A Machine Learning Approach for Gender Identification of Greek Tweet Authors

Spiros Baxevanakis
sprbax@gmail.com
Department of Informatics, Ionian University
Corfu, Ionian Islands, Greece

Stelios Gavras
stelgavras@gmail.com
Department of Informatics, Ionian University
Corfu, Ionian Islands, Greece

Despoina Mouratidis
c12mour@ionio.gr
Department of Informatics, Ionian University
Corfu, Ionian Islands, Greece

Katia Lida Kermanidis
kerman@ionio.gr
Department of Informatics, Ionian University
Corfu, Ionian Islands, Greece

ABSTRACT

Digital communities and social media are widely used and produce a huge amount of information every second. Text analysis has been widely used by researchers and machine learning (ML) engineers for automating the author profiling task. Author profiling can be used in marketing and business intelligence frameworks but also remains a strong factor in crime investigations gaining more insight regarding the suspect. In this paper we describe the process we used in obtaining a new Twitter corpus and propose an ML approach to determine the gender of Greek author's tweets. The best result (0.7 accuracy) was obtained using SVMs and TF-IDF encoding.

CCS CONCEPTS

• **Information systems** → *Social networking sites*; Data analytics; • **Social and professional topics** → *Gender*; • **Computing methodologies** → **Natural language processing**.

KEYWORDS

author profiling, gender identification, behavioral analysis, machine learning, text classification, data mining, social media

ACM Reference Format:
Spiros Baxevanakis, Stelios Gavras, Despoina Mouratidis, and Katia Lida Kermanidis. 2020. A Machine Learning Approach for Gender Identification of Greek Tweet Authors. In *The 13th PErvasive Technologies Related to Assistive Environments Conference (PETRA '20), June 30-July 3, 2020, Corfu, Greece.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3389189.3397992

1 INTRODUCTION

Text classification is an important task with diverse areas of application such as spam detection systems, language identification, genre classification and sentiment analysis. The area of text classification that distinguishes between classes of authors is called author profiling. Author profiling is an interesting problem in a variety of areas, including forensics, security and marketing. For example, automatically identifying authors or analyzing characteristics of authors is also useful for marketing intelligence where specific information about current and potential customers is of high importance. This can help businesses to have suitable marketing strategy and develop products to meet customer demands. In this work we deal with gender identification which can be viewed as a two-class text classification problem.

The ability to be able to identify a person's gender yields great promise. For example, a polling firm that wants to extract the opinions of men and women from a random populous, would benefit vastly from an automated gender identification system because it would allow for a significant increase in sample size. It should be noted however that such a system would need to be very accurate in order to avoid higher error rates.

We summarize the related articles in two sections, various classification tasks and author profiling tasks. Two commonly used ML algorithms for text classification are Support Vector Machines (SVMs) [12] and Naive Bayes. The Naive Bayes classifier is a highly scalable algorithm, requiring a number of parameters linear in the number of variables. Additionaly, given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier.

There are numerous different learning algorithms that can be used for text classification with Artificial neural networks (ANN) and Convolutional Neural Network (CNN) showing very promising results in the field. Most of these techniques are based on words. In [24], the authors present a character-level CNN which seems to be an effective method for text classification with the assumption that that CNNs do not require knowledge regarding the syntactic or semantic structure of a language [23]. Another work worth mentioning is the character-level Recurrent Neural Network (RNN) of [9] who showed that a simple character-level RNN can be a useful model for text classification. Their work was a part of the author identification task at PAN 2015. The highest accuracy of the algorithm was 69% [9, 16, 23, 24].

Among the first studies that performed author profiling with machine learning was the work of [8]. In the aforementioned paper a combination of linguistic features has been used in order to discern the author's profile.

PETRA '20, June 30-July 3, 2020, Corfu, Greece

ACM ISBN 978-1-4503-7773-7/20/06…$15.00
https://doi.org/10.1145/3389189.3397992

The big scientific event, PAN [5], during the last years has arranged several author profiling tasks with various researchers considering aspects like gender, age, native language and personal interests. Support Vector Machines are the most popular models [10, 11, 13, 14, 17, 19], while some authors prefer logistic regression and random forests [7]. In [7] Amasyali et al, investigate various algorithms concluding that random forests (RF) [15] are the most reliable.

George K. Mikros et al. [21] presented a methodology for identifying the author's gender using the text of Greek twitter authors. The dataset was split by text length measured in number of words. This resulted into 10 subdatasets, out of which 3 feature groups were calculated. The author's achieved an accuracy of 88,3% using SVMs while investigating the effect of the text's length on the prediction accuracy of both Random Forests and SVM classifiers. The key thing to note is that, the dataset the authors used was initially developed in [18] for the purposes of authorship attribution and contains only 10 authors which were picked based on the number of followers and twitter activity.

Herein, we introduce a machine learning framework in order to predict the gender of Greek author's tweets. While author profiling using Natural Language Processing (NLP) is a very literature wealthy field, no similar work using Greek text has been addressed before. The key contribution of this research is that the data were obtained from a random twitter population.

The remainder of this paper is organized as follows. Section 2 describes the data acquisition process, the final dataset, the tools and our procedure (the feature set, the prepossessing process etc.). Section 3 describes experiment details and the results. Finally, Section 4 present the conclusion and directions for future research.

2 METHODS

The goal of the present work was to obtain a solid dataset containing tweets in the Greek language as well as corresponding author characteristics. To that end, the authors of this work used the twitter Application Programming Interface (API) [4] and the Python programming language [6]. The API allows anyone with a developer account to programmatically extract and analyze twitter data such as tweets, users characteristics, trends, topics and others. It is crucial to note however, that the API **does not** provide sensitive personal user information such as age, and, additionally, that twitter profiles do not have a field in which the user can report their gender. These facts mark proper data acquisition tasks of similar nature as nontrivial.

Data Acquisition. Firstly, keywords were selected at random from a pool of - at that time - trending ones. Then they were provided to the API which in turn extracted 500 tweets containing them. Each tweet was written by a different author. After the annotating procedure, which is described in the corresponding paragraph, 463 distinct authors remained from whom we extracted, using the API, a maximum of 100 tweets from each user's timeline. This process resulted in the final dataset which contains 45848 tweets/rows obtained from 463 users with 99.023 mean number of tweets per user. This methodology ensured the collected dataset is devoid of any noise regarding the target characteristic.

Target Characteristic. One of the most important decisions that authors had to take when initiating this research was the one regarding the target characteristic - in particular if it was going to be gender or age or the author. Empirically, it is well known that a significant percentage of ordinary twitter users do not provide their correct age due to various psychological factors that will not be analyzed in this study. In order to avoid a dataset of questionable validity, the gender is used as the target characteristic.

Data Annotation. In contrast to other similar studies where annotation is usually left to trusted professionals, here the annotation was done by the authors. This decision is supported by the fact that no professional knowledge or experience is needed in order to deduce a twitter user's gender from profile information. The authors worked as adversaries and took into consideration each profile's biography field and image. Profiles on which the authors did not reach to a consensus were disregarded from the dataset along with their tweets.

Final Dataset Description. The final dataset contains 45848 rows and 6 columns. Each row represents a particular tweet and has the following attributes: (author's twitter handle, author's name, twitter issued profile identification number, tweet link, tweet text, author's gender). It is important to consider that only the gender and text attributes were used as algorithm inputs in this study, all other features were collected on the premises of future research [1].

2.1 Data Preprocessing

An important task of any text analysis research is removing the stop words, a nontrivial task in the Greek language. This is mainly due to two reasons. First, almost all the Greek words include tonal signs which provide pronunciation information to the reader. Moreover, modern Greek was only standardized in the last 40 years. As a result, in some tweets authors used stopwords from other dialects of Greek such as Katharevousa, Koine, Ancient or they used words with polytonic signs instead of the standardized monotonic equivalents. As it becomes evident, the list of stop words is more complicated to extract than its English counterpart and the number of possible combinations make the task of removing Greek stop words, from randomly collected text found on social media, formidable. In order to tackle this issue the authors compiled a "super list" of stop words from many different sources, including, but not limited to, Natural Language Toolkit (NLTK) Greek stop words, GitHub projects and other sources we found online [1–3]. The resulting list used in this paper consists of 2175 words.

2.2 Feature Extraction

Once the data was preprocessed, the features used in the classification algorithms were extracted. In this paper we used two Bag-of-Words encoders which transform the data into a document-term matrix. Here the tweets are referred to as documents and the set of all distinct words in all documents make up the terms. Since machine learning algorithms cannot work with text directly - text needs to be converted to numbers. To that end we used the Count and TF-IDF encoders from [20]. These encoders provide a simple

[1] Both code and dataset can be provided from the authors upon request.

way to both tokenize a collection of text documents and build a vocabulary of known words.

The Count encoder transforms a collection of documents in a document-term matrix in which each element represents the number of times that the term appears in the document.

On the other hand, TF-IDF is a more shopisticated numerical statistic which is calculated for every document-term pair and reflects the importance of the term to the document with respect to a corpus. It provides word scores which try to highlight interesting words, e.g. frequent in a document but not across documents.

The size of the matrices that were created in both scenarios had dimensions equal to 45848x95853 meaning *number of documents* rows and *vocabulary size columns*.

2.3 Classification Parameters

In our experiments we used the scikit-learn [20] python library and the classifiers were configured with the default settings. Namely:

Multinomial Naive Bayes. alpha: 1, *fit prior*: True, *class prior*: None.

Support Vector Machine (SVM). Linear and RBF kernels were used, both with C = 1.

K-Nearest Neighbors. After running experiments with k ranging from 1 to 9 (with step 2) we found that k=1 yields the best accuracy. Thus that is what is presented in Table 1. All our experiments used the euclidean distance to determine the nearest neighbors.

Random Forests. We used 100 trees in the forest with the "gini" split criterion and no maximum depth limit.

3 DISCUSSION & RESULTS

In this paper we investigated a classical author profiling problem using Greek tweets. More specifically, we conducted data extraction and annotated the author's gender in more than 45000 tweets. Four different algorithms were tested in the experiments and the accuracy metric is used. Accuracy is the fraction of the correct predictions that the model achieves. Unequal distribution of the classification classes is observed (imbalanced dataset). In order to improve the classification accuracy, we address the class imbalance problem in the data by randomly removing male instances; we stop when both classes have the same number of observations. Table 1 shows the accuracy performance of different algorithms using the two Bag-of-Words encoders. It is worth mentioning that any kind of hyperparameter tuning in the aforementioned algorithms did not improve accuracy in any statistically significant way. By performing this classification we are able to achieve a very promising accuracy using only the tweet's text.

The Multinomial Naive Bayes classifier using the Count encoder achieved the best score for balanced data (69%) and imbalanced data (68%) among all the tested classifiers. KNN classifier and SVMs performed quite well, (68% for balanced / 66% for imbalanced using SVM linear, 67% for balanced and imbalanced using SVM and 67% for balanced / 64% for imbalanced using KNN). When the TF-IDF representation was used, the highest accuracy value 70 was obtained using SVM with linear kernel on imbalanced data. It is important to note that generally all classifiers achieved better results when the TF-IDF representation was used.

Table 1: Experiment Results by Accuracy

Encoding	Counts		TF-IDF	
Data	Balanced	Imbalanced	Balanced	Imbalanced
Naive Bayes	69%	68%	69%	68%
SVM (linear)	68%	66%	67%	70%
SVM (rbf)	67%	67%	68%	69%
KNN (k=1)	56%	57%	56%	44%
Decision Tree	67%	64%	64%	63%

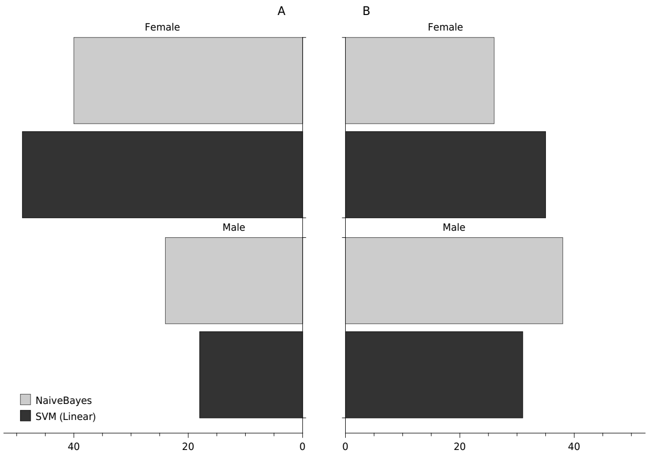

Figure 1: Incorrectly identified instances as percentages of the total instances. Imbalanced data on the left, balanced on the right.

Figure 1 presents the percentage of misclassified instances by the two best performing classifiers (NaiveBayes and SVM with linear kernel) in balanced and imbalanced data. The Naive Bayes classifier before balancing the data misclassifies 24% for Male class and 40% for Female class of the total incorrectly classified instances. For balanced data, a 14% increase in accuracy was observed (from 24% to 38%) for the Male class, while accuracy decreased 14% (from 40% to 26%) in the Female class. Furthermore, the SVM classifier achieved the highest accuracy on imbalanced data. It classifies incorrectly 18% male and 49% female instances. For balanced data, SVM misclassifies 31% male and 35% female instances. On imbalanced data both classifiers misclassied more female than male instances. Recently Sezerer et al [22] presented at 2018 PAN a convolution neural network producing competitive results on three languages with an average accuracy of 75.1% on local runs and 70.23% on the submission run.

4 CONCLUSION AND FUTURE WORK

In this paper an ML approach is presented for predicting the gender of Greek tweet authors. The model takes as input Greek text and predicts the gender of the author. The predictions are based only on the text and the assumption that it holds information about the gender. Enriching the input to the model is a future part of this work. This can be done by providing additional features like image data or specific statistics of tweeting behaviour. The combination of gender and age in the dataset is also a very promising framework but needs to be addressed very carefully because, empirically speaking, the age feature derived from twitter tends to be unreliable ground truth. Concluding, SVM produces more accurate results than the rest of the selected algorithms. However, such a complex classification requires more confident results and therefore Neural Networks shall be applied instead as part of a future work.

REFERENCES

[1] (accessed January 10, 2020). *Stopwords Github.* https://github.com/stopwords-iso/stopwords-el

[2] (accessed January 10, 2020). *Stopwords NLP Law project.* https://github.com/xtsimpouris/gr-nlp-law

[3] (accessed January 10, 2020). *Stopwords nltk.* https://www.nltk.org/index.html

[4] (accessed March 1, 2020). *Twitter API.* https://developer.twitter.com/

[5] (accessed March 25, 2020). *pan.* https://pan.webis.de/

[6] (accessed March 25, 2020). *Python Programming Language.* https://www.python.org/

[7] M Fatih Amasyalı and Banu Diri. 2006. Automatic Turkish text categorization in terms of author, genre and gender. In *International Conference on Application of Natural Language to Information Systems*. Springer, 221–226.

[8] Shlomo Argamon, MoSezerershe Koppel, James W Pennebaker, and Jonathan Schler. 2009. Automatically profiling the author of an anonymous text. *Commun. ACM* 52, 2 (2009), 119–123.

[9] Douglas Bagnall. 2015. Author identification using multi-headed recurrent neural networks. *arXiv preprint arXiv:1506.04891* (2015).

[10] Roy Bayot and Teresa Gonçalves. 2016. Author Profiling using SVMs and Word Embedding AveragesâĂŤNotebook for PAN at CLEF 2016. (2016).

[11] Konstantinos Bougiatiotis and Anastasia Krithara. 2016. Author Profiling using Complementary Second Order Attributes and Stylometric Features.. In *CLEF (Working Notes)*. 836–845.

[12] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20, 3 (1995), 273–297.

[13] Rodwan Bakkar Deyab, José Duarte, and Teresa Gonçalves. [n.d.]. Author Profiling Using Support Vector Machines Notebook for PAN at CLEF 2016. ([n. d.]).

[14] Daniel Dichiu and Irina Rancea. 2016. Using Machine Learning Algorithms for Author Profiling In Social Media.. In *CLEF (Working Notes)*. 858–863.

[15] Tin Kam Ho. 1995. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, Vol. 1. IEEE, 278–282.

[16] Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* (2014).

[17] Ilia Markov, Helena Gómez-Adorno, Grigori Sidorov, and Alexander F Gelbukh. 2016. Adapting Cross-Genre Author Profiling to Language and Corpus.. In *CLEF (Working Notes)*. 947–955.

[18] George K. Mikros and Kostas Perifanos. 2013. Authorship Attribution in Greek Tweets Using Author's Multilevel N-Gram Profiles. In *AAAI Spring Symposium: Analyzing Microtext*.

[19] Mart Busger op Vollenbroek, Talvany Carlotto, Tim Kreutz, Maria Medvedeva, Chris Pool, Johannes Bjerva, Hessel Haagsma, and Malvina Nissim. 2016. Gronup: Groningen user profiling. In *Working Notes of CLEF, CEUR Workshop Proceedings*. 846–857.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python . *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[21] George K Mikros-Kostas Perifanos. 2015. Gender identification in Modern Greek tweets. *Recent Contributions to Quantitative Linguistics* 70 (2015), 75.

[22] Erhan Sezerer, Ozan Polatbilek, ÃŰzge Sevgili, and Selma Tekir. 2018. Gender Prediction From Tweets With Convolutional Neural Networks.

[23] Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*. 649–657.

[24] Chunting Zhou, Chonglin Sun, Zhiyuan Liu, and Francis C. M. Lau. 2015. A C-LSTM Neural Network for Text Classification. *CoRR* abs/1511.08630 (2015). arXiv:1511.08630 http://arxiv.org/abs/1511.08630