A Turkish Dataset for Gender Identification of Twitter Users

Erhan Sezerer
İzmir Institute of Technology,
Computer Engineering,
İzmir, Turkey
erhansezerer@iyte.edu.tr

Ozan Polatbilek
İzmir Institute of Technology,
Computer Engineering,
İzmir, Turkey
ozanpolatbilek@iyte.edu.tr

Selma Tekir
İzmir Institute of Technology,
Computer Engineering,
İzmir, Turkey
selmatekir@iyte.edu.tr

Abstract

Author profiling is the identification of an author's gender, age, and language from his/her texts. With the increasing trend of using Twitter as a means to express thought, profiling the gender of an author from his/her tweets has become a challenge. Although several datasets in different languages have been released on this problem, there is still a need for multilingualism. In this work, we propose a dataset of tweets of Turkish Twitter users which are labeled with their gender information. The dataset has 3368 users in the training set and 1924 users in the test set where each user has 100 tweets. The dataset is publicly available[1].

1 Introduction

Author profiling is the characterization of an author through some key dimensions such as gender, age, and language. Among these profiling tasks, gender identification is different from authorship attribution problem in that it is a higher level abstraction, unlike authorship attribution where the candidate set of authors is unavailable a priori (Cheng et al., 2011). In gender identification from tweets, the difficulty lies in working with short text messages rather than using traditional text documents. Further, tweets are informal in their nature. Moreover, social media users have a tendency to hide their identity, to fake gender information. Thus, gender identification from the tweets of Twitter users is a challenging problem.

Author profiling is organized as a shared task in the PAN Workshop series as part of the CLEF conferences. The shared task releases a corpus and an evaluation framework to provide a lab environment to participants and measure their performances. In PAN 2013, the problem is stated as to identify age and gender from anonymous texts that are in English and Spanish (Pardo et al., 2013). A similar corpus construction effort takes place as part of the PAN 2017 task on gender and language variety identification in Twitter. In terms of methodological novelties; varying language use in tweets by the same user, retweet facility, possibility to retrieve tweets by region, validation through other types of data (photo, profile info, etc.) are considered specific to Twitter (Pardo et al., 2017). Also a dataset for Twitter user gender classification is released in Kaggle in 2015 [2].

There are several works focused on this problem. (Daneshvar and Inkpen, 2018) give Latent Semantic Analysis (LSA)-reduced forms of word and character n-grams into Support Vector Machine (SVM) and achieve state-of-the-art performance on PAN 2018 challenge (Pardo et al., 2018) for gender classification from text. Recently, neural network-based models have been proposed to solve this problem. In literature, CNN (Sezerer et al., 2018) or RNN (Takahashi et al., 2018), (Kodiyan et al., 2017) is used on this task. In the PAN 2018 challenge, using both textual and image data, (Takahashi et al., 2018) obtain state-of-the-art performance by proposing a model architecture where they process text through RNN with GRU cells.

Gender classification problem is addressed in Turkish language as well. (Talebi and Köse, 2013) use Naive Bayes, SVM, and K-nearest neighbour classifiers on a dataset composed of Facebook comments of Turkish users.

In this work, we contribute to the problem of author gender identification by sharing a corpus in Turkish for Twitter user gender classification. Although several datasets in different languages have been released on this problem, there is still a need for multilingualism.

[1] https://cloud.iyte.edu.tr/index.php/s/5DhqdlUCCdB60qG

[2] https://www.kaggle.com/crowdflower/twitter-user-gender-classification

Proceedings of the 13th Linguistic Annotation Workshop, pages 203–207
Florence, Italy, August 1, 2019.

In the remaining part of the paper, in Section 2, we explain the construction of the corpus in detail. Then, in Section 3, we present baseline performances on this dataset. Finally, we conclude the paper with some observations and insights regarding Twitter usage.

2 Dataset

We have compiled a corpus of Twitter for gender classification. Users are annotated as "male" or "female" and the corpus is publicly available.

2.1 Data Collection

In order to have a balanced collection with respect to each gender, we used common names from each gender as search filters (Pardo et al., 2017). In the determination of common names, we referred to websites that suggest names to male/female babies and a name database of Turkish Language Agency (Tr. Türk Dil Kurumu). After constructing the name database, we eliminated names that appear on the name list of both genders and also some names that are known as unisex. In the end, the size of the name database was 507 for female, 589 for male.

We used Twitter Web API[3] to search for names in Twitter. From the resulting set of user accounts that are retrieved from search queries, we selected the ones which have 200 tweets and 20 photos at minimum. The motivation behind this is that in order to identify gender, we need active users who have sufficient number of tweets on their own, and photos are taken to supply a different type of data to help annotators in their task. After retrieving those users, they are auto-labeled by their name's gender category.

Furthermore, in the selection of users we considered the presence of retweets. Since a retweet is not written by the original author, it may belong to a gender other than the user's gender. Thus, we selected those users that have at least 200 tweets of which 100 at minimum are not retweets. As a result, out of 12212 users that are collected from Twitter, only 8211 of them meet this criterion and are available to be labeled by annotators. Since we told annotators not to annotate if they are not sure, only 8071 of them are labeled.

[3] developer.twitter.com/en/docs/tweets/search/overview

2.2 Dataset Labeling

To guide the annotators, we have created a set of label categories (0-5) to control for correct/incorrect gender attribute, language of tweets, bot/human account, account belonging to a real individual, and account containing inappropriate content. Some label categories have subcategories to have more specific class labels inside each category for prospective Twitter classification tasks.

To guide the annotators, we have created several labels for users where each label corresponds to the type of rejection or acceptation. The labels are:

"0": If the automatically assigned gender is correct.

"1": If annotator thinks that the automatically assigned gender is wrong. Couples' account also fall into this category since both of them may contribute to the tweets.

"2": If the user mostly uses any language other than Turkish.

"3": If the user is a bot, or tweets are auto-generated texts. Here the definition of bot is extended to include "meaningless texts" (some computer viruses cause an account to generate meaningless texts in order to boost a certain hashtag).

"4a": If the user is a parody account or a sharing account like "funny cats", "funny joke each day" etc.

"4b": If the account is a fan page or an account that pretends to be a celebrity (Annotators are told to check whether the user is a real celebrity on the Internet).

"4c": If the user is a celebrity who doesn't tweet on his/her own (some celebrity or business people create a Twitter account and hire a PR (Public Relations) company to tweet on behalf of them).

"4d": If the user is not a human but a corporate identity (there are non-human accounts, such as company, political party, etc. on Twitter).

"5a": If the user is under 18 (An adult is defined as any person over 18 in Turkey, so if a clue like birthday or high/elementary school information is obtained about users being under-aged, user is discarded).

"5b": If the account has content involving nudity, sex, or prostitution (here nudity doesn't only rely on basic nudity but revealing body parts in favor of prostitution or finding partner).

If an LGBT+ person is found, the user is rejected with code 1 and commented as "neither".

The reason behind is that it's not possible to identify their gender or how they identify themselves by just looking at their tweets and profile pictures. Their status on the Twitter is used to detect whether they identify themselves as LGBT+ or not.

For this labeling task, we asked 22 people who are native speakers to help us. The annotators mostly consist of university students and academic personnel. To guide the annotation process, labels with their detailed descriptions are given to annotators and 400 users are assigned to each of them. The annotators are told to read all tweets of the user and they were able to check their status info and profile picture to be more sure about labeling. The annotators are also told not to label a user if they are not sure about their decision. They were given 6 weeks to finish labeling but to not let them feel pressure, that period is extended to 3 months. To control the consistency of annotations, each annotator is provided with randomly selected 20 users with ground-truth labels and a performance of 80% accuracy was expected on this set to accept his/her labels. The reason behind this threshold is that auto-labels turned out to be approximately 66% accurate on the ground-truth data and as (Nguyen et al., 2014) suggest humans can only achieve approximately 90% accuracy on this subject. So we expect from the annotators to surpass the auto-labels and perform close to 90% with a small margin of error to humans. Only one annotator failed to reach this accuracy, and his/her data are re-assigned to another annotator.

2.3 Post Processing

After the annotation phase, we received feedback from annotators that some accounts tweet some auto-generated texts, such as "az önce bir fotoğraf paylaştı" (eng. "Just shared a photo") or "Günlük istatistiğim, Takipçi: " (eng. "Daily statistics, followers:"). Using these feedbacks, we extracted the specified auto-generated texts and deleted those tweets including them from the dataset. After deletion, users who still have more than 100 tweets on their own are kept in the dataset. Lastly, in order to balance gender classes, some users are randomly discarded from females. Resulting ratio of females in the dataset is 0.53 and the total size of the dataset is 5292. We wanted to keep the test dataset size high (training/test dataset size ratio close to 2) thus we randomly partitioned the dataset as a training set of 3368 users and the rest as the test users which are 1924 in total. Additionally, to hide the true identity of the users, the user ids are hashed with the MD5 hash algorithm (Rivest, 1992).

Label	number of users	ratio
0	5803	0.718
1	427	0.052
2	111	0.013
3	153	0.018
4a	81	0.010
4b	389	0.048
4c	332	0.041
4d	615	0.076
5a	56	0.006
5b	104	0.012

Table 1: Distribution of labels in the dataset before partitioning

2.4 Findings on Behaviour of Turkish Twitter Users

As can be seen from Table **??**, we had to reject approximately 30% (1-5b) of the collected data due to non-human activities or other issues stated previously. This rate is quite higher than we expected and most of the rejections were because of non-real-human accounts (3-4d). This indicates that Twitter is getting more like a medium of advertisement. Moreover, this high rate can be attributed to Twitter's search algorithm. As a result of a search query, Twitter returns highly visible accounts that are related to it. Besides company accounts, since celebrities and people who act like a celebrity have more daily interaction than a regular user, they have a high ranking in the result set of queries.

On the other hand, we rejected more than a half of the total collected data due to insufficient number of tweets. Accounts that have less than 100 tweets of their own are discarded. Our experience in creating a dataset from Twitter shows that one needs to sample twice as much as s/he desires.

Additionally, the rate of bots is approximately 2% which shows that each sampling from the Twitter will have at least 2% noise if not eliminated by hand. This is observed among Turkish users only, it needs to be investigated in other languages.

Baseline Method	Accuracy
Random	0.5000
Bag-of-Words	0.7232

Table 2: Baseline Scores for Proposed Dataset

3 Baselines

To determine what to expect from the dataset, we created some baseline scores. Baselines are methods that define a lower bound for prediction performance. The performances of our baseline methods are given in Table **??**.

3.1 Random Baseline

Random Baseline is accepted as a reference point and its score is widely stated in each new dataset release. Random baseline score depends on the number of classes. Since there are two classes in this dataset, random assignment of classes will get approximately 50% accuracy.

3.2 Bag-of-Words

As a more advanced baseline, bag-of-words model is selected to obtain a more realistic lower bound. In the implementation of this baseline, we lowercased all words and tokenized them with NLTK (Loper and Bird, 2002) tool. Then, stop word removal and term frequency calculation are performed on the training dataset. In the frequency calculation; each mention, hashtag, and URL is labeled as <MENTION>, <HASHTAG>, and <URL> respectively. After getting frequencies, we selected the most frequent 1000 words as bag-of-words and represented all documents as a vector of 1000 frequent words. We used SVM (Cortes and Vapnik, 1995) with linear kernel as a classifier and got an accuracy score of 72.32%.

4 Conclusion

In this work, we propose a new dataset for gender classification from tweets of Twitter users. The language of tweets is Turkish and the dataset is annotated by native Turkish speakers. Random subsets of the annotations are cross-checked to validate the performance of each annotator. The dataset has 3368 users in the training set and 1924 users in the test set where each user has 100 tweets. Additionally, we run the traditional bag-of-words approach with a standard classifier and got 72.32% accuracy score as a baseline.

As a result of this dataset construction experience, we also share some insights and evidences about trends of Turkish Twitter users. We have seen that 17.5% of the users were non-real-human accounts, which shows that Twitter is more than a social media platform for some users. Also nearly 2% of the users were bots, which implies that for a random dataset selection from Twitter, there will be at least 2% noise coming from bot accounts.

Acknowledgments

We would like to thank Onur Keklik, Özge Sevgili Ergüven, Damla Yaşar, Berkay Can, Ceren Atik, İskender Ülgen Oğul, Arif Kürşat Karabayır, Ali Verep, Oğuzhan Öztürk, Mustafa Berkay Özkan, Ahmet Şemsettin Özdemirden, Elif Duran, Hasan Para, Emrah İnan, Cenk Tekir, Zuhal Tekir, Onur Tekir, Algın Poyraz Arslan, and Elgun Jabrayilzade for their great contribution on labeling the dataset.

We also would like to thank anonymous reviewers for their helpful comments and reviews.

The Titan V used for this research was donated by the NVIDIA Corporation.

References

Na Cheng, R. Chandramouli, and K.P. Subbalakshmi. 2011. Author gender identification from text. *Digital Investigation*, 8(1):78 – 88.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Mach. Learn.*, 20(3):273–297.

Saman Daneshvar and Diana Inkpen. 2018. Gender identification in twitter using n-grams and LSA: notebook for PAN at CLEF 2018. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018.*

Don Kodiyan, Florin Hardegger, Stephan Neuhaus, and Mark Cieliebak. 2017. Author profiling with bidirectional rnns using attention with grus. In *CLEF*.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, pages 63–70, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dong Nguyen, Dolf Trieschnigg, A. Seza Doğruöz, Rilana Gravel, Mariet Theune, Theo Meder, and Franciska De Jong. 2014. Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. In *Proceedings of COLING*

2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 1950–1961. Dublin City University and Association for Computational Linguistics.

Francisco M. Rangel Pardo, Paolo Rosso, Manuel Montes y Gómez, Martin Potthast, and Benno Stein. 2018. Overview of the 6th author profiling task at pan 2018: Multimodal gender identification in twitter. In *CLEF*.

Francisco M. Rangel Pardo, Paolo Rosso, Moshe Koppel, Efstathios Stamatatos, and Giacomo Inches. 2013. Overview of the author profiling task at PAN 2013. In *Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013.*

Francisco Manuel Rangel Pardo, Paolo Rosso, Martin Potthast, and Benno Stein. 2017. Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter. In *CLEF*.

R. Rivest. 1992. The md5 message-digest algorithm.

Erhan Sezerer, Ozan Polatbilek, Özge Sevgili, and Selma Tekir. 2018. Gender prediction from tweets with convolutional neural networks: Notebook for PAN at CLEF 2018. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018.*

Takumi Takahashi, Takuji Tahara, Koki Nagatani, Yasuhide Miura, Tomoki Taniguchi, and Tomoko Ohkuma. 2018. Text and image synergy with feature cross technique for gender identification: Notebook for PAN at CLEF 2018. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018.*

M. Talebi and C. Köse. 2013. Identifying gender, age and education level by analyzing comments on facebook. In *2013 21st Signal Processing and Communications Applications Conference (SIU)*, pages 1–4.