes FilhoJosé, Ahirton Batista Lopes FilhoRodrigo PastiRodrigo, PastiLeandro De CastroLeandro De Castro . Gender Classification of Twitter Data Based on Textual Meta-Attributes Extraction

[7] Nikolaos Aletras, Benjamin Paul Chamberlain . Predicting Twitter User Socioeconomic Attributes with Network and Language Information