Measuring Gender: A Machine Learning Approach to Social Media Demographics and Author Profiling

Erik-Robert Kovacs(✉), Liviu-Adrian Cotfas, and Camelia Delcea

Department of Economic Informatics and Cybernetics, Bucharest University of Economic Studies, 010552 Bucharest, Romania
{erik.kovacs,camelia.delcea}@csie.ase.ro, liviu.cotfas@ase.ro

Abstract. Social media has become a preeminent medium of communication during the early 21st century, facilitating dialogue between the political sphere, businesses, scientific experts, and everyday people. Researchers in the social sciences are focusing their attention on social media as a central site of social discourse, but such approaches are hampered by the lack of demographic data that could help them connect phenomena originating in social media spaces to their larger social context. Computational social science methods which use machine learning and deep learning natural language processing (NLP) tools for the task of author profiling (AP) can serve as an essential complement to such research. One of the major demographic categories of interest concerning social media is the gender distribution of users. We propose an ensemble of multiple machine learning classifiers able to distinguish whether a user is anonymous with an F1 score of 90.24%, then predict the gender of the user based on their name, obtaining an F1 score of 89.22%. We apply the classification pipeline to a set of approximately 44,000,000 posts related to COVID-19 extracted from the social media platform Twitter, comparing our results to a benchmark classifier trained on the PAN18 Author Profiling dataset, showing the validity of the proposed approach. An n-gram analysis on the text of the tweets to further compare the two methods has been performed.

Keywords: author profiling · gender identification · ensemble methods · social media analysis · COVID-19

1 Introduction

The present economic and social environment is characterized by a series of unexpected events having a major impact on the lives of people across the world [1, 2]. In this context, social media has become a meeting ground where people connect in real time, sharing ideas and information regarding the events that alter their daily lives [3, 4]. At the same time, social media has become an ideal data source that can be explored by both researchers and policy makers when trying to better understand the issues, fears and information needs of society. In the process of addressing these issues, knowing the

N. T. Nguyen et al. (Eds.): ICCCI 2023, LNAI 14162, pp. 337–349, 2023.
https://doi.org/10.1007/978-3-031-41456-5_26

audience is an important step for devising adequate policy [5]. Among the demographic characteristics of the users posting in social media, gender plays an important role since events can affect women and men differently [6, 7].

The specific problem we aim to solve with our contribution is the lack of information regarding the underlying demographics of text data sampled from social media sources. This lack of insight has widely been cited as an intrinsic issue with computational social science research [8–10]. An estimation of the actual underlying demographics would be valuable as it would allow an evaluation of how representative the sample is. Additionally it would allow a very granular approach to computational social science that could generate deeper insights into the many factors that correlate with certain opinions, such as the opposition towards vaccination [9, 11]. The development of a comprehensive demographic classification methodology, associated with the availability of appropriate training data, would push the field towards becoming a valuable complement to traditional methods such as surveys, with the advantage of being able to leverage a vastly superior number of data points.

As a result, the aim of our contribution is to suggest an improved method for estimating the gender distribution of a sample of online texts gathered from the microblogging platform Twitter using computational tools. We use publicly available datasets to train a series of classifiers for two sub-problems: the identification of a given name as opposed to a surname or other English word, and the identification of the gender of that given name. We obtain the best results using random forest (RF) classifiers for both sub-problems. We compare our approach to a baseline inspired by the PAN18 Author Profiling task [12] using the text component of the provided dataset. We validate our approach on domain data gathered by Banda et al. [13] during the COVID-19 pandemic, obtaining a gender distribution that matches the true estimated distribution of Twitter users [14]. Additionally, we extract and compare the top n-grams for each gender for the purpose of analyzing if there are meaningful differences among genders in the discourse related to the COVID-19 pandemic. We release the composite dataset used for the given name identification sub-problem for further research use.

The paper is structured as follows: Sect. 2 provides a brief literature review which supports the need for the current study, while Sect. 3 describes the data and methods used in the current approach. Section 4 analyzes the performance of the gender identification approach, with a focus on the results obtained on the selected COVID-19 dataset. The paper ends with concluding remarks and further research directions.

2 Related Work

Natural language processing (NLP) uses computational tools for operating on inputs in natural language. The field has evolved tremendously during the past few years, seeing the introduction of the Transformer neural network architecture [15] and the development of large transfer learning models such as BERT [16].

At the same time, there has been increased interest from fields associated with the social sciences in using these computational tools to analyze various aspects related to public opinion [1, 5, 9, 17]. The challenge with these approaches is that in most of the studies, demographics information is missing, the analysis being conducted on the entire dataset, without considering any differences that could exist in terms of gender, age, or ethnicity [18]. Thus, it is difficult to connect any findings back to the social context in which the social media discourses studied arose in the first place.

The NLP task that is concerned with extracting such information from text data is known as author profiling (AP) [19]. It aims to identify details about the user such as gender, age, native language, etc. [19–21]. An important subtask of AP is gender identification. This can be defined formally as the task of finding tuple $<a,g>$ given any sample of text x_i, where a is the author and g is the gender, $g \in \{female, male\}$. We have identified two main approaches to gender identification: intrinsic gender identification, when x_i is one document out of a corpus X of annotated documents $<x_i, g_i>$, and metadata-based gender identification, when the document x_i is a piece of information concerning the author, such as their name, occupation, place of employment, preferred pronouns, etc.

Approaches to gender identification can be grouped from a technical standpoint into dictionary-based [22], classical machine learning [19], and deep learning [21]. A comprehensive review that compares the results achieved on the PAN18 Author Profiling dataset[1] by the approaches described in 23 papers focusing on gender detection is included in [12]. Out of these, the best accuracy (82.21%) has been achieved by Daneshvar and Inkpen [23], where the authors used a Support Vector Machine classifier with a combination of different n-grams as features.

The gender identification method we propose is a metadata-based one, as by using the Twitter API it is possible to retrieve the public name field of any tweet's author. Inferring a person's gender from their name is possible because most European names are inherently gendered. It is important to note that this is not applicable to all languages and cultural contexts; for instance, not all Mandarin Chinese names can be assigned a gender [24]. Thus, great care must be taken to avoid using the name-based approach when dealing with non-European contexts and languages. In such cases, domain knowledge should be used to determine how well the approach fits local naming customs.

3 Data and Methods

3.1 Domain and Training Data

The domain data on which we aimed to validate our approach is the Large-Scale COVID-19 Twitter Chatter Dataset made available by Banda et al. [13]. This dataset contains 1.2 billion tweets related to COVID-19, collected between January 2020 and June 2021, presented in the form of a list of tweet IDs that can be used to retrieve each individual tweet from the Twitter API [13].

Due to issues of scale, we have further reduced the number of tweets by restricting the timeframe to the period between January 2021 and March 2021.We retrieved the name of the user who posted each tweet for gender identification. Our curated domain data contained 44,248,682 tweets from 6,999,706 distinct users. One of the difficulties with using the user's name to identify their gender is that on Twitter, the name field is free text; as such, it can also contain non-name related tokens, such as titles, job-related

[1] https://pan.webis.de/data.html.

information, political affiliation, preferred pronouns, etc. in addition to allowing the user to simply use a pseudonym. Nevertheless, after cleaning up the date and removing special characters, we have observed that many of the most common unigrams ranked by term frequency appear to be personal names (see Table 1.), lending credence to the fact that many users prefer to use actual human names instead of other signifiers. At the same time, the incidence of given names decreases when bigrams and trigrams are considered, with email addresses ("gmail com"), pandemic-related ("wear mask") and political messages ("black lives matter", "president elect") becoming more common.

Table 1. Top-10 n-grams ranked by term frequency found in the name field.

Type	N-grams
Unigrams	dr (41718), david (32122), john (31899), michael (26795), chris (24236), james (22820), kumar (21670), singh (20656), paul (20322), mark (18835)
Bigrams	gmail com (2895), wear mask (2079), stan account (2055), president elect (1962), lives matter (1845), black lives (1816), kinda ia (1816), yoongi day (1418), semi ia (1163), de la (1139)
Trigrams	black lives matter (1631), name cannot blank (577), wear damn mask (210), hu tao haver (192), sb ikalawang yugto (186), de la cruz (151), happy yoongi day (123), bts paved way (108), de la torre (102), certified lover boy (100)

The data appears to contain the name information, simply requiring special preprocessing. Pre-trained named-entity recognition (NER) models can be used to identify given name – surname tuples, but as the name order used on Twitter might be variable or interspersed with tokens such as "dr" or "mr", these might not generalize well to arbitrary data and might skew the resulting distribution in ways we cannot easily explain, account or compensate for. Thus, we propose a novel, machine learning approach to given name identification, in which we evaluate each token individually and classify it as a given name or not.

Because to the best of our knowledge this exact technique has not been applied in the literature, we know of no publicly available dataset relevant to this task. Nevertheless, the n-gram analysis from Table 1 suggests the presence of at least four categories of signifiers in the field: given names ("david", "john", "michael"), surnames ("singh", "de la *x*"), other English words ("dr", "name", "cannot"), and non-English words ("hu tao haver", "sb ikalawang yugto"). As such, a dataset containing these classes of tokens can be constructed from other publicly available datasets and annotated automatically.

For this purpose, we have merged three separate datasets: for given names, we used the Gender by Name Data Set available in the University of California Irvine Machine Learning Repository, containing 147,270 personal names and the associated biological sex of the persons bearing those names, dating from between 1880 and 2019 and gathered from the US, the UK, Canada, and Australia [25]. For surnames, we used the Wiktionary Names Appendix Scraped Surnames Dataset containing 45,136 surnames from persons across the world, gathered from Wiktionary[2]. Finally, for arbitrary English words, we used the 1/3 million most frequent words corpus [26][3], containing the top 333,331 words used in the English language ranked by term frequency. Because certain tokens were present in more than one dataset, we removed all duplicates. The merged dataset, consisting of 476,089 tokens, can be accessed at: https://github.com/erkovacs/measuring-gender-a-machine-learning-approach-social-media-demographics-author-profiling-data. For the gender identification problem, we have used the Gender by Name Data Set individually.

3.2 Preprocessing and Tokenization

We have substituted all special characters from the data with their English phonetic transcriptions using the software package unidecode[4]. In addition, we substituted all punctuation with the character "*" and lowercased all the tokens in the dataset. We applied this same preprocessing to tokens in all categories, and for both sub-problems.

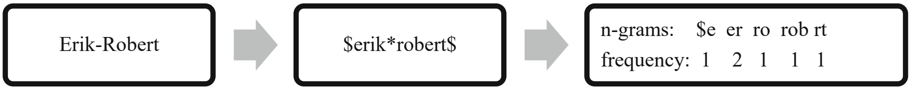

Fig. 1. Data representation steps with a toy feature set consisting of bigrams and trigrams.

For feature representation, we have used a character-level tokenization scheme based on the one proposed by Malmasi and Dras [27]. This tokenization scheme can capture sub-word structures that encode gender information at the level of names using a compact alphabet composed of unigram, bigram and trigram features with the addition of special tokens "*" mentioned above and "$", marking the beginning or end of a string [27]. After building this feature set and transforming the data, we used the most representative 2048 features to build a document-term matrix for each token, as shown in Fig. 1.

3.3 Ensemble Classifier

The final ensemble pipeline we propose consists of two classifiers and decision points (Fig. 2.). The first one is the given name identification step, which takes a list of tokens for each author and classifies each token as a given name or surname/other word. Binary classification is sufficient for our purposes in this case because we are only interested in whether the token is a given name or not. The tokens that are classified as given names are kept in the list; all other tokens are removed.

Users with zero tokens identified are labelled as "anonymous" and will not be evaluated for their gender during the next step. For the remaining users, each given name

[2] https://github.com/solvenium/names-dataset.
[3] https://norvig.com/ngrams/.
[4] https://pypi.org/project/unidecode/.

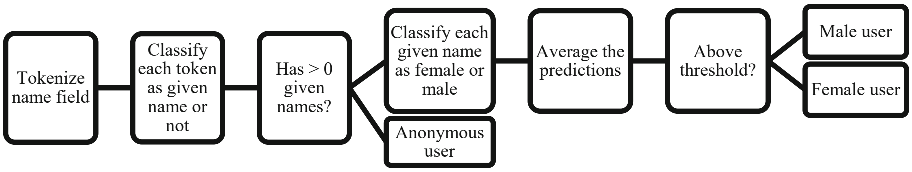

Fig. 2. Illustration of the ensemble method steps and decision points.

identified in the previous step is identified as *female* or *male* and the predictions averaged over the number of tokens. For simplicity the decision threshold was set at 0.5.

3.4 Benchmarking

In order to compare our metadata-based approach to an intrinsic gender identification approach, we have decided to use the English text component of the PAN18 Author Profiling dataset [12], containing 4,900 users annotated as *female* or *male* and 100 tweets from each user. We have chosen this dataset because of its proximity to our own application and because it has also been collected from Twitter. The original task envisioned using multimodal techniques for author profiling, with the data also including several pictures for each user, but we considered that this machine vision element does not fit the scope of our work and as such, we limited ourselves to using the text component only for a fair comparison.

We fine-tuned a BERT [16] classifier on the full English part of the dataset for 10 epochs, with a maximum sentence length of 16 tokens (the average sentence length being 20.45 tokens), and a learning rate of 55e–6, chosen by running several training cycles with different learning rates we have experimented with in the past [9]. To match our approach from Sect. 3.3, we merged the train and the test data made available by the organizers and performed 5-fold cross validation. The model obtained an F1 score of 79.40% and an accuracy of 79.38%, an above-average performance considering the results reported by Rangel et al. [12]. We then removed all duplicate tweets, being left with 12,432,935 unique tweets, and used the model trained on the PAN18 AP dataset to predict the gender of the user for each of these.

4 Results

4.1 Classifier Evaluation

We have trained multiple classifiers, using both classical machine learning and deep learning, for each of the two sub-problems: random forest (RF), support vector machine (SVM), multinomial naïve Bayes (NB), a feedforward neural network (FFN), a recurrent neural network (RNN), and a long short-term memory network (LSTM). We compared the classifiers using the F1 score (Eq. 3), computed as the harmonic mean between precision (Eq. 1) and recall (Eq. 2). All values given are mean values obtained over 5-fold cross-validation.

$$recall = \frac{TP}{TP + FN} \tag{1}$$

$$precision = \frac{TP}{TP + FP} \tag{2}$$

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{3}$$

For the given name identification sub-problem, we have obtained the best results using the RF classifier with n-gram counts as features (see Table 2). Despite experimenting with different architectures and hyperparameter tuning, the deep learning models have not been able to surpass the performance of some of the classical machine learning algorithms. This highlights the continued relevance of these models, especially if well-fitting feature sets can be found. It is worth mentioning that these models are much faster to train than their deep learning counterparts and have much more modest hardware requirements.

Table 2. Classifiers performance in the case of the given name identification sub-problem.

Classifier	Vectorization	Precision	Recall	F_1-score	Accuracy
RF	TF	86.47%	81.27%	**90.24%**	90.27%
SVM	TF-IDF	81.27%	75.76%	86.91%	87.09%
NB	TF	74.00%	77.87%	84.78%	84.69%
FFN	TF	82.56%	78.35%	80.35%	88.25%
RNN	TF	81.40%	76.01%	78.59%	87.20%
LSTM	TF	80.73%	73.70%	77.05%	86.39%

For the gender identification sub-problem, it is also the RF classifier with n-gram count features that obtained the best results (see Table 3). The same underperformance in the case of the deep learning models can be seen here as well. It is likely that the selected feature set [27] is not well-suited to these models.

In comparison to a purely dictionary-based approach, we expect this classifier to capture sub-word structures common to given names, allowing it to generalize better to new data. To test this hypothesis, we have gathered a list of 70 fictional character names from the online computer game World of Warcraft[5]. This game takes place in a medieval fantasy setting and as such most characters have invented names that reflect their in-game culture and ethnicity. Nevertheless, most of these names contain the same sub-word structures as real-life names, allowing the classifier to obtain an 84.29% accuracy. A sample of the predictions, both accurate and inaccurate, can be seen in Table 4.

In both cases it should be noted that all classifiers had good performance, lending credence to the hypotheses that human given names are sufficiently morphologically distinct from other English words to be easily learnable by the classifiers, and that gender information is encoded at the level of the form of given names.

[5] https://wowwiki-archive.fandom.com/wiki/Major_characters.

Table 3. Performance of the classifiers in the case of the gender identification sub-problem.

Classifier	Vectorization	Class	Recall	F_1-score	Balanced accuracy
RF	TF	*female*	91.66%	**89.22%**	88.44%
		male	85.23%		
SVM	TF-IDF	*female*	89.53%	87.07%	86.27%
		male	83.02%		
NB	TF	*female*	85.47%	83.81%	83.18%
		male	80.88%		
FFN	TF	*female*	88.52%	82.43%	85.96%
		male	83.40%		
LSTM	TF	*female*	88.64%	81.61%	85.28%
		male	81.93%		
RNN	TF	*female*	88.81%	80.73%	84.52%
		male	80.24%		

Table 4. Examples of predictions on fantasy names.

Name	Actual gender	Predicted gender
Sen'jin	male	male
Chen Stormstout	male	male
Sylvanas Windrunner	female	female
Uther the Lightbringer	male	male
Daelin Proudmoore	male	female
Velen	male	female
Ysera	female	male
Varian Wrynn	male	female

4.2 Discussion

By applying the best performing classifier to the COVID-19 domain data, described in Sect. 3.1, the predicted distribution of the users by gender is as follows: 30.76% are anonymous users, 33.71% have been classified as female, and 35.53% have been predicted as male (Fig. 3).

When excluding anonymous users, the predicted distribution (48.69% female, 51.31% male) is very close to the empirically-observed distribution (43.60% female, 56.40% male) [14]. The distribution obtained by the benchmark model is 49.58% female and 50.42% male, which is also very close. Note however that the benchmark model cannot distinguish anonymous users at all.

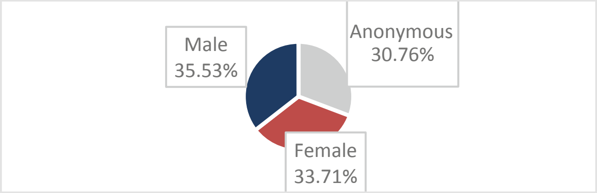

Fig. 3. The distribution of users by gender.

It is noteworthy that among the anonymous users, Black Lives Matter-related messages and COVID-19-related messages appear to have been preeminently featured in their names (see Table 5).

Table 5. Top 10 n-grams by term frequency from the name field of users marked as anonymous

Type	N-grams
Unigrams	mr (8923), com (7944), dr (7814), news (7416), health (5258), black (4956), de (4567), el (4538), blm (4495), big (4056)
Bigrams	gmail com (2013), black lives (1652), lives matter (1629), wear mask (1205), commissions open (461), high school (439), mental health (369), public health (324), blm acab (314), yahoo com (314)
Trigrams	black lives matter (1437), wear damn mask (205), black lives still (119), lives still matter (118), boop bop beep (63), grammy nominated bts (60), please wear mask (51), wear fucking mask (49), trans lives matter (47), new year new (46)

At the same time, the n-gram analysis performed on both the predictions produced by our pipeline and the benchmark model reveals that our approach has significantly more female names in the top 15 n-grams (except for "mike") than the benchmark model, which actually has many male names (see Table 6). The same phenomenon, albeit less pronounced, can be seen at the level of the users predicted as male. This issue in the case of the benchmark classifier can be caused by the fact that its predictions take only the text into account, and as such many users who are of a given gender but have an anonymous Twitter presence have been included, resulting in an incorrect correlation. It is also possible that the model simply did not generalize well from the PAN18 dataset to the domain data. Limitations such as these show that the two approaches can be used either independently or as complements, depending on the aims of the research and the available data.

Furthermore, an analysis of the top-20 n-grams has been performed on the tweets for which the author has been classified as male or female. In the case of unigrams and bigrams it has been observed that the top-20 n-grams are highly specific to the topic of the dataset, namely the the COVID-19 pandemic (e.g., "'covid", "vaccine", "coronavirus", "pandemic", "death", "covid vaccine", "covid pandemic", "covid vaccination", "wear mask", "get covid", etc.), with the same n-grams being present in the tweets written by both female and male authors.

Table 6. Top 15 n-grams by term frequency from the name field of users.

Gender	Ensemble classifier	Benchmark classifier
Female	dr (21325), mike (18218), sarah (17390), maria (11622), mary (11558), lisa (11521), laura (11172), taylor (10213), michelle (9108), kelly (8949), karen (8809), jennifer (8636), emily (8598), anna (8527), marie (8501)	news (193919), com (64199), dr (59689), john (39664), health (39660), david (34122), michael (27250), patch (26813), covid (25038), world (24087), james (23164), md (22733), paul (21763), mark (21532), iweller (20243)
Male	david (30317), john (29899), michael (25146), chris (22962), james (20781), kumar (20030), paul (18929), mark (17932), alex (16357), dr (16266), singh (14323), daniel (14191), andrew (14002), matt (13706), joe (12831)	news (222578), bot (83877), corona (72025), update (70978), corona update (70011), update bot (70011), corona update bot (70011), dr (55677), com (47832), david (43410), john (42967), covid (32276), michael (31729), health (27166), james (26695)

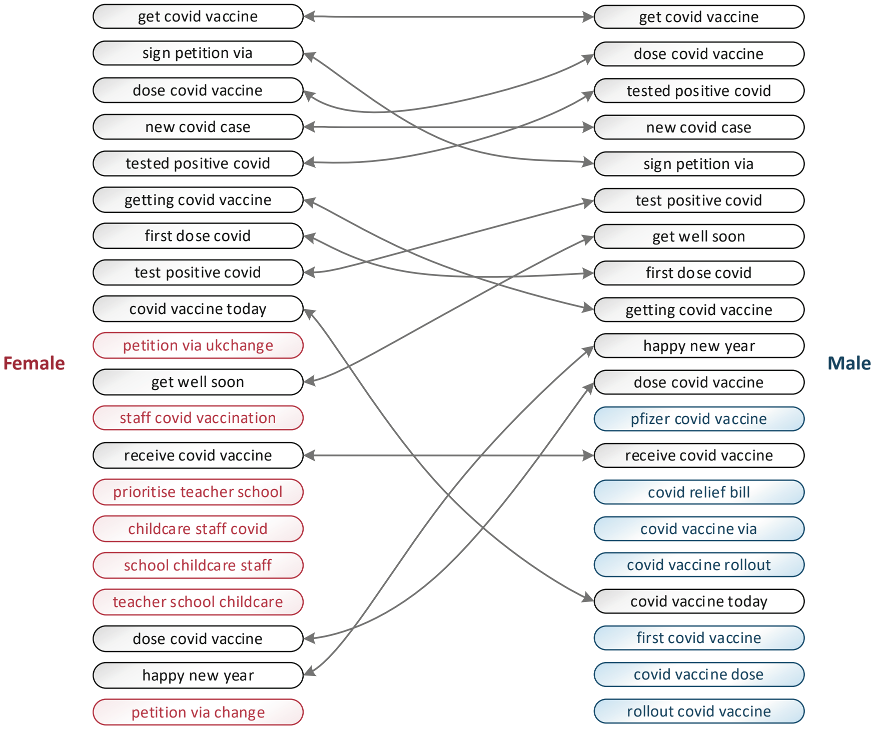

Fig. 4. Top 20 trigrams by TF-IDF score from the text of the tweets.

A significant difference between female and male written tweets becomes visible in the case of the top-20 trigrams. Thus, while the first nine trigrams are common in the discourse of both genders, as they pertain to general topics related to COVID-19 (e.g., "get covid vaccine", "new covid case", "tested positive covid", etc.), for the rest, differences can be noted among genders. While female authors focused their speech on encouraging the signing of a petition (e.g., "petition via ukchange", "petition via change") and expressed concern regarding the safety of children (e.g., "prioritise teacher school", "teacher school childcare", "school childcare staff", "childcare staff covid", "staff covid vaccination"), the discourse of male authors revolves around relief funds (e.g., "covid relief bill") and the vaccination process (e.g., "pfizer covid vaccine", "covid vaccine via", "covid vaccine rollout", "first covid vaccine", "covid vaccine dose", "rollout covid vaccine"). The common trigrams are depicted in grey in Fig. 4, while the ones that are specific to female and male authors are represented with red and blue respectively.

On the other hand, if the same n-grams analysis is performed on the tweets for which the gender of the authors has been determined using the benchmark classifier, no significant differences can be distinguished.

5 Conclusion

Correctly identifying the differences in gender discourse can be of the utmost importance in shaping the right information campaigns. The present approach is most relevant in situations where more details are available regarding the authors, such as the text of their tweets or profile photos, as a complementary analysis tool, where it can be incorporated as an important component of a multimodal gender detection approach that also considers the traditional or stylistic text features extracted from the tweets, as well as the results of the profile photos analysis.

One of the limitations of the study is that it, by necessity, does not consider the full complexity of gender within society. Our approach is also unable to detect instances of gender deception, or the use of pseudonyms that do in fact conform to standards of human given names. The approach also does not distinguish between anonymous users and organizational users, which state the name of a company, product, or institution as their name. Finally, as our reference benchmark does not leverage its full potential because we have omitted using the machine vision element; it is possible that its performance could be improved by complementing it with image-based data (though the authors report mixed results from such attempts [12]). These issues can be solved in the future by extending or modifying our approach.

Funding. This study was co-financed by The Bucharest University of Economic Studies during the PhD program.

References

1. Öztürk, N., Ayvaz, S.: Sentiment analysis on Twitter: a text mining approach to the Syrian refugee crisis. Telemat. Inform. **35**(1), 136–147 (2018). https://doi.org/10.1016/j.tele.2017.10.006

2. Ruz, G.A., Henríquez, P.A., Mascareño, A.: Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers. Future Gener. Comput. Syst. **106**, 92–104 (2020). https://doi.org/10.1016/j.future.2020.01.005
3. D'Andrea, E., Ducange, P., Bechini, A., Renda, A., Marcelloni, F.: Monitoring the public opinion about the vaccination topic from tweets analysis. Expert Syst. Appl. **116**, 209–226 (2019). https://doi.org/10.1016/j.eswa.2018.09.009
4. Kullar, R., Goff, D.A., Gauthier, T.P., Smith, T.C.: To tweet or not to tweet—A review of the viral power of twitter for infectious diseases. Curr. Infect. Dis. Rep. **22**(6) (2020). Art. no. 14. https://doi.org/10.1007/s11908-020-00723-0
5. Cristescu, M.P., Nerisanu, R.A., Mara, D.A., Oprea, S.-V.: Using market news sentiment analysis for stock market prediction. Mathematics **10**(22), 4255 (2022). https://doi.org/10.3390/math10224255
6. Flor, L.S., et al.: Quantifying the effects of the COVID-19 pandemic on gender equality on health, social, and economic indicators: a comprehensive review of data from March, 2020, to September, 2021. Lancet **399**(10344), 2381–2397 (Jun.2022). https://doi.org/10.1016/S0140-6736(22)00008-3
7. Vloo, A., et al.: Gender differences in the mental health impact of the COVID-19 lockdown: longitudinal evidence from the Netherlands. SSM - Popul. Health **15**, 100878 (2021). https://doi.org/10.1016/j.ssmph.2021.100878
8. Cascini, F., et al.: Social media and attitudes towards a COVID-19 vaccination: a systematic review of the literature. eClinicalMedicine **48**, 101454 (2022). https://doi.org/10.1016/j.eclinm.2022.101454
9. Kovacs, E.-R., Cotfas, L.-A., Delcea, C.: COVID-19 vaccination opinions in education-related tweets. In: Bilgin, M.H. Danis, H., Demir, E. (eds.) Eurasian Business and Economics Perspectives. EBES, vol. 24, pp. 21–41. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-15531-4_2
10. Cotfas, L.-A., Delcea, C., Roxin, I., Ioanăş, C., Gherai, D.S., Tajariol, F.: The longest month: analyzing COVID-19 vaccination opinions dynamics from tweets in the month following the first vaccine announcement. IEEE Access **9**, 33203–33223 (2021). https://doi.org/10.1109/ACCESS.2021.3059821
11. Cotfas, L.-A., Delcea, C., Gherai, R.: COVID-19 vaccine hesitancy in the month following the start of the vaccination process. Int. J. Environ. Res. Public Health 18(19) (2021). Art. no. 19. https://doi.org/10.3390/ijerph181910438
12. Rangel, F., Rosso, P., Montes-y-Gómez, M., Potthast, M., Stein, B.: Overview of the 6th Author Profiling Task at PAN 2018: Multimodal Gender Identification in Twitter
13. Banda, J.M., et al.: A large-scale COVID-19 Twitter chatter dataset for open scientific research—An international collaboration. Epidemiologia **2**(3) (2021). Art. no. 3. https://doi.org/10.3390/epidemiologia2030024
14. Global Twitter user distribution by gender 2022. Statista. https://www.statista.com/statistics/828092/distribution-of-users-on-twitter-worldwide-gender/. Accessed 16 Dec 2022
15. Vaswani, A., et al.: Attention is all you need. In: Guyon, I., et al. (eds.) Advances in Neural Information Processing Systems. Curran Associates, Inc. (2017). https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
16. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186. Association for Computational Linguistics, Minneapolis, June 2019. https://doi.org/10.18653/v1/N19-1423
17. Cotfas, L.-A., Delcea, C., Gherai, R., Roxin, I.: Unmasking people's opinions behind mask-wearing during COVID-19 pandemic—A Twitter stance analysis. Symmetry **13**(11), 1995 (2021). https://doi.org/10.3390/sym13111995

18. (Zack) Hayat, T., Lesser, O., Samuel-Azran, T.: Gendered discourse patterns on online social networks: a social network analysis perspective. Comput. Hum. Behav. **77**, 132–139 (2017). https://doi.org/10.1016/j.chb.2017.08.041
19. Sezerer, E., Polatbilek, O., Tekir, S.: A Turkish dataset for gender identification of Twitter users. In: Proceedings of the 13th Linguistic Annotation Workshop, pp. 203–207. Association for Computational Linguistics, Florence, August 2019. https://doi.org/10.18653/v1/W19-4023
20. Soler, J., Wanner, L.: A semi-supervised approach for gender identification. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), pp. 1282–1287. European Language Resources Association (ELRA), Portorož, May 2016. Accessed 12 Dec 2022. https://aclanthology.org/L16-1204
21. Ouni, S., Fkih, F., Omri, M.N.: Bots and gender detection on Twitter using stylistic features. In: Bădică, C., Treur, J., Benslimane, D., Hnatkowska, B., Krótkiewicz, M. (eds.) ICCCI 2022. CCIS, vol. 1653, pp. 650–660. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16210-7_53
22. Bartl, M., Leavy, S.: Inferring gender: a scalable methodology for gender detection with online lexical databases. In: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, pp. 47–58. Association for Computational Linguistics, Dublin, May 2022. https://doi.org/10.18653/v1/2022.ltedi-1.7
23. Daneshvar, S., Inkpen, D.: Gender identification in Twitter using N-grams and LSA. In: Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018). CEUR-WS (2018)
24. van de Weijer, J., Ren, G., van de Weijer, J., Wei, W., Wang, Y.: Gender identification in Chinese names. Lingua **234**, 102759 (2020). https://doi.org/10.1016/j.lingua.2019.102759
25. Rao, A.: UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences (2017). http://archive.ics.uci.edu/ml
26. Norvig, P.: Natural language corpus data. In: Beautiful Data, pp. 219–242. O'Reilly Media (2009)
27. Malmasi, S.: A data-driven approach to studying given names and their gender and ethnicity associations. In: Proceedings of the Australasian Language Technology Association Workshop 2014, Melbourne, Australia, pp. 145–149, November 2014. Accessed 12 Dec 2022. https://aclanthology.org/U14-1021