*Intelligent Equipment and Special Robots*
*Q. Zhang (Ed.)*

*doi:10.3233/ATDE240259*

# Bot and Gender Profiling Task of Twitter Using RoBERTa Pre-Training Model

Yutong SUN[a1] and Hui NING[b]
[a]*Heilongjiang Institute of Engineering, Harbin, Heilongjiang Province, China*
[b]*Harbin Engineering University, Harbin, Heilongjiang Province, China*
ORCiD ID: Yutong Sun https://orcid.org/0000-0002-5105-1006;
Hui Ning https://orcid.org/0000-0001-5266-6558

**Abstract.** This paper mainly describes the Bots and Gender Profiling task. Specifically, given a Twitter information, determine whether its author is a social bot or a human. In case of human, identify her/his gender. On the whole, this paper regards this as a classification task, using the RoBERTa pre-training language model to extract the emotional semantic features of the tweet, and identify the duplication rate of its content, and combine a variety of statistical features (including words+character n-grams, emoji and the forwarding rate of tweets), jointly recognize the human and bot categories.

**Keywords.** RoBERTa Model; Bots and Gender Profiling; Emotional Information; Multi-language Data set

## 1. Introduction

The correlation of social media in our daily life is getting higher and higher. While we get real-time news through social media, we have also accompanied some social users to maliciously manipulate online dialogue and release false news. The purpose of establishing social robots is to turn the original clumsy online processes into automation, rather than relying on manual completion. The main purpose of now is to be used for commercial propaganda. Products or advertising information are published regularly every day to attract corresponding customer groups.

After investigation, social robots are mainly targeted at human influence, but they will be divided into two parts according to their specific goals. Some are called normal social media robots to replace artificially and help users solve problems quickly and efficiently. The other part is called deceptive social media robots. They can pretend to be humans, affecting users' thinking methods for business, political or ideological purposes. For example, robots can exaggerate the popularity of products by publishing positive evaluations, and can also destroy the reputation of competing products through negative evaluations. In terms of expansion, the harmfulness of the political field will be greater, and the German politics party strictly refuses to use social robots for publicity in the presidential campaign.

[1] Corresponding Author: Yutong SUN, Heilongjiang Institute of Engineering, e-mail: 2961272047@qq.com

Therefore, the potential dangerous behavior promoted by automatic users (also known as 'deceived social robots' in online social networks (also known as robots) is of great research significance [1].

Specifically, the study of this article comes from the Bots and Gender Profiling tasks by the PAN@CLEF conference [2]. This research focuses on whether a Twitter information is given todetermine whether the author who published the information is a social bot or a human. In case of human, identify her/his gender. By monitoring a large number of robot social media accounts in advance, it has certain practical significance to curb the spread of deceit information.

In Chapter 2, we will analyze the relevant work of the Bot and Gender tasks. In the third chapter, we will introduce the methods proposed in this paper in detail, including text pre-processing, feature calculation methods, and classification prediction models. In Chapter 4, we mainly described the baseline methods in the experiment and the comparative analysis of the experimental results. In the end, Chapter 5 mainly summarizes this study and explains the direction of future research work.

## 2. Related Work

In previous studies, many researchers have explored the impact of different methods on the performance of social robot identification tasks. In fact, the Bots and Gender Profiling task can be attributed to the scope of the author profiling. Therefore, the research of this project focuses on locking and monitoring the abnormal target user group through the perspective of author analysis, and minimizes the impact of the spread of harmful information in social media from the source.

Most researchers have chosen N-Gram features and statistical characteristics based on text-based style for classification prediction. Ashraf et al.[3] use the total of 27 Language Independent stylometry features (contains 18 characters statistical features and 9 Twitter emoticons). Based on the Linear SVC, the profiling task of bot authors are jointly carried out.

Bacciu et al.[4] uses an integrated model to solve the problem of BOT detection. In bot profiling, selecting the statistical characteristics of text -based on the SVM model; in terms of gender prediction of human authors, a simple NB model is used to predict.

In addition, some researchers have proposed some more novel feature extraction methods. For example, [5] proposes to use the honeypots system to identify the online spam sender. To this end, they deployed honeypots as a false website and served as a trap for spam senders. The experiment found that the collected spam data features, such as text content, friends information or publishing mode. They use these observed features to provide information for machine learning classifiers in order to identify those with high availability spam senders. The accuracy is high and the misunderstanding rate is low.

Dickerson et al. [6] proposed a framework for collecting, pre -processing, annotations, and analysis of robots in Twitter. They have extracted a few features such as like, forwarding, user response and mentioning, URL, or Follower Friend. They found that human users can create more novel content, while robot users rely more on text information that rely on forwarding or URL sharing. Botometer is an online tool for robot detection. It can extract about 1,200 features for the given Twitter account to describe the personal information, friends, social networks, and emotional information of the account.

Compared with machine learning classifiers, some researchers are turning to deep learning algorithms. For example, using LSTM analysis at tense text data collected from Twitter, and reports that F1 scores F1 on the data set of 0.8732.

## 3. Model

### *3.1 Data Set*

This paper uses the data set released by the Bots and Gender Profiling task at the Pan@CLEF Conference in 2019 [1]. The XML file of each author (Twitter user) contains 100 tweets published. Each XML file corresponds to the only true authors ID. In the truth text, the list of authors and the group, the first column corresponds to the author ID, and the second column contains classification and gender labels.

### *3.2 Experimental Steps*

The model structure diagram of this experiment are detailed in Figure 1.

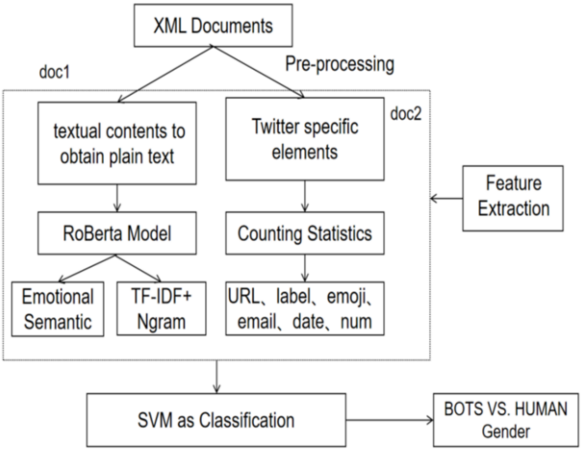


**Figure 1.** The model structure diagram of this experiment

**Text Pre-processing**: Pre-processing operation of the data set, divide the text into two parts: Document1 and Document2. Among them, the DOC1 part is the plain text content (filter stopping words). The DOC2 contains the specific statistical elements in Twitter. For example: labels, emoji, reserved words, emails, date, or numbers.

**Feature Extraction (this article has been extracted in total three types of typical features):**

A. The RoBERTa model uses the RoBERTa-Base pre-training model to build methods, and uses the transformer mechanism in the RoBERTa pre-training language model to learn and generates the embedded information based on the word, words, and sentences in the data text.

B. Based on the Document1 file obtained after pre-processing, the word and character n-gram method based on the TFIDF weighted algorithm to learn the statistical characteristics of the word sequence.

C. The biggest difference between bots and humans on social media information is whether there are emotional fluctuations. Many factors related to emotions are the key to identifying bots. In addition, the content of the information sent by the robot may be more monotonous, the content repetition rate is high, and the conversion rate is low. Therefore, this paper considers calculating the number of occurrences of certain types of elements to measure the personalized style of the tweet.

**Classification Prediction**: Integrate the above mentioned emotional semantic characteristics with personalized statistical characteristics, and re-send the feature vector obtained after stitching to the SVM model for classification. It is predicted whether the user is a bot or a human.

## 4. Experimental Results

### *4.1 Evaluation*

The performance of the Bots and Gender Profiling task will be ranked according to accuracy. For each language, calculate the accuracy rate separately. In this paper, the accuracy rate includes two aspects: 1. calculate the accuracy of identifying robots and humans. 2. For humans, calculate the accuracy of identifying men and women. The definition of the evaluation index is the ratio of the correct number of the prediction to the total number of predictions.

### *4.2 Based Methods*

(1) Bacciu method

In the control group one, using multiple groups of SVM integrated models to test whether the English tweet users are robots, and only provide a single SVM model for users who write in Spanish to predict. For gender testing, the two languages use a single SVM architecture, but the pre-processing operation of the tweet data set is made in different ways.

(2) Espinosa method [7]

In the control group two, the data is first processed, which mainly includes data cleaning operations and use character extraction feature Big-Gram. And try to use several features and machine learning model SVM and Linear SVM to make predictions.

(3) kosmajac method [8]

In the control group three, the user's behavioral data and statistical diversity are used to classify and predict the user. For gender identification tasks, the feature selection mainly uses a set of text statistical information, as well as syntax information and original words.

### *4.3 Experimental Results*

In order to verify the prediction of the model proposed in this paper for the Bots and Gender Profiling tasks, we used the 5-fold cross validation method to segment the data set. The training date set accounts for 80% and the verification date set accounts for 20%. Table 1 shows the effects of different feature combinations on social robot recognition and human gender accuracy under the Roberta pre-training language model.

**Table 1.** The effects of different feature combinations on the Bots and Gender Profiling

| Method | Feature Extraction Methods | es | | en | |
|---|---|---|---|---|---|
| | | B/H | Gender | B/H | Gender |
| Method-1 | Semantic | 0.930 | -- | 0.894 | -- |
| Method-2 | Semantic+n-gram | 0.938 | -- | 0.913 | -- |
| Method-3 | Semantic+n-gram+Emotion | 0.944 | -- | 0.924 | -- |
| Method-4 | Semantic+n-gram+<br>Emotion+number of RT | 0.950 | 0.845 | 0.930 | 0.788 |

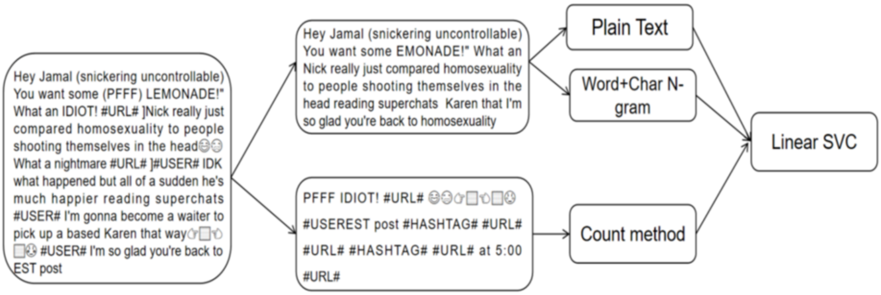
By analyzing the four groups in Table 1, different characteristic combinations have different degrees of improvement of experimental performance. The Method-4 method is more abundant than the other three methods, which contains the expression information of some behavioral data and emotional tendencies of Twitter users in social networks, which has improved in terms of experimental performance.

At the same time, this paper compares the evaluation results of the Bacciu method, Espinosa method, and Kosmajac method on the PAN@CLEF 2019 data set. Table 2 displays the experimental results of the method of the paper and traditional baseline methods.

By analyzing the experimental results of this method and the baseline method, the RoBERTa pre-training language model proposed in this paper and the feature extraction models such as n-gram can extract the core ideas of tweeting. The expression characteristics, the number of tweets, and emotional information contained in the tweet can also help us analyze whether the user has the common characteristics of social robots. Figure 2 depicts the specific experimental procedure of the Method-4.

**Table 2.** The experimental results of this method and traditional baseline methods

| Method | es | | en | |
|---|---|---|---|---|
| | B/H | Gender | B/H | Gender |
| Method-4(our) | 0.950 | 0.845 | 0.930 | 0.788 |
| Bacciu method | 0.943 | 0.841 | 0.908 | 0.776 |
| Espinosa method | 0.841 | 0.841 | 0.768 | 0.717 |
| Kosmajac method | 0.922 | 0.793 | 0.896 | 0.749 |



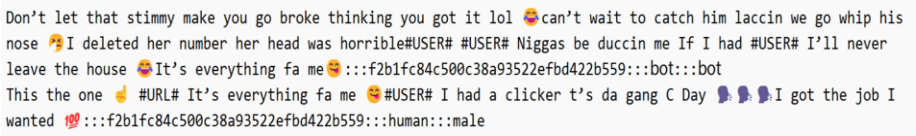
Hey Jamal (snickering uncontrollable) You want some (PFFF) LEMONADE!" What an IDIOT! #URL# ]Nick really just compared homosexuality to people shooting themselves in the head😀😀 What a nightmare #URL# ]#USER# IDK what happened but all of a sudden he's much happier reading superchats #USER# I'm gonna become a waiter to pick up a based Karen that way👉□👈 □😡 #USER# I'm so glad you're back to EST post

Hey Jamal (snickering uncontrollable) You want some EMONADE!" What an Nick really just compared homosexuality to people shooting themselves in the head reading superchats Karen that I'm so glad you're back to homosexuality

PFFF IDIOT! #URL# 😀😀👉□👈□😡 #USEREST post #HASHTAG# #URL# #URL# #HASHTAG# #URL# at 5:00 #URL#

**Figure 2.** The specific experimental procedure of the Method-4

This experiment implemented the core algorithm of the bot and gender profiling, so the final output is only a truth file, which displays the results of identifying the authors of real tweets, as shown in Figure 3. The result file is divided into four columns, using ':::' intervals. Among them, the first column is the tweet information downloaded from Twitter, the second column is the user ID, the third column is the identified user identity (bot or human), and if the result is a human tag, the fourth column will display gender attributes.



```
Don't let that stimmy make you go broke thinking you got it lol 😂can't wait to catch him laccin we go whip his
nose 🤦I deleted her number her head was horrible#USER# #USER# Niggas be duccin me If I had #USER# I'll never
leave the house 😂It's everything fa me😋:::f2b1fc84c500c38a93522efbd422b559:::bot:::bot
This the one ☝ #URL# It's everything fa me 😋#USER# I had a clicker t's da gang C Day 🗣🗣🗣I got the job I
wanted 💯:::f2b1fc84c500c38a93522efbd422b559:::human:::male
```

**Figure 3.** The bot detection results in the truth file

On the whole, the accuracy of the models based on the RoBERTa pre-training ideas and traditional statistical characteristics is improved compared to traditional baseline methods, which confirms the effectiveness of the method proposed in this paper.

## 5. Conclusion

This paper proposes social bot profiling methods based on RoBERTa pre -training language models and traditional statistical characteristics. This paper mainly starts from three aspects. (1) Social media corpus contains the problem of language diversity. Need to choose the appropriate cross-language analysis technology to extract the semantic information presented in the tweet. (2) In the experiment, compared with human beings, the behavioral differences in social networks enable the extraordinary features to distinguish social bot. (3) Through research, the biggest difference between bots and humans publishing social media information is whether there are emotional fluctuations. This paper extracts information related to emotional emoticons in tweets, and information such as tweets. From the experimental results, the deep learning model proposed in this paper is better than the traditional baseline methods.

For future work, we will continue to study how to capture and excavate the differences that can reflect the differences between bots and human thinking.

## Acknowledgments

This work is supported by the Youth Science Foundation of Heilongjiang Institute of Technology (No.2021QJ08).

## References

[1] Lee K, Caverlee J and Webb S, 2010. Uncovering social spammers: social honeypots+ machine learning. In: Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval.

[2] Rangel F, and Rosso P, 2019. Overview of the 7th Author Profiling Task at PAN 2019: Bots and Gender Profiling. In: CLEF 2019 Labs and Workshops, Notebook Papers.

[3] Ashraf S, Javed O, Adeel M, 2019. Bots and Gender Prediction Using Language Independent Stylometry-based Approach Notebook for PAN at CLEF 2019. In: CLEF 2019 Labs and Workshops, Notebook Papers.

[4] Bacciu A, Morgia L, Nemmi E. N, 2019. Bot and Gender Detection of Twitter Accounts Using Distortion and LSA Notebook for PAN at CLEF 2019. In: CLEF 2019 Labs and Workshops, Notebook Papers.

[5] Gilani Z, Farahbakhsh F, Tyson G, 2017. Of bots and humans (on twitter). In: Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.

[6] Dickerson G.P, Kagan K and Subrahmanian VS, 2014. Using sentiment to detect bots on twitter: Are humans more opinionated than bots? In :Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social of the 2014 IEEE/ACM International Conference on Advances in Social.

[7] Espinosa D.Y, Gómez-Adorno H and Sidorov G, 2019. Bots and Gender Profiling using Character Bigrams. In: CLEF 2019 Labs and Workshops, Notebook Papers.

[8] Kosmajac D and Keselj K, 2019. Twitter User Profiling: Bot and Gender Identification. In:CLEF 2019 Labs and Workshops, Notebook Papers.