Tweet Author Gender Identification

PAN 2016 Task

Serghei Nicvist, Daria Bogatireva
Informatics and System Engineering Department
Technical University of Moldova
Chișinău, Moldova
serejanxxx@gmail.com, daria.914@mail.ru

Victoria Bobicev
Informatics and System Engineering Department
Technical University of Moldova
Chișinău, Moldova
victoria.bobicev@ia.utm.md

***Abstract* — The paper presents an experiment of tweet's author gender detection. We used PAN 2016 data and task description and have build an application that decides whether an analysed tweet has been written by man or woman. Multiple texts' characteristics are used as features in the application, such as: references to pictures, to web pages, to other people, emojis, hashtags and a number of words that are associated with tweets written by women and men respectively. For 100 random tweets we obtained average accuracy 0.61. This is good result although it is not as good as the best one in PAN 2016 task.**

***Key words* — User Generated Content, text automate analysis, user profiling.**

I. Introduction

It is widely known that the growing volume of online user generated content (UGC) is a reach source for various sociological, psychological, political and marketing studies.

Due to its less strict privacy policy Twitter (https://twitter.com/twitter) is the most studied social net. The main reason is that its messages unlike other leading social nets as, for example, Facebook, are public by default. First launched on the Internet in March 2006, Twitter is a platform on which users can connect and communicate with short messages of just 140 characters. The number of characters was doubled in 2017. Everyone can see the posted tweets and use them in their research.

Author profiling is a problem of growing importance in applications in forensics, security, and marketing. E.g., from a forensic linguistics perspective one would like being able to know the linguistic profile of the author of a harassing text message (language used by a certain type of people) and identify certain characteristics (language as evidence). A series of online organised shared tasks, so called challenges are organised annually by PAN (https://pan.webis.de/tasks.html) for fostering of digital text forensics research. Anyone can participate in these tasks using the data prepared by the organisers and developing their own soft in order to solve the proposed tasks.

We selected one of the tasks, namely gender identification in English tweets organised in 2016. The data is available on the site (https://pan.webis.de/clef16/pan16-web/author-profiling.html) along with explanations and instructions. The page contains the results of the 2016 task which are relatively low: the best system obtained the overall Accuracy = 0.5258.

In this paper we present our attempt to solve one of the tasks of author profiling, namely gender identification of tweets' author basing solely on tweet's text.

II. Related Work

Analysis of User Generated Content (UGC) is being one of the hottest topics for the last decade. The texts generated by users are a reach source of various types of information. The obtained information is used in marketing application to target the appropriate audience. For example [2] used a hybrid text-based and community-based method for the demographic estimation of Twitter users, where these demographics were estimated by tracking the tweet history and clustering of followers/followees. The experiments were carried out on 10000 Twitter users and demonstrated good results even for users who only tweet infrequently.

Sociology is interested in so called „user profiling" which is the process of constructing a user's profile using his or her publicly and voluntarily shared social data [3]. A large amount of these data, including one's language, location and interest, is shared through social media and social network. The process includes detection user's gender, ages, origin, social group, interests, psychological features, etc. Altogether, this information can construct a person's social profile. In [3], a supervised machine learning approach was presented which categorizes Twitter users based on three important features: Tweet-based, User-based and Time-series based, into six interest categories - Politics, Entertainment, Entrepreneurship, Journalism, Science & Technology and Healthcare. The authors obtained up to 89.82% accuracy in classification experimenting with different traditional classifiers like Support Vector Machines, Naive-Bayes, k-Nearest Neighbours, Decision Tree and Logistic Regression.

[5] presented an exhaustive study of gender identification on the base of Russian texts. Four annotated datasets, including deception texts were used in the experiments. Seven types of features were investigated including the simplest ones as character n-grams, various word representations with their morphological characteristics, word-to-vec, and semantic representation using psycho-social dictionary. Machine learning models included Support Vector Machine, Decision

Trees, Gradient boosting and Neural Nets. The best obtained accuracy of 64% was reported with tf-idf on character n-gram features an Gradient Boosting model. The authors admitted that it is extremely difficult to obtain better accuracy and this would be possible only in case of considerable increase of the volume of training texts.

Twenty two teams participated in PAN 2016 author profiling task [6]. Pre-annotated tweets, blogs and social media texts in English, Spanish and Dutch were offered for training and evaluation. The participants run their soft on virtual machines provided by the organisers and the results were evaluated automatically. The best overall accuracies were around 0.52 although for gender discrimination task in some cases accuracy reached 0.75. For tweets, however, the best gender identification result was 0.73.

III. The Task and the Data

The paper presents the solution of the PAN 2016 task on author profiling, namely gender identifying. PAN has been (https://pan.webis.de/tasks.html) organizing a series of shared task evaluations for fostering of digital text forensics research. Shared tasks are computer science events that invite researchers and practitioners to work on a specific problem of interest, the task. The series started in 2011 with Authorship Attribution task formulated as follows: given a document and a set of candidate authors, determine which of them wrote the document (https://pan.webis.de/clef11/pan11-web/author-identification.html). Anyone may participate in the task. The organisers ask for the online registration and every registered participant receives the data set for the task. The participants were supposed to create a script which execute this task atomatically and run this script on the provided data. The results are evaluated by the organisers. Finally, all participants are presented in a table online along with their results sorted in descending order everyone to see the best performers. The evaluation is followed by a workshop where all participants are invited to present a paper describing their methodology. This year, the tasks are organised again and the evaluation will take place in April-May (https://pan.webis.de/clef18/pan18-web/index.html). The data, the description of the tasks, the obtained results and the workshop proceedings from the previous years (2011-2017) are presented on the PAN site so everyone may try to solve them again and to compare their results with the obtained previously.

We selected the PAN 2016 Author Profiling task. Author profiling is a problem of growing importance in applications in forensics, security, and marketing. E.g., from a forensic linguistics perspective one would like being able to know the linguistic profile of the author of a harassing text message (language used by a certain type of people) and identify certain characteristics (language as evidence). The focus is on author profiling in social media since the organizers are mainly interested in everyday language and how it reflects basic social and personality processes. The focus of 2016 shared task is on cross-genre age and gender identification. That is, the training documents will be on one genre (e.g. Twitter, blogs, social media...) and the evaluation will be on another genre (e.g. Twitter, blogs, social media...).

The task has been described as follows:

The organizers provide participants with a training data set that consists of Twitter tweets in English labelled with age and gender.

Due to Twitter's privacy policy the organizers cannot provide tweets directly, but only URLs referring to them. The participants have to download them using downloadable software provided by the organizers. The participants are expected to extract gender information only from the textual part of a tweet and to discard any other meta-information that may be provided by Twitter's API.

Downloaded archive pan16-author-profiling-training-dataset-2016-04-25.zip contained data for three languages: English, Dutch and Spanish. We worked only with English dataset as there are more lexical sources for this language. The folder with training dataset contained 436 xml files named after tweets' author ID, for example: 00db29c2dc1d87c8f07b72d753f7f2c0.xml. The file contained xml tags coding this author tweets as for example:

<document id="363394000847249408" url="https://twitter.com/sparCKL/status/363394000847249408"/>

The files with this information have been used to download the tweets.

The last file in the folder called truth.txt contained the information about gender and ages of each tweets' author in the following form (ID:::GENDER:::Ages), for example:

3bb08fb1b8b3d0d35bd70e1753840d2c:::MALE:::35-49

Each ID is associated with one xml file and can be used to connect tweets and gender.

Thus, we selected to solve the task of gender detection of the author of tweets on the base of their texts. In this case we have a task of text classification in two classes: male and female. Each tweet has to be assigned to exactly one of these two classes.

The evaluation metrics suggested by the task organisers was *Accuracy* [4] calculated as follows:

$$Acc = (\Sigma \text{ True Positive} + \Sigma \text{ True Negative}) / \Sigma \text{ Total},$$

where

True Positive is the number of tweets assigned to the gender (for example, male) which were indeed written by the same gender author;

True Negative is the number of tweets not assigned to the gender which were indeed written by the opposite gender author;

Total is the total number of tweets used for evaluation.

In the other words, Accuracy is the percent of correctly identified tweets among the total number of tweets used for evaluation.

IV. The Applied Method and the Used Features

Downloaded documents (dataset) were converted from .xml file format to JSON for better readability. An example of JSON coded document is presented below.


```
{
    "filename":"00db29c2dc1d87c8f07b72d753f7f2c0",
    "sex":"MALE",
    "age":"35-49",
    "tweets":[
        {"text":"@zulahni Oh, honey. Don't force us to stage an intervention. Look at your life. Look at your choices.
#sassygayfriend","url":"https:\/\/twitter.com\/sparCKL\/status\/363394000847249408"},
        {"text":"@BrianEnigma Re: ARG Tools 2.1: bad news, the HTML entity is actually \"&mdash\" (no \"e\"). #htmlisweird","url":"https:\/\/twitter.com\/sparCKL\/status\/363396349510684672"},
        ...
    ]
}
```


We used only tweet's text in our classification algorithm.

For the further work of the algorithm, we counted for each message (tweet) in each document:

- Total number of male words;
- Total number of female words;
- Total number of image links;
- Total number of links (URL);
- Total number of references to groups;
- Total number of references to people;
- Total number of hashtags;
- Total number of emoji.

All this data is used as features for classification.

After that, we counted:

- Total unique words for both genders
- Total common words for both genders
- Words frequency for both genders

Based on this data, a table of comparisons was created; a fragment is presented in Table 1.

Preference means that target feature or word is mostly used by this gender. Once we have calculated a table of comparison based on our dataset, we can start to predict the gender of any other message (tweet) from our data.

Gender prediction consists of the following steps:

- Counting main features (total number of links, hashtags, words, etc.)
- Finding words, which could represent gender.
- Final prediction of author's gender on the base of the features observed in the previous steps. An example of calculations is presented below.

```
[text] => @zulahni Oh, honey. Don't force us to stage an intervention. Look at your life. Look at your choices. #sassygayfriend

FEATURES:
    [totalPics] => unknown
    [totalLinks] => unknown
    [totalPeopleReferences] => male
    [totalGroupReferences] => unknown
    [totalHashtags] => female
    [totalEmoji] => unknown
    [honey] => female
    [force] => male
    [us] => male
    [to] => male
    [stage] => female
    [an] => male
    [intervention] => male
    [at] => male
    [your] => female
    [life] => female
    [choices] => female

PREDICTION:
  [gender] => male
  [predictedGender] => male
```

TABEL I. Comparison between Male nad Female Couters

feature	*male count*	*female count*	*Preference*
totalPics	5631	3055	male
totalLinks	75258	68645	male
totalPeopleReferences	90265	68552	male
totalGroupReferences	800,	1186	female
totalHashtags	53531	59770	female
totalEmoji	4924	5803	female
word "oh"	979	821	male
word "honey"	42	48	female
word "don't"	2389	1624	male
word "force"	141	86	male
word "us"	2288	1965	male
etc.	...	...	...

Text has no image links, links, group references and emoji, thus they are marked as unknown and these features are not used in final calculations.

Based on our knowledge from dataset, we can say, that people references are mostly used by males (see main table of comparison).

We count all found features for male and female genders. If there are more male features then the predicted gender is male, otherwise female is predicted. If there is the same number of features for each gender, the predicted gender is the most frequent one in our dataset.

V. Results and Discussion

In order to test our algorithm we selected random tweets from the corpus and run gender detection part over these tweets. Twenty tweets were selected and the algorithm detected gender of their authors and the accuracy was calculated. This evaluation process was repeated five times. Thus, we obtained the following results for the five iterations of the evaluation:

set 1: 0.6%
set 2: 0.5%
set 3: 0.75%
set 4: 0.6%
set 5: 0.6%
Average: 0.61%

Thus, for 100 random tweets we obtained average accuracy 0.61. This is good result although it is not as good as the best one in PAN 2016 task.

From the other hand the baseline is 0.5 which means that if we randomly assign the class (male or female) to a tweet we will reach 50% accuracy.

The low accuracy can be explained by two main reasons. First, this is really difficult task to detect author gender on the base of only one tweet. One tweet is only 140 characters long and can be rather neutral. For example, in the tweet: "Y'all remember when we had to copy and paste to quote RT." is almost impossible to tell the author gender. Instead, in this example: "Spent a nice day with my son and hubby!" or this: "Communicate with girls who have a questionable degree of motivation - my hobby" we can do this much easier.

VI. Conclusion and Future Work

The paper presents an experiment of tweet's author gender detection. We used PAN 2016 data and task description and have build an application that decides whether an analysed tweet has been written by man or woman. Multiple text's characteristics are used as features in the application, such as: references to pictures, to web pages, to other people, emojis, hashtags and a number of words that are associated with tweets written by women and men respectively. The application simply counts the found features and decides by voting who is most probable author: man or woman. Obviously, this is a quite simple algorithm and it may be upgraded to achieve better results.

First of all, we detect author's gender on the base on only one tweet which in many cases is almost impossible. Otherwise, we may collect a relatively sufficient number of tweets by one author to detect with high confidence the author's gender. The organisers provided tweet's authors information; tweets were even grouped by authors. Exactly 1000 tweets were collected for each author and this is a corpus sufficiently large to be able to detect the gender.

The other way to obtain better results is to change the methodology of work with the features. We may use various statistical methods or machine learning techniques such as models based on Bayes' formula, Decision lists, Support Vector Models, Lazy Learning (k-NN models) and many others to improve the obtained results.

The third source of improvement is the selected feature set. We use relatively small number of features while we could make a list of all words, hashtags, emojis and other tweet elements and analyse which of them are used mostly by men or women. There are also various methods of statistical dependency calculation which we could apply to our features and select the most influential for our two classes.

All these improvement we plan as our future work.

References

[1] M.A. Raghuram, K. Akshay and K. ChandrasekaranEfficient User Profiling in Twitter Social Network Using Traditional Classifiers. Intelligent Systems Technologies and Applications pp. 399-411, 2015.

[2] Kazushi Ikeda, Gen Hattori, Chihiro Ono, Hideki Asoh, Teruo Higashino. "Twitter user profiling based on text and community mining for market analysis." Elevier, Volume 51, October 2013, pp. 35-47.

[3] Kanojea, Sumitkumar; Mukhopadhyaya, Debajyoti; Girase, Sheetal. "User Profiling for University Recommender System using Automatic Information Retrieval". Procedia Computer Science. 2016, 78: 5–12.

[4] Powers, David M. W. "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation". Journal of Machine Learning Technologies. 2 (1): 37–63, 2011.

[5] Alexander Sboev, Ivan Moloshnikov, Dmitry Gudovskikh, Anton Selivanov, Roman Rybka, Tatiana Litvinova. Automatic gender identification of author of Russian text by machine learning and neural net algorithms in case of gender deception. Procedia Computer Science, Volume 123, 2018, Pages 417-423.

[6] Francisco Manuel Rangel Pardo, Paolo Rosso, Ben Verhoeven, Walter Daelemans, Martin Potthast, and Benno Stein. Overview of the 4th Author Profiling Task at PAN 2016: Cross-Genre Evaluations. CLEF 2016 Evaluation Labs and Workshop Working Notes Papers, CEUR Workshop Proceedings Volume 1609, Evora, Portugal, 750-784, 2016.