Natural Language Processing Journal 4 (2023) 100018

Contents lists available at ScienceDirect

Natural Language Processing Journal

journal homepage: www.elsevier.com/locate/nlp

Gender prediction with descriptive textual data using a Machine Learning approach☆

Babatunde Onikoyi[1], Nonso Nnamoko[2], Ioannis Korkontzelos [*,3]

Edge Hill University, United Kingdom

ARTICLE INFO

Keywords:
Gender prediction
Gender classification
Machine Learning
Twitter
Natural Language Processing
Pre-trained word embeddings

ABSTRACT

Social media are well-established means of online communication, generating vast amounts of data. In this paper, we focus on Twitter and investigate behavioural differences between male and female users on social media. Using Natural Language Processing and Machine Learning approaches, we propose a user gender identification method that considers both the tweets and the Twitter profile description of a user. For experimentation and evaluation, we enriched and used an existing Twitter User Gender Classification dataset, which is freely available on Kaggle. We considered a variety of methods and components, such as the Bag of Words model, pre-trained word embeddings (GLOVE, BERT, GPT2 and Word2Vec) and machine learners, e.g., Naïve Bayes, Support Vector Machines and Random Forests. Evaluation results have shown that including the Twitter profile description of a user significantly improves gender classification accuracy, by 10% approximately. Stanford's GLOVE embedding model, pre-trained on 2 billion tweets, 27 billion tokens and a vocabulary size of 1.2 million words, achieved the highest gender prediction accuracy, considering both the tweets and the profile description of a user. Statistical significance has been assessed using McNemar's two-tailed test.

1. Introduction

Social media are central in online communication and community building. Users of Twitter, Instagram, TikTok, YouTube and Facebook, allow users to interact remotely (Gruzd et al., 2011) and exchange multimedia content (Lu and Hsiao, 2010). Twitter is estimated to have 217M daily active users who send 500M tweets per day. 23% of its users are adults, of which 70.4% are male, 29.6% female, and 38.5% are in the age range 25–34 (Aslam, 2022). This statistics demonstrate Twitter's wide popularity and thus its suitability as a social media data source.

Tweets mainly consist of unstructured text, a source of hidden information, invaluable to corporate decision making and profit generation. Companies who analyse consumer behaviour data to produce behavioural insights perform better than their competition, by an 85% increase in sales and 25% in gross margin (Brown et al., 2017). Organisations make more profit by offering gender-tailored products or services. For example, men trade equities more often than women, getting poorer returns than women (Barber and Odean, 2001). Equity companies may assume that men are more interested than women to improve their profit, and target them to grow their customer base.

Targeting particular users is challenging as the majority do not share their personal data online. Gender classification identifies gender-specific patterns and features in tweets, to automatically predict a user's gender. In simple terms, it uses Text Mining and Natural Language Processing (NLP) to spot differences in writing style and vocabulary usage and guess users' gender (in this paper, male or female).

In this paper, we propose a gender classification method that applies NLP and Machine Learning (ML) methods on a user's tweets and profile description, i.e., a short "about me" summary. For experimentation and evaluation, we enriched Twitter User Gender Classification dataset.[4] To transform text into vectors, we employed four word embedding models: Global Vectors for Word Representation (GLOVE) (Pennington, 2014), Bidirectional Encoder Representations from Transformer (BERT) (Devlin et al., 2018), Generative Pretrained Transformer 2 (GPT2) (Wolf et al., 2020) and Word2Vec (Google, 2019). We experimented with popular ML algorithms: Naive Bayes (NB) (Scikit-learn, 2019b), Random Forest (RF) (Scikit-learn, 2018a), Decision Tree (DT) (Scikit-learn, 2009), Logistic Regression (LR) (Scikit-learn, 2014), Support

☆ This paper uses the Twitter User Gender Classification dataset, available at https://kaggle.com/datasets/crowdflower/twitter-user-gender-classification.

* Corresponding author.
E-mail addresses: onikoyib@edgehill.ac.uk (B. Onikoyi), nnamokon@edgehill.ac.uk (N. Nnamoko), yannis.korkontzelos@edgehill.ac.uk (I. Korkontzelos).
[1] ORCID: 0000-0002-0019-9123
[2] ORCID: 0000-0002-5064-2621
[3] ORCID: 0000-0001-8052-2471

[4] Freely available: kaggle.com/datasets/crowdflower/twitter-user-gender-classification.

https://doi.org/10.1016/j.nlp.2023.100018
Received 3 February 2023; Received in revised form 19 May 2023; Accepted 23 May 2023

Vector Machine (SVM) (Scikit-learn, 2018b), XGBoost (XGB) (Chen and Guestrin, 2016), Bagging (Scikit-learn, 2019a) and Voting Ensemble Classifiers (Scikit-learn, 2021) (hard and soft voting that considers all mentioned Algorithms).

1.1. Research significance

The proposed method for gender prediction using descriptive textual data has significant potential to benefit both businesses and end-users. Some potential applications include:

Accurate gender prediction on Twitter which can assist businesses in targeting each customer in a personalised manner. Businesses can design more effective marketing campaigns, leading to increased profits and customer satisfaction.

Further, determining user gender and linking it with content and the way that it is expressed can aid crime detection and investigations. For instance, in cases of cybercrime, where user's identity, gender prediction can help to narrow down the pool of potential suspects.

In addition, the method can help identifying posts generated by bots. This process can guarantee that social media analysis about the products or services of a business is based on reliable data. This is crucial for market research, where data accuracy directly affects decision making.

Finally, the method is applicable to other social media, e.g., Facebook and Instagram, and also to social science research for analysing inter-gender communication patterns. Identifying the language differences of male and female social media users can lead to more effective communication strategies and a better understanding of gender dynamics.

In summary, the proposed gender prediction method has significant application potential. It can benefit businesses and end-users by improving accuracy in data collection, enhancing marketing campaigns, aiding criminal investigations and advancing social science research.

1.2. Research motivation

Mining social media data attracts interest as it contains potentially valuable business insights. Gender prediction can help businesses personalise their marketing campaigns to advertise products and services effectively, resulting in increased profits and customer satisfaction (Chen and Skiena, 2014). However, accurate gender prediction based solely on social media data is challenging, because data is unstructured and noisy.

To address the challenge, we propose a gender identification method that considers both the tweets of a user and their profile description. We employ popular machine learning techniques and evaluate their performance on our expanded version[5] of Twitter User Gender Classification dataset. Our experiments show that including the Twitter profile description of a user significantly improves gender classification accuracy. The proposed approach has the potential to help businesses understand their target audience better (Joulin et al., 2016). It aligns with previous research that has shown how gender prediction can help businesses tailor their marketing strategies and improve customer engagement (Dobscha, 2019).

The more popular social media become, the more imminent the need for accurate gender prediction. According to the Pew Research Center, 65% of adults in the United States use social media (Perrin and Anderson, 2021) and Twitter had over 330 million active users as of 2019 (Dixon and 27, 2022). Gender prediction has potential applications in computing and business, e.g., targeted advertising and improving user experience on social media. Gender-targeted advertising has been shown to be more effective (Zarouali et al., 2022). Gender-specific content and recommendations improves user experience on social media (Joulin et al., 2016).

[5] Freely available at: data.mendeley.com/datasets/6x9srbfp6w.

1.3. Contributions

The contribution of this research are two-fold:

1. **A novel gender identification method** with improved accuracy: We have experimentally shown that **considering a user's profile description** on Twitter in addition to their tweets, **always enhances the accuracy** of gender identification. Combining tweets and profile description is more informative about the user and performs better than each of these sources in isolation. This conclusion is **novel** and generalises for all ML classifiers considered in this research work. It can have implications for gender-related research, including but not limited to **demographic studies**, **marketing** and **advertising**, **political campaigns** and **social media monitoring**.
2. **A new, larger dataset**: This research work involves the creation of a **new dataset** that contains more instances and textual volume than existing ones. A larger dataset allows to extract more **fine-grained features** than from existing datasets. The enriched dataset *[citation concealed for blind review]* is freely available for academic research purposes on Mendeley Data repository.

Overall, the contributions of this work have the potential to advance Twitter analysis and social media research, in general. Our novel dataset and findings can help to analyse data more accurately and comprehensively, leading to a deeper understanding of the role of social media in society.

In the remaining of this paper, Section 2 focuses on reviewing the state-of-the-art. Section 3 presents the methods used for data extension, pre-processing and building ML classifiers. The experimental outcomes are presented and discussed in Section 4, and Section 5 concludes the paper with some future work dimensions.

2. Related work

This section provides a comprehensive review of related research in four sections. Section 2.1 reviews contemporary research on gender classification using textual data, with a particular focus on the availability of relevant datasets for experiments and the challenges associated with them. Section 2.2 describes various feature extraction, data preprocessing methods, and machine learning techniques that have been employed in the field, and discusses their impact on classification accuracy. Finally, Section 2.3 delves into the challenges of gender classification on social media platforms, discussing the issues arising from user behaviour, profile standardisation, and the need for more diverse and representative training data. Through this literature review, we aim to provide a solid foundation for understanding the current state of research and the potential directions for future advancements in gender classification on social media platforms.

2.1. Classification with textual data

Recent research has explored gender classification/prediction based on text from social media. Research works that have used Twitter as a data source mainly consider the profile names, the content of tweets, or profile descriptions of users. Vashisth and Meehan focused on this task, using the dataset we used as our base data, but considered the content of tweets only. The best accuracy, approximately 57%, was achieved using Word2Vec and LR (Vashisth and Meehan, 2020). In this paper, we achieved a higher accuracy, considering data expansion and profile descriptions.

Liu and Ruths used the names of users as a gender classification feature, exploring the potential correlation between the first name and the gender of a user. A large dataset of gender-labelled Twitter users was created and used. An overall best accuracy of 87% was achieved in experiments that employed an SVM classifier and a 10-fold cross-validation setting (Liu and Ruths, 2013). Due to the unavailability of

the dataset, we were not able to directly apply our method for a direct comparison.

To consider the profile names of Twitter users for gender prediction, related research methods involve the creation of a dictionary of names classified as male or female (Vicente et al., 2015; Alowibdi et al., 2013a). The main shortcoming of this method is that nowadays users often use unisex names, making gender classification of names difficult or impossible. Users sometimes use pet names or abstract nicknames, making the dictionary method even less effective.

Ankit and Saleena, Vashisth and Meehan generated features from the content of tweets and achieved accuracy ranging from 47% to 97% (Ankit and Saleena, 2018; Vashisth and Meehan, 2020). Unfortunately, the employed dataset is not freely available, and thus its important properties for the task at hand are unknown. For example, it is unclear if the dataset contains usernames that map to well-known male or female names. Further, data pre-processing is not described in enough detail for reproduction.

In addition to the above studies, several other works have investigated gender classification with textual data. Burger et al. proposed a method for gender classification that considers the user's self-reported gender information as well as the linguistic and behavioural features of their social media activity (Burger et al., 2011). This approach shows promise for improving the accuracy of gender classification on social media platforms.

Alowibdi et al. proposed a method for gender classification using linguistic inquiry and word count (LIWC) features extracted from user-generated content on Twitter. The authors achieved an accuracy of 76.1% on a dataset of 6000 users (Alowibdi et al., 2013b). This approach demonstrates the potential for improving gender classification accuracy by leveraging user-generated content from social media platforms.

Furthermore, Park et al. investigated the effect of gender bias on the performance of gender classification models trained on social media data. The authors showed that gender bias in the training data can result in significant performance disparities between male and female users, with female users being more likely to be misclassified. They proposed a debiasing method that can reduce such disparities and improve gender classification accuracy on social media platforms (Park et al., 2015). This research highlights the importance of addressing bias in gender classification models and the need for more diverse and representative training data.

2.2. Extracting features for classification tasks

Published experimental results on gender classification strongly rely on the classification features, feature pre-processing and feature representation. The base dataset, that was extended in the present study, has previously been used in Angeles et al. (2021), Vashisth and Meehan (2020). The authors used the Bag of Words model (BOW) and Term Frequency-Inverse Document Frequency (TF-IDF) weighting and achieved accuracy ranging between 47% and 60% (Angeles et al., 2021; Vashisth and Meehan, 2020). We have considered BOW for our experiments. In other research for gender classification, the use of word embeddings has been shown to detect extra semantic characteristics and minimise dimensionality (Vashisth and Meehan, 2020). Accordingly, we have considered the best performing word embedding models (57% in Vashisth and Meehan (2020)), i.e., Word2Vec and GLOVE.

Further research has considered the fact that corpora, datasets and other linguistic resources do not always fully reflect and cover the variability of expression and linguistic style of different parts of the population (Bamman et al., 2014). The study used a new corpus of 14k Twitter users to explore associations between gender, language style, and social networks. Gender classification experiments yielded an accuracy of 88%.

Following extended previous research in document classification, we used a variety of ML classifiers: NB, RF, DT, LR, SVM and XGB. We also considered bagging and Voting Ensemble Classifiers, that used hard and soft voting to combine all the above classification models (Vashisth and Meehan, 2020; Bamman et al., 2014; Ankit and Saleena, 2018; Mouthami et al., 2013). In our experiments, SVM and LR, which also got satisfactory results in Ankit and Saleena, seem to be effective in combination with the baseline BOW model (Ankit and Saleena, 2018).

Moreover, attention mechanisms have been employed in the context of gender classification, as they can capture the contextual and semantic relationships within the text. For instance, Vaswani et al. introduced the Transformer model, which incorporates self-attention mechanisms to process input text more effectively (Vaswani et al., 2017). This model has been shown to improve performance in various natural language processing tasks, including gender classification.

Another avenue of exploration is the incorporation of additional sources of information, such as emojis and hashtags, to improve gender classification performance. Wijeratne et al. demonstrated that emojis could be effectively used as features for various classification tasks, including sentiment analysis and gender classification (Wijeratne et al., 2017). Similarly, incorporating hashtag information may also provide valuable insights into users' interests and preferences, further enhancing the classification model's performance.

2.3. Challenges of gender classification on social media platforms

Gender classification on social media platforms faces several challenges, such as the use of unisex names, pet names, and abstract nicknames by users. Such names can make gender classification of names difficult or impossible. Pavalanathan and Eisenstein found that users often use non-standard capitalisation and misspellings, which can further complicate the task of gender classification (Pavalanathan and Eisenstein, 2015). Additionally, Bamman et al. demonstrated that linguistic resources do not always fully reflect and cover the variability of expression and linguistic style of different parts of the population, further complicating the gender classification task.

Addressing these challenges requires the development of more advanced NLP and machine learning techniques that can capture the semantic and contextual information in user-generated content, as well as more diverse and representative training data. By building upon the methods and techniques examined in the literature review, future research can continue to improve the accuracy and effectiveness of gender classification models while addressing potential biases and challenges.

Furthermore, the insights gained from these studies can be applied to other classification tasks on social media platforms, enriching our understanding of user behaviour and preferences. In this context, gender classification on social media platforms plays a crucial role in shaping the future of research and development in both academia and industry.

In conclusion, gender classification on social media platforms is an important task with applications in advertising, social media platform development, and research. Recent research, as discussed in the literature review, has explored the use of advanced NLP and machine learning techniques, feature extraction methods, attention mechanisms, and additional sources of information to overcome challenges and improve gender classification accuracy on social media platforms.

3. Methodology

3.1. Datasets

Due to the lack of large manually annotated datasets for the gender prediction task on Twitter, we decided to develop a large dataset by expanding the only freely-available dataset with gender annotation. The Twitter User Gender Classification dataset is available on Kaggle and contains data for 20,050 users with one random tweet each. Each entry consists of the tweet's textual content, the date it was posted, the

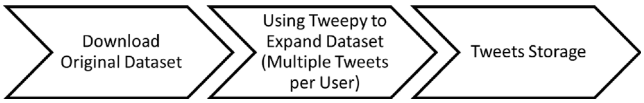

Fig. 1. Data extraction (text input).

Table 1
Statistics of the original and expanded dataset.

	Dataset	
	Original	Expanded
Total	20,050	296,108
Male	6194	159,486
Female	6700	136,622
Organisation	5942	–
Unknown	1214	–

Table 2
Statistics of the training and test part of the expanded dataset.

	Dataset		
	Training	Test	Total
Before cleaning	189,769	106,339	296,108
After cleaning	169,647	96,297	265,944
Male gender	97,545	47,572	145,117
Female gender	72,102	48,725	120,827

username of the author, their gender, their profile description and other features such as gender confidence and sidebar colour (Eight, 2016).

We expanded the Twitter User Gender Classification dataset with 296,108 more tweets, authored by authors of tweets in the original dataset. Using Tweepy (Roesslein, 2020) to access the Twitter API, we expanded the database by generating more tweets for each male and female users in the original dataset. Fig. 1 shows the data extraction process that was followed.

For the purpose of this paper, we only use a subset of the information available for each tweet: the user's id, postage date and textual content and the user's username, gender and profile description.

The Twitter User Gender Classification dataset contains gender annotations for male and female gender, brand and unknown users. Brand gender has been used to annotate tweets authored by organisations or businesses. For the purposes of this paper, we only used tweets with male and female gender annotations, reducing the original size of 20,050 tweets to 12,894, of which 6194 tweets are posted by male users and 6700 by female users. This set 12,894 tweets was the basis for the expansion process. Table 1 shows the distribution of users gender on the original and the expanded dataset. Table 2 shows how our data was split into training and test partitions and the effect of cleaning, which is discussed in detail in the following section.

3.2. Data pre-processing

Textual data from social media, especially Twitter, is unstructured and contains noisy tokens, such as hashtags and user mentions, which need to be cleaned to improve its quality and usefulness for training machine learning models. According to Wang et al., data pre-processing methods prepare data for further processing, verify its integrity and consistency, reduce data noise, fill in missing values, and structure it in databases.

To facilitate data cleaning, we created the following additional fields associated to each row in the dataset:

- **TweetsAlone (TA)**: the textual content of tweets made per user
- **TweetsDesc (TD)**: a concatenation of TA and the user's profile description
- **Desc (Desc)**: the user's profile description
- **CleanTweetsAlone (CTA)**: pre-processed version of TA
- **CleanTweetsDesc (CTD)**: pre-processed version of TD
- **CleanDesc (CD)**: pre-processed version of Desc

Pre-processing consisted of the following stages:

1. conversion of all text to lowercase
2. removal of special, non-ASCII characters
3. removal of stopwords using the Natural Language Toolkit (Bird et al., 2009)
4. removal of emoticons, retweets and favourites, hashtags, URLs and usernames starting with "@"
5. removal of duplicates
6. tokenisation using NLTK
7. Gender values were converted to binary: 1 for male and 0 for female

After the pre-processing stage, CTA, CTD and CD were processed further, as described in Section 3.3.

Table 3 displays two example tweets, one authored by a male and one by a female. Gender has been binary-coded, in preparation for machine learning algorithms. Tweet texts and descriptions have been concatenated in the TweetsDesc column. The cleaned version of the concatenation is shown in the CleanTweetsDesc (CTD) column.

3.3. Bag of words model and ML algorithms

After pre-processing, CTA, CTD and CD were converted into machine readable vectors, using the BOW model, which is a technique for transforming a text snippet into a vector consisting of word frequencies. The method is straightforward and adaptable, and it may be used to extract information from text snippets in a variety of ways. In Natural Language Processing, the BOW model is a highly prevalent element of sentence and document extraction procedures (Goldberg, 2017). The computed BOW vectors were used in succession to train a variety of Machine Learning classifiers: NB, RF, DT, LR, SVM and XGB. We also considered bagging and Voting Ensemble Classifiers, that used hard and soft voting to combine all the above classification models.

3.4. Pre-trained embeddings models

In language modelling, word embedding methods represent words or sentences in multi-dimensional vectors of real numbers. Vectors reflect the frequency of occurrence or co-occurrence of words or phrases in a corpus. A neural network's non-linearity, as well as the network's capacity to quickly integrate pre-trained word embeddings, frequently result in higher classification accuracy (Goldberg, 2015).

Instead of the BOW model, text snippets can be vectorised using pre-trained word embeddings models. For this research, we used several popular pre-trained models, GLOVE, BERT, GPT2 and Word2Vec, aiming to improve accuracy scores. Table 4 shows the results of applying GLOVE with dimensionality 200, to the sample data shown in Table 3, as a preparation step for ML classifiers.

The best performing model was GLOVE with dimensionality 200, trained on a corpus of 27 billion tokens and a vocabulary of 1.2 million distinct words. The model is trained specifically on Twitter data and achieved an accuracy of 70% in combination with the RF classifier.

Using these pre-trained embeddings models, each word was represented with the corresponding vector of the model. As a result, each CTA, CTD and CD (discussed in Section 3.2) was represented as a collection of vectors, each of which corresponded to a CTA, CTD or CD word. To combine the vectors for the words in a text snippet into a single vector representing the whole snippet, we experimented with several vector aggregation functions, per vector dimension: mean, sum, minimum and maximum. For example, using the mean function, the value of the nth dimension of the vector that represents a text snippet is the mean of all the nth dimension values of the vectors that correspond to the words in the snippet. Our experiments showed that averaging the

Table 3
Example of raw data vs. clean data.

	unit_id	Gender	Created	Description	Name	Text	TweetsDesc	CleanTweetsDesc
0	8.52*e*+17	0	2017-04-12 06:26:40+ 00:00	#RIP Dad #RIP Jalil #$$856‰Ïö ▪ ü302	_jeremiah_	Not Even tired õŸ¤¦õŸ‡â€â™,ï,	Not even tired #RIP Dad #RIP Jalil #$$ 856 ‰Ïö ▪ ü302	Even tired rip dad rip jalil
1	1.440*e*+18	1	02/10/2021 14:56	penn state alum #classof2015	_amira_	Every time I think about Karen at that Reasonably Shady party I be in tears õŸ˜ õŸ˜õŸ˜õŸ˜õŸ˜ õŸ˜	Every time I think about Karen at that Reasonably Shady party I be in tears õŸ˜ õŸ˜õŸ˜õŸ˜õŸ˜ õŸ˜ penn state alum #classof2015	Every time think karen reasonably shady party tear penn state alum classof

Table 4
Example of embedded data.

[	[	0.10790485	−0.04668283	0.20857024	...	0.14885022	0.29652923	−0.3108302	]
	[	−0.05644412	−0.15535885	0.2699906	...	−0.21408299	0.03377729	−0.24691796	]]

Table 5
Result of one of the best performing ML classifiers: an RF classifier that considers the tweets and descriptions of all users in the dataset. The GLOVE 27B 200d embedding has been used for feature encoding (Table 7, row 3, column 4).

Predictions:	[0. 0. 0. ... 1. 1. 1.]			
Accuracy score:	70%			
	Precision	Recall	f1-score	Support
Female	77%	56%	65%	47 572
Male	66%	84%	74%	48 725
macro avg	72%	70%	69%	96 297
weighted avg	71%	70%	69%	96 297
Accuracy			70%	96 297

vectors (i.e., mean) performed better than the other three aggregation methods.

To represent words that were unknown to a particular embeddings model, we used zero (0) vectors of the same dimensionality as the model. For example, when using glove.twitter.6B.300d, unknown words were represented with zero vectors of 300 dimensions. To facilitate the look-up of the embedding for a given word, all pre-trained word embeddings were loaded in memory of a dictionary data structure, for easy reference and fast retrieval.

3.5. Evaluation metrics and statistical testing

We used the Scikit-Learn library to compute our evaluation classification metrics which are: Accuracy, Precision, Recall and F-score (Pedregosa et al., 2011). In this paper, all evaluation results and associated visualisations are reported using Accuracy.

ML classifier predictions are evaluated against the real, manually assigned labels, producing an evaluation table as shown in Table 5. The table displays the predictions and accuracy score as well as precision, recall, f1-score, and the number of test instances per class and on average.

We also used McNemar's statistical significance test to compare pairs of methods and gauge the significance of any improvements. Statistical significance testing can verify that the reported improvements are real and not random, and thus are expected to generalise (Bifet et al., 2015). We applied McNemar's statistical test to compare the performance of the best trained classifier GLOVE (RF) with the baseline (SVM), using the CTD version of the data (see Section 3.2). In this application, the test evaluates the extend to which the two models agree or disagree in their predictions.

4. Experiments, results and discussion

As mentioned in Section 3.1 and shown in Table 2, the extended dataset contains activity evidence on Twitter for 296,108 users. Data pre-processing, discussed in Section 3.2, resulted in a reduction of the data size to 265,944 instances.

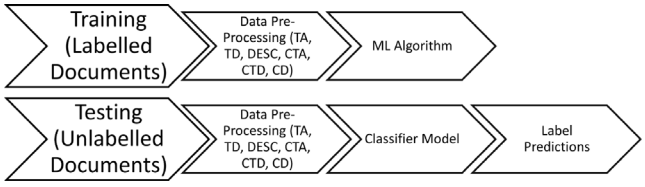

Fig. 2. Training and testing setup.

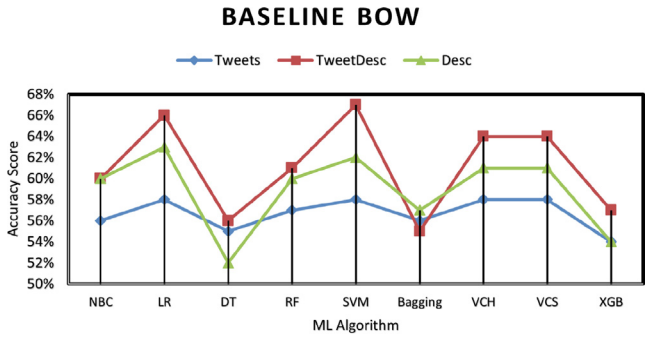

Fig. 3. Baseline results using the Bag of Words model.

After vectorisation, the data was passed through the following ML Classifiers: NB, RF, DT, LR, SVM, XGB and Bagging. Our data was split into 65% for training and 35% for testing the classifiers. Fig. 2 shows a graphical representation of the process.

The test partition was confirmed to contain instances other than those in the training set. The test set is balanced, i.e., it contains an almost equal number of instances from both classes, granting a Most Frequent Class (MFC) baseline of 50% for evaluation purposes. Angeles et al. (2021) and Vashisth and Meehan (2020) used the same base dataset as us with their train and test split set to 75:25 and 80:20, respectively. Our best ML performance is higher, even with a smaller training set of 65%.

We proceeded by vectorising the data through both the BOW and TF-IDF models, followed by evaluating several ML algorithms. The results obtained from this process can be seen in Figs. 3 and 4. It can be observed that the combination of tweets and descriptions performs better than just tweets.

We tried concatenating all multiple tweets and description of each user as one data instance but this did not improve classification accuracy. We observed that the ML model was not performing better because it was only reading one long string of text per user.

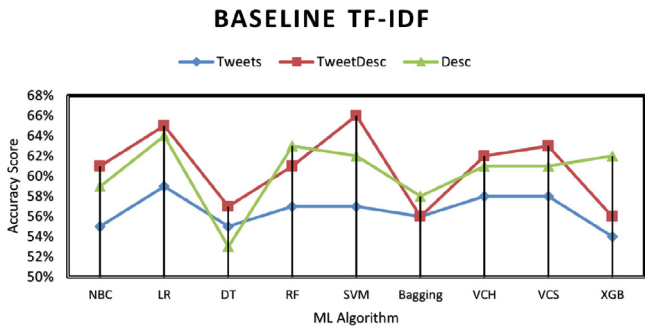

Fig. 4. Baseline results using the TF-IDF model.

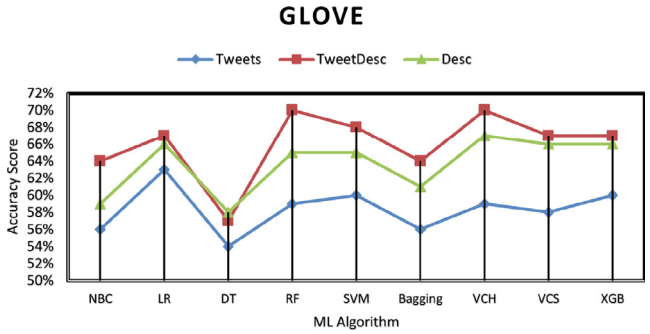

Fig. 5. Results of GLOVE pre-trained embeddings combined using the Mean function.

Concatenating all activity of a user into a single data instance, significantly reduces the number of instances in the dataset, leading to increased data sparsity. In addition, the textual content of each instance becomes too rich and diverse, which may have an effect on the classifier's ability to distinguish between genders. The best performing model, the SVM algorithm, achieved an accuracy of 67%, which is over 10% higher than other research using the same data and similar techniques, such as Angeles et al. (2021), Vashisth and Meehan (2020). We used McNemar's statistical test to confirm the statistical significance of this improvement (P value ≈ 0.0408).

In our next experiments, we vectorised the data using pre-trained word embedding models, i.e., GLOVE, BERT, GTP2 and Word2Vec. The models provided individual word vectors that were aggregated to compute a document representation using the mean, sum, minimum and maximum per vector dimension, as explained in Section 3.4. Averaging GLOVE vectors for individual words performed best. Its results are shown in Fig. 5. Its best accuracy of 70%, achieved when using a RF classifier, is higher than the best accuracy for the BOW model, 68%, which was achieved using an SVM. McNemar's statistical test confirmed the statistical significance of this improvement (P value was less than 0.0001).

The results in Fig. 5 also confirm that combining tweets and user profile description performs better than using just tweets. These results are almost 20% better than previously published results in Angeles et al. (2021), Vashisth and Meehan (2020).

The remaining pre-trained word embedding models, i.e., BERT, GTP2 and Word2Vec, achieved an overall best accuracy of 66%, each using XGB, SVM and LR. These results have also been achieved by combining tweets, tweets and description with just descriptions data and are shown in Figs. 6–8.

By analysing only the user profile description data, we observed that the accuracy scores were comparable to those obtained when combining both the profile description and tweet content. This suggests that gender prediction can be achieved with a certain degree of confidence even in the absence of users' tweets. Moreover, this approach can potentially be extended to other social media platforms where

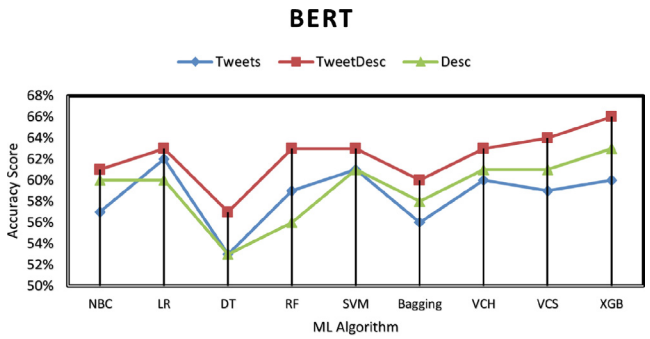

Fig. 6. Bidirectional Encoder Representations from Transformers (BERT) pre-trained model applied on CleanTweetsAlone (CTA), CleanTweetsDesc (CTD) and CleanDescription (CD).

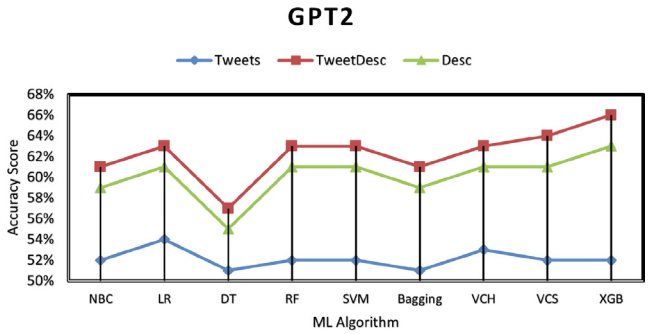

Fig. 7. Generative Pretrained Transformer 2 (GPT-2) pre-trained model applied on CleanTweetsAlone (CTA), CleanTweetsDesc (CTD) and CleanDescription (CD).

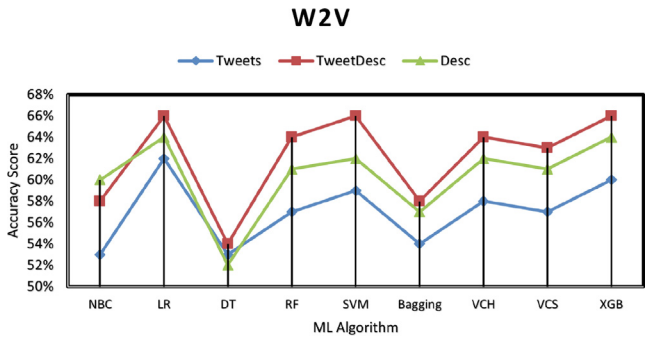

Fig. 8. Word2vec (W2V) pre-trained model applied on CleanTweetsAlone (CTA), CleanTweetsDesc (CTD) and CleanDescription (CD).

user profile descriptions are available, but not their conversations, such as YouTube. This demonstrates the versatility and adaptability of our method in addressing the gender prediction task across various online contexts.

Fig. 3 depicts the baseline BOW model applied on tweets (TA) and tweets with description (TD). It is confirmed that TA performs always worse than TD, as all TA results are less than 59%. This can be attributed to the fact that TD uses a combination of personalised features, yielding results ranging from 51% to 70%, and the combination of embedding models. Tables 6–8 show all experimental results.

4.1. Discussion

Our experimental results provide an interesting comparison of the performance of different machine learning algorithms on the data using various word embedding techniques, as seen in Table 9. The results suggest that the choice of word embedding technique can significantly

Table 6
Tweets - Data results: Accuracy of machine learning models considering various embeddings, when taking into account tweet information only.

Embeddings/Vectors	Machine learning methods								
	NBC	LR	DT	RF	SVM	Bagging	VCH	VCS	XGB
BERT - Large	**57%**	62%	53%	**59%**	**61%**	**56%**	**60%**	**59%**	**60%**
GPT2 - Large	52%	54%	51%	52%	52%	51%	53%	52%	52%
Mean - GLOVE 27B 200d	56%	**63%**	54%	**59%**	60%	**56%**	59%	58%	**60%**
Word2vec	53%	62%	53%	57%	59%	54%	58%	57%	**60%**
Baseline - BOW	56%	58%	**55%**	57%	58%	**56%**	58%	58%	54%

Table 7
Tweets and description (TweetsDesc) - Data results: Accuracy of machine learning models considering various embeddings, when taking into account tweet information alongside user profile description.

Embeddings/Vectors	Machine learning methods								
	NBC	LR	DT	RF	SVM	Bagging	VCH	VCS	XGB
BERT - Large	61%	63%	**57%**	63%	63%	60%	63%	64%	66%
GPT2 - Large	61%	63%	**57%**	63%	63%	61%	63%	64%	66%
Mean - GLOVE 27B 200d	**63%**	**67%**	**57%**	**70%**	**68%**	**64%**	**70%**	**67%**	**67%**
Word2vec	58%	66%	54%	64%	66%	58%	64%	63%	66%
Baseline - BOW	60%	66%	56%	61%	67%	55%	64%	64%	57%

Table 8
Description (Desc) - Data results: Accuracy of machine learning models considering various embeddings, when taking into account user profile description information only.

Embeddings/Vectors	Machine learning methods								
	NBC	LR	DT	RF	SVM	Bagging	VCH	VCS	XGB
BERT - Large	**60%**	60%	53%	56%	61%	58%	61%	61%	63%
GPT2 - Large	59%	61%	55%	61%	61%	59%	61%	61%	63%
Mean - GLOVE 27B 200d	59%	**66%**	**58%**	**65%**	**65%**	**61%**	**67%**	**66%**	**66%**
Word2vec	**60%**	64%	52%	61%	62%	57%	62%	61%	64%
Baseline - BOW	**60%**	63%	52%	60%	62%	57%	61%	61%	54%

Table 9
Results comparison: Accuracy levels of ML models considering various embeddings and baseline feature representations (BOW and TF-IDF) with tweet information only, user description only, or both. The combination of tweets and user description performs best in the majority of cases.

	Model	Tweets	Description	Machine learning methods				
				LR	SVM	NB	RF	XGB
BOW	Vashisth and Meehan (2020)	✓	✗	54%	53%	54%	48%	55%
	Our model	✓	✗	58%	58%	56%	57%	54%
	Our model	✗	✓	63%	62%	**60%**	60%	54%
	Our model	✓	✓	**66%**	**67%**	**60%**	**61%**	**57%**
TF-IDF	Angeles et al. (2021)	✓	✗	–	59%	61%	–	–
	Our model	✓	✗	59%	57%	55%	57%	54%
	Our model	✗	✓	64%	62%	59%	**63%**	**62%**
	Our model	✓	✓	**65%**	**66%**	**61%**	61%	56%
W2Vec	Vashisth and Meehan (2020)	✓	✗	57%	53%	–	48%	55%
	Our model	✓	✗	62%	59%	53%	57%	60%
	Our model	✗	✓	64%	62%	**60%**	61%	64%
	Our model	✓	✓	**66%**	**66%**	58%	**64%**	**66%**
GLOVE	Vashisth and Meehan (2020)	✓	✗	54%	53%	–	48%	52%
	Our model	✓	✗	63%	60%	56%	59%	60%
	Our model	✗	✓	66%	65%	59%	65%	66%
	Our model	✓	✓	**67%**	**68%**	**63%**	**70%**	**67%**

All experiments in this Table used the same initial dataset (i.e., before expansion). Note that Angeles et al. (2021) tuned the parameters of their SVM and NBC models, using character and syntax-based meta-attributes. In this paper and in Vashisth and Meehan (2020), the parameters of ML algorithms have not been tuned. Thus, our findings are more comparable to those of Vashisth and Meehan (2020).

impact the accuracy of the classification model and experiment. The results are presented in terms of accuracy percentages for each algorithm and each type of text data (Tweets, TweetDesc, and Desc).

The baseline models using Bag of Words (BOW) and Term Frequency-Inverse Document Frequency (TF-IDF), shown in Figs. 3 and 4, achieved moderate accuracy levels ranging from 52% to 66% across the different models and datasets. However, when GloVE was used, the accuracy of the models increased, especially for the TweetDesc dataset, as shown in Fig. 5. The best performing models using GloVE achieved accuracy levels ranging from 64% to 70%. Word2Vec also showed improved performance over BOW, with accuracy levels ranging from 52% to 66%, which can be seen in Fig. 8.

BERT and GPT2, two advanced deep learning models, demonstrated even higher accuracy compared to BOW and Word2Vec, but still lower than GloVE embeddings. For instance, the best-performing models using BERT achieved accuracy levels ranging from 60% to 66%. GPT2 also demonstrated notable performance gains, particularly for the TweetDesc dataset, with accuracy levels ranging from 61% to 66%, as shown in Figs. 6 and 7.

Overall, our experiments indicate that the use of advanced word embedding techniques, such as GloVE, BERT, and GPT2, can result in significant improvements in the accuracy of classification models for Twitter data. These results suggest that for classification tasks on Twitter data, a suitable word embedding technique should be selected

carefully. Additionally, these findings also highlight the potential of deep learning models in improving classification accuracy, particularly when using more complex text sources. Table 9 confirms this.

4.2. Experimental setup and data provision

All methods have been implemented in Python 3.9, using Jupyter Notebook IDE and numerous libraries, such as Tweepy API for data extraction, NLTK for data pre-processing and sklearn for ML classification.

The data used in our experiments has been made freely available on Mendeley Data.[6] We have also developed a Python script for re-hydrating tweets in the dataset using the provided tweet IDs.

5. Conclusion

This paper focused on gender prediction on Twitter, using a freely available dataset of 20,050 Twitter users with one random tweet per user. The dataset has been expanded to 296,108 which contained multiple random tweets per user, and experimentation has been conducted, considering a wide variety of pre-trained word embedding models and Machine Learning algorithms. It was observed that the best performing model is GloVE, pre-trained on 2 billion tweets, that contain 27 billion tokens and a vocabulary of 1.2 million distinct words trained. The best performing ML algorithm when combined with GloVE is Random Forest, achieving an accuracy of 70%. This accuracy level is higher in comparison to other published research that used the same base dataset and similar methods as mentioned in 4.

Our experiments have shown that the combination of the tweets of a user with the user's Twitter profile description achieves over 10% higher accuracy than applying the same methods on just the tweets of the user. The statistical significance of this superior performance has been confirmed by McNemar's statistical test between our best baseline Bag-of-Words model and the best pre-trained method GLOVE for both Tweets and Tweets with Description. For Tweets, the two-tailed P value was 0.0408 which by conventional criteria denotes a statistically significance difference, whereas for Tweets and Description, the two-tailed P value is less than 0.0001 which denotes extreme statistical significance.

Concerning further research as future work, it would be interesting to investigate the effect of adding more features to the textual data, such as the number of emoticons or stop words that each gender uses, the average length of sentence per user or even the average number of parts of speech used in the sentences. These features could improve the gender identification capability of ML models, using the same source information. A higher classification power would probably make this method applicable to the task of automatically identifying human users as opposed to bots, which is similar. This could be a good step in identifying automated spam accounts and reducing their effect on the communication of human users on social media.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Authors would like to thank the Department of Computer Science at Edge Hill University, UK, for providing resources and time for the design and implementation of this research work.

[6] The data set is freely available at: data.mendeley.com/datasets/6x9srbfp6w.

References

Alowibdi, J.S., Buy, U.A., Yu, P., 2013a. Empirical evaluation of profile characteristics for gender classification on Twitter. In: 2013 12th International Conference on Machine Learning and Applications, Vol. 1. IEEE, pp. 365–369.

Alowibdi, J.S., Buy, U.A., Yu, P., 2013b. Language independent gender classification on Twitter. In: 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013). pp. 739–743.

Angeles, A., Quintos, M.N., Octaviano, M., Raga, R., 2021. Text-based gender classification of Twitter data using Naive Bayes and SVM algorithm. In: TENCON 2021 - 2021 IEEE Region 10 Conference. TENCON, IEEE, Piscataway, pp. 522–526.

Ankit, Saleena, N., 2018. An ensemble classification system for Twitter sentiment analysis. Procedia Comput. Sci. 132, 937–946.

Aslam, S., 2022. Twitter by the numbers: Stats, demographics & fun facts. https://www.omnicoreagency.com/twitter-statistics/.

Bamman, D., Eisenstein, J., Schnoebelen, T., 2014. Gender identity and lexical variation in social media. J. Socioling. 18 (2), 135–160.

Barber, B.M., Odean, T., 2001. Boys will be boys: Gender, overconfidence, and common stock investment. Q. J. Econ. 116 (1), 261–292.

Bifet, A., de Francisci Morales, G., Read, J., Holmes, G., Pfahringer, B., 2015. Efficient online evaluation of big data stream classifiers. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '15, ACM, pp. 59–68.

Bird, S., Klein, E., Loper, E., 2009. Natural language processing with Python. http://deposit.d-nb.de/cgi-bin/dokserv?id=3281322&prov=M&dok_var=1&dok_ext=htm.

Brown, B., Kanagasabai, K., Pant, P., Pinto, G.S., 2017. Capturing value from your customer data. https://search.proquest.com/docview/2372094689.

Burger, J.D., Henderson, J., Kim, G., Zarrella, G., 2011. Discriminating Gender on Twitter. Technical Report, MITRE CORP BEDFORD MA BEDFORD United States.

Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16.

Chen, Y., Skiena, S., 2014. Building sentiment lexicons for all major languages. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 383–389.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Dixon, P.b.S., 27, J., 2022. Twitter mau worldwide 2019.

Dobscha, S., 2019. Handbook of Research on Gender and Marketing. Edward Elgar Publishing.

Eight, F., 2016. Twitter user gender classification. https://www.kaggle.com/crowdflower/twitter-user-gender-classification.

Goldberg, Y., 2015. A primer on neural network models for natural language processing. CoRR abs/1510.00726.

Goldberg, Y., 2017. Neural Network Methods in Natural Language Processing, first ed. Morgan & Claypool Publishers, San Rafael, pp. 1–309.

Google, 2019. Google code archive - long-term storage for google code project hosting. https://code.google.com/archive/p/word2vec/.

Gruzd, A., Wellman, B., Takhteyev, Y., 2011. Imagining Twitter as an imagined community. Am. Behav. Sci. (Beverly Hills) 55 (10), 1294–1318.

Joulin, A., Grave, E., Bojanowski, P., Mikolov, T., 2016. Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759.

Liu, W., Ruths, D., 2013. What's in a name? using first names as features for gender inference in twitter. In: 2013 AAAI Spring Symposium Series. pp. 10–16.

Lu, H.-P., Hsiao, K.-L., 2010. The influence of extro/introversion on the intention to pay for social networking sites. Inf. Manage. 47 (3), 150–157.

Mouthami, K., Devi, K.N., Bhaskaran, V.M., 2013. Sentiment analysis and classification based on textual reviews. In: 2013 International Conference on Information Communication and Embedded Systems. ICICES, IEEE, pp. 271–276.

Park, G., Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Kosinski, M., Stillwell, D.J., Ungar, L.H., Seligman, M.E., 2015. Automatic personality assessment through social media language. J. Personal. Soc. Psychol. 108 (6), 934.

Pavalanathan, U., Eisenstein, J., 2015. Confounds and consequences in geotagged Twitter data. arXiv preprint arXiv:1506.02275.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al., 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12 (Oct), 2825–2830.

Pennington, J., 2014. GloVe: Global vectors for word representation. https://nlp.stanford.edu/projects/glove/.

Perrin, A., Anderson, M., 2021. Share of U.S. adults using social media, including Facebook, is mostly unchanged since 2018.

Roesslein, J., 2020. Tweepy: Twitter for Python!. https://github.com/tweepy/tweepy.

Scikit-learn, 2009. Decision Trees documentation. https://scikit-learn.org/stable/modules/tree.html.

Scikit-learn, 2014. Logistic Regression documentation. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html.

Scikit-learn, 2018a. Random Forest Classifier documentation. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

Scikit-learn, 2018b. Support Vector Machines documentation. https://scikit-learn.org/stable/modules/svm.html.

Scikit-learn, 2019a. Bagging Classifier documentation. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.BaggingClassifier.html.

Scikit-learn, 2019b. Naive Bayes documentation. https://scikit-learn.org/stable/modules/naive_bayes.html.

Scikit-learn, 2021. Voting Classifier documentation. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.VotingClassifier.html.

Vashisth, P., Meehan, K., 2020. Gender classification using Twitter text data. In: 2020 31st Irish Signals and Systems Conference. ISSC, IEEE, pp. 1–6.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.

Vicente, M., Batista, F., Carvalho, J.P., 2015. Twitter gender classification using user unstructured information. In: 2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). IEEE, pp. 1–7.

Wang, H., Ma, C., Zhou, L., 2009. A brief review of machine learning and its application. In: 2009 International Conference on Information Engineering and Computer Science. IEEE, pp. 1–4.

Wijeratne, S., Balasuriya, L., Sheth, A., Doran, D., 2017. Emojinet: An open service and api for emoji sense discovery. In: Proceedings of the International AAAI Conference on Web and Social Media, Vol. 11. pp. 437–446.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T.L., Gugger, S., Drame, M., Lhoest, Q., Rush, A.M., 2020. Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Association for Computational Linguistics, Online, pp. 38–45.

Zarouali, B., Dobber, T., De Pauw, G., de Vreese, C., 2022. Using a personality-profiling algorithm to investigate political microtargeting: assessing the persuasion effects of personality-tailored ads on social media. Commun. Res. 49 (8), 1066–1091.