Gender Detection on the Web

Xiaoqi Ma

RWTH Aachen University
xiaoqi.ma@rwth-aachen.de

Abstract. Historical data and user demographic information are essential in personalized web applications. However, for new users, there is no historical data and it is infeasible to directly obtain personal information from users due to privacy concerns. Thus, the ability to predict user demographics, particularly gender, plays a vital role. After literature reviews, we conclude three main method groups which exploit textual features, visual features and behavioral features respectively. Gender classification could be trained on textual features based on the fact that females and males have huge differences in linguistic writing style. And the face image as well as the semantic meaning in images could be utilized to detect gender. For behavioral features, such as web browsing history, web search queries, and e-commerce data, they could be explored to build a gender predictor and it has been an active research topic on this type of features. In this seminar paper, various methods and performance measurement for each method group will be discussed.

Keywords: Gender detection · Textual · Visual · Behavioral

1 Introduction

With the incredible growth of the Internet and remarkable emergence of popular social media platforms, like Facebook, Twitter, a substantial amount of user-generated data are available on the web, which boosts the appearance of web services. Relying on customized web services, the effectiveness of many web applications and user experiences can be enhanced. For example, e-commerce websites could conduct market basket analysis by discovering user behavior, in order to provide precise item recommendation for specific users. Besides, for commercial purposes, search engines are heavily contingent upon personalized services, also known as targeted advertisement, which is regarded as one of the main technologies to raise the effectiveness and profits of digital marketing. What's more, personalized services are beneficial to attach membership and establish trust in community [1].

It comes as no surprise that personalization could be based on historical data, e.g. previous web browsing history or shopping records, which, nonetheless, requires that users have registered in the web application or have used it before [2]. But for new users or guest users, the historical-data driven approach is not applicable since there is no prior knowledge. For those who are newcomers to the system, demographic attributes could be explored to offer personalized services. Demographic attributes could be age, gender, education level and etc. And gender is treated as a striking feature to represent user's characteristic among all of them, which is also the attribute that needs to be further investigated in this seminar paper. However, nowadays users are reluctant to expose their personal information due to privacy concerns, as well as the fact that law constrains the leak of sensitive user information on social media platforms [3]. Due to the privacy protection on the web, it is not feasible to obtain user gender information directly, therefore, predicting user's gender on the web naturally becomes an achievable way, which is an interesting topic that considerable research efforts have been devoted to [4].

On the web, there are multiple features could be employed to improve gender prediction accuracy. During the literature review, three dominating research method groups are discovered. One method group focuses on textual data written by users, such as blogs, movie reviews, website comments and tweets. In a sense, sentimental analysis and stylometry identification could be cast on those textual data, in order to generate key features to detect gender on the web [4]. The second method group concentrates on visual data. Features adopted to predict gender could be derived from face image or social media images by interpreting the semantics of the images [5]. The method group that using users'behavioral data to detect gender is also demonstrated in a number of studies. The behavioral data includes but not limited to web browsing history, web search queries and user clickstream data [6]. With regard to increasing gender prediction accuracy, a specific feature or some feature combinations are going to be discovered.

In the following, related work review will be described in chapter 2. Chapter 3 characterizes several existing state-of-art methods in each method group respectively, followed by chapter 4, revealing the corresponding experimental results. In chapter 5, a short summary will be covered as well as some implications for future work.

2 Related Work

As stated before, there are three main method groups, each favors one type of features. In this section, several related work reviews outlining each method group to detect gender on the web are discussed.

2.1 Textual features

Linguistic differences, especially writing style or speaking style, lie between female and male users. To be specific, character usage, writing syntax, functional words and word frequency could be regarded as linguistic features [7] and previous studies have already shown those distinctions. Deitrick [7] found out from emails that females favor emotionally language and incline to use more adjectives and adverbs, while males make more typos and commit more grammatical errors. Similarly, after investigating tweets by utilizing sociolinguistic-based model and N-gram feature model, Rao *et al.* [8] drew the conclusion that females tend to use more emoticons, ellipses as well as repeated exclamation. In addition, features derived from blogs text facilitated to identify gender from weblogs, which have been studied by Herring *et al.* [9], and Yan *et al.* [10]. Prior research also implied that first names could be employed as the key feature to predict gender. Mislove *et al.* [11] mapped those self-reported names of Twitter users to a name database reported by the U.S. Social Security Administration in order to detect gender. In [12], authors first evaluated several widely accepted name-based gender detection methods, then proposed a mixed method that combined name-based features with image-based features. Apart from textual data, word usage features extracted from conversational data were considered to predict gender for individual speakers [13].

2.2 Visual Features

In computer vision filed, automatic gender detection from face image has been intensively investigated, and an overview of existing state-of-art face recognition approaches was demonstrated in [14], where a CNN-based ensemble model was proposed for gender recognition, setting up a new record performance of 97.31% accuracy on the LFW datasets. Although with high accuracy to predict gender based on the face image, in more general cases, face recognition alone is insufficient for precise gender inference due to effects of image occlusion, image blur or other technical reasons [5]. Therefore, Merler *et al.* developed a method to estimate gender through analyzing the semantics implied in those pictures posted by users on social media. It also should be pointed out that a stacked-SVM gender classifier was implemented by You *et al.* [15], which was built on top of topics modeling.

2.3 Behavioral features

Besides the methods relied on textual and visual features, it is noteworthy knowing that substantial researches have been conducted on behavioral features. And there are various behavioral features could be utilized, such as web-browsing history, web-search history, web clickstream data and etc. Hu *et al.* focused on webpage click-through data to detect user's demographics, including gender. First, Bayesian Framework was employed on webpage features which are content-based and categorical-based to obtain the gender tendency of web pages. Then authors assumed similar web browsing behaviors of users indicated similar gender tendencytowards the web pages after smoothing approach processing. In this case, they could simply build a gender classifier based on the web pages visited by users [6]. T. M. Phuong *et al.* followed a similar approach, they extracted topic-based features, time features and sequential features from web browsing history to train a stacked-SVM classifier to predict gender. One interesting finding was males inclined to switch between different webpage categories more regularly [4]. Other features based on their catalog viewing information on e-commerce system, such as shopping records, items viewed, and time of access, were adopted to train the gender classifier [2]. In paper [16], by mapping both Facebook likes data and Bing search queries to ODP(Open Directory Project), the coarse-grained common representation was able to build the gender predictor.

3 Methods

The main topic of this seminar paper is how to precisely classify user gender on the web. After literature reviews, we have found three dominating method groups to approach the topic, by exploring textual features, visual features and behavioral features respectively. Nevertheless, it is not achievable to simply compare model metrics (accuracy, F1 score and etc.) obtained from the model evaluation from each method group by reason of the diverse data applied for training. Therefore, the techniques exploited to predict gender on the web are introduced separately for each method group in the following.

3.1 Textual features model

As previously discussed, the textual data such as tweets, blogs, movie reviews, could be utilized to train a gender classifier. Twitter, one of the most popular social networking forums, naturally becomes a great source for researchers to investigate on. And recent developments of referring gender from tweets have attracted much attention. One novel method to detect gender of Twitter users was proposed by Rao et al. [8], who introduced stacked-SVM-based classification algorithms over a vast of features gleaned from Sociolinguistic model and N-gram model.

Though predicting gender from their first names had high accuracy as described in [12], many Twitter users tend not to use their real names, causing difficulty in distinguishing those true names. As a result, only the message content of the tweets or reply for the tweets were crawled and manually annotated by two independent annotators, served as the training datasets.

In [17], it has concluded that people will have their own gender characteristics while using the language either from physiology or psychology, causing the sociolinguistic differences, which facilitates to detect gender. In this case, the utterance-choosing differences could be treated as features, including vocabulary difference, syntactic difference and etc. As displayed in Table 1, some sociolinguistic features had been acquired from tweets messages. Then an SVM [18]based binary classifier was built based on these sociolinguistic features [8]. Besides, the N-gram language model was widely used to analyze texts, which was also included in the paper. First, the authors preprocessed on texts, like digits normalization, text lowercase, tokenization, emoticons and specific punctuation preservation, to produce unigrams and bigrams. Then they brought out an SVM-based classification model employing unigrams and bigrams. Finally, the best performing classification model discovered was to use stacked-SVM to combine predictions from Sociolinguistic and N-gram models accompany with their prediction weights [8].

Table 1. List of Sociolinguistic features [8]

Feature	*Examples*
Simleys	A list of emoticons compiled from the Wikipedia
OMG	Abbreviation for “Oh My God”
Ellipses	“....”
Possesive bigrams	my_XXX, our_XXX
Repeated alphabets	niceeee, noooo waaaay
Laugh	LOL, ROTFL, haha
Exasperation	Uhg, mmmm, hmmm, ahh
Agreement	yea, yeah, ohya
Honorifics	dude, man, bro
Excitement	A string of exclamation symbols (!!!!)
Pulzzled Punct	A combination of any number of ? and ! (!??!!)

3.2 Visual features model

Accordingly, visual features could be derived from images, either from the profile picture or from the pictures posted by users on popular image-based social media platforms, like Instagram and Pinterest. In order to infer gender, one simple method is to extract the human face from the profile image, then apply the face image detection method. Surprisingly, this method was able to achieve the record performance around 97% F1 score as stated in [14], by using ensemble CNN-based model. However, face image detection from a profile image is not sufficient for gender prediction and cannot be fully trusted. For one reason, the profile image might contain multiple faces of different gender, leading inaccurate prediction without the knowledge of the real account holder. Besides, if the image is blurry or the face is not visible, the prediction results might get impacted. In an extreme situation, it is infeasible to make face detection if there is no face in the profile image.

Considered that those pictures posted by users on social media are tied in with their gender. Therefore, gender prediction based on semantic analysis of social media images was proposed to alleviate the above detection problems. The idea was illustrated in [15], by first extracting the concept of the image based on the content, then mapped it to predefined semantics, including 33 categories (see Table 2), such as animal, art, sports and etc. And the framework for building user-level visual topic distribution in different categories was shown in Figure 1. Later, they utilized the visual content features to train SVM classifiers on top of the semantic model. To be more specific, each user could be denoted by a feature vector with length 33, each element in the vector representing one category. The vector was calculated by accumulating the number of images belonging to each category then normalized by the number of images posted by all training users with respect to categories. Finally, the feature vector was fed to the SVM classifiers.

Table 2. List of Categories

Animals	Architecture	Art	Cars & Motorcycles	Celebrities	Design	DIY & Crafts
Education	Film, Music & Books	Food & Drink	Gardening	Geek	Hair & Beauty	Health & Fitness
History	Holidays & Events	Home Decor	Humor	Illustrations & Posters	Kids	Mens Fashion
Outdoors	Photography	Products	Quotes	Science & Nature	Sports	Tattoos
Technology	Travel	Weddings	Womens Fashion	Other		

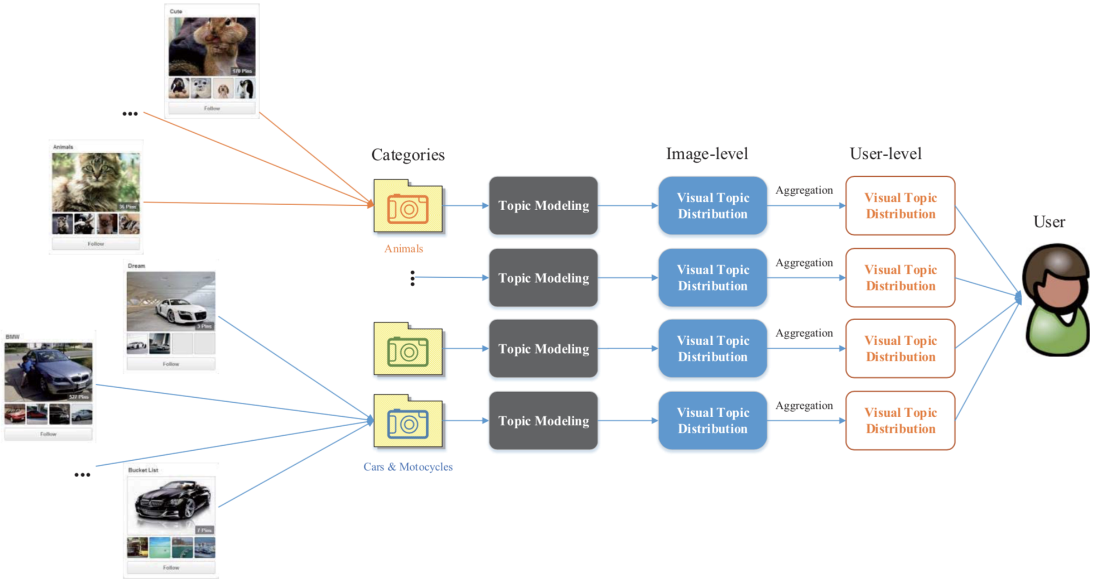
Cute
Animals
Dream
BMW
Bucket List
Categories
Animals
Cars & Motocycles
Topic Modeling
Topic Modeling
Topic Modeling
Topic Modeling
Image-level
Visual Topic Distribution
Visual Topic Distribution
Visual Topic Distribution
Visual Topic Distribution
Aggregation
Aggregation
Aggregation
Aggregation
User-level
Visual Topic Distribution
Visual Topic Distribution
Visual Topic Distribution
Visual Topic Distribution
User

Fig. 1. Framework for building user-level visual topic distribution in different categories

3.3 Behavioral features model

It is not hard to observe that there are plenty of behavioral features being exploited to detect gender on the web, including web search history, web search queries, e-commerce data and etc. In this seminar paper, it is intended to have a deep look into the method that made use of Facebook likes data and web search queries data to predict the gender of Bing search users.

The idea was that they aimed to develop a coarse-grained common representation which matched Facebook Likes against search queries (see Figure 2). And the key intuition was that whether a user is interested in some specific categories or not relies on the gender and other demographics, independent of the web applications. In this case, the representation for each user was a 219-dimensional feature vector, denoting 219 various categories on the web within the Open Directory Project (ODP). Then a gender classifier was trained based on the Facebook Likes data, and afterwards to predict the gender of users by applying the classifier to the Bing search queries [16].

To be more detailed, for each Facebook like data, the title of the corresponding liked item or event was extracted and fed into the Bing search engine. After obtaining the top 10 results returned from the search engine, assigning three categories for each result then fitting the results to feature vectors. Repeat the same process for each Facebook like data and each user. In general, for each user, a 219-dimensional feature vector should be maintained, which also requires normalization properly. Then apply the logistic regression model with L2 regularization to fit the input feature vectors. Following the same concept, a 219-dimensional feature vector was preserved for each Bing search user. Afterwards, we could predict the gender of the Bing search users using the pre-trained logistic regression model.

In a nutshell, we could train a gender classifier based on user Facebook Likes data, then tested it on users distinguished by their search queries.

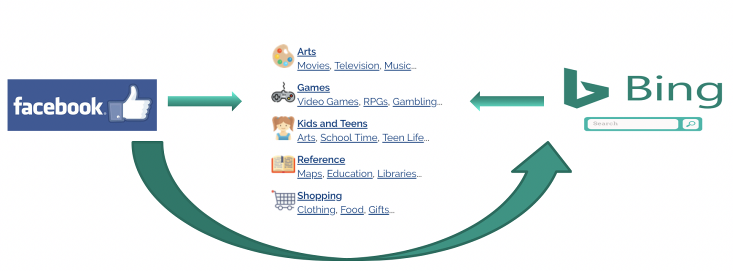
facebook
Arts
Movies, Television, Music...
Games
Video Games, RPGs, Gambling...
Kids and Teens
Arts, School Time, Teen Life...
Reference
Maps, Education, Libraries...
Shopping
Clothing, Food, Gifts...
Bing
Search

Fig. 2. Common representation between Facebook likes and Bing search queries

4 Results & Discussion

In this section, experiments were conducted to evaluate the performance of methods stated in previous section respectively. Various experiment results and evaluation metrics were exposed in the following.

4.1 Results for Textual features model

It needs to be mentioned that partial results are directly borrowed from the original paper [8], and reorganized in the below. In Table 3, we could find that stacked-SVM method which combined both sociolinguistic feature and N-gram feature outperformed the other two, achieving 72% accuracy. Also from Table 4, it could be simply acknowledged that females tended to use emoticons, ellipses more often. In addition, females were prone to use OMG four times more than males. And males were less possible to use repeated character and exclamation in tweets. Besides, an overview of performance measurement on textual-features based classification model is listed in Table 5. As you may notice, word embedding techniques like BOW and Glove could facilitate to predict gender with accuracy 82%. And the last row showed that if just first names were used as input features, the prediction accuracy reached around 82%. However, if we combined the features derived from face images and first names, the accuracy could rise to 92%. It gives us some indications that we could combine several types of features.

Table 3. Gender detection results [8]

Model	*Accuracy*
Sociolinguistic	71%
N-gram	69%
Stacked-SVM	72%
Prior	50%

Table 4. Sociolinguistic features expressed as relative frequency of females and males [8]

Features	Example	#female/#male
Emoticons	:), :D	3.5
Ellipses	....	1.5
Character repetition	nooo waaay	1.4
Repeated exclamation	!!!!	2.0
Puzzled punctuation	!?!!??!	1.8
OMG	Oh My God	4.0

Table 5. Textual features based related work

Author	Data	Methods	Performance(F1)
E.Vasilev [19]	Reddit Comments	BOW, Glove, CNN	82%
D. Rao et al., [8]	Tweets	Sociolinguistics-features, N-gram-features	72%
F. Karimi et al., [12]	Names/face Images	Name/Image based detection	82%/92%

4.2 Results for Visual features model

As stated in the paper [15], 10-fold cross validation was conducted to evaluate the performance of image semantic analysis in terms of accuracy, precision, recall and F1 score. Results could be found in Table 6. It was found out that this method had higher performance while detecting female users than detecting male users, which could be argued that female users tended to post more pictures, leading more stable results for prediction. Overall the average accuracy and F1 score measure were around 66%, which didnt perform well enough. However, a similar idea was also experimented by Merler [5] and the performance was improved to 76% by enlarging the training dataset and employing some additional visual information, such as the colors. Also, an overview of related work about exploiting visual features to detect gender is listed in Table 7. In this table, we observed that face image detection method performed best, but it suffered many challenges as explained before. Therefore, one possible way to imporve gender prediction is to combine the face image detection and image semantic analysis.

Table 6. Performance of image semantic analysis

Class	Accuracy	Precision	Recall	F1-score
Female	76%	63%	76%	69%
Male	55%	70%	55%	61%
Average	66%	66%	66%	65%

Table 7. Visual features based related work

Author	Data	Methods	Performance(F1)
Q. You et al., [15]	Social network images	Stacked-SVM	66%
M. Merler et al., [5]	Social media images	SVM	76%
G. Antipov et al., [14]	Face images	Ensemble CNN	97%

4.3 Results for Behavioral features model

The Facebook Likes data were gathered from myPersonality Facebook, and the search queries were collected from Bing Query Logs. As described in the method section, there were two classifiers, one for females detection and the other one for males detection. Classification models were trained on Facebook Likes data, then we applied classifiers to Bing search queries data to detect gender. In practice, the authors also tested models on test data of Facebook Likes in order to compare the performance with testing results on Bing search queries data. For measuring the model performance, receiver operating characteristic (ROC) curve was used, which is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. Just to know, in the binary classification tasks, the model performs better if the area under the ROC curve is closer to 1, which indicates that this classification is perfect if the area is 1.

The experiments results could be tracked back from [16]. It showed that the average accuracy for gender detection achieved 83% and 80% on Facebook test data and Bing queries test data respectively. Not surprisingly, the performance was slightly better when tested on Facebook since the models were trained on Facebook data. Nevertheless, the relative difference was not significant, which further stood by the assumption that the relation between user interests and user demographics was independent of Facebook and Bing search engine. Accordingly, an overview of related work about exploiting behavioral features to infer gender is listed in Table 8.

Table 8. Behavioral features based related work

Author	Data	Methods	Performance(F1)
Hu, et al., [6]	Webpage click-through data	Bayesian Framework	80%
B. Bi et al., [16]	Facebook likes & Bing search queries	Logistic regression	80%
T. M. Phuong et al., [4]	Web browsing history	SVM, LDA, N-gram	81%
Zhong et al., [1]	Location checkins	Tensor Factorization	81%

5 Implications

In this section, previous work will be summarized and the implications for future study will be discussed. In addition, some critical thinking will be cast on after reviewing current existing literature.

First, it has been acknowledged that many web applications offer personalized services based on not only historical data but also the demographics of users. Due to the lack of user information, the demographic detection method comes into mind. Among the user characteristics, gender is one of the easiest attributes that could be predicted from the web, which is also the focus of this seminar paper. Though other demographics are not explored here, they still deserve further studying.

For textual feature analysis, almost all papers concentrate on English texts and to study the linguistic differences between females and males. Though it might achieve reasonable results in English, it is not perfect and might suffer from other language systems, like Japanese [22]. Since Japanese and English are in different language system, their linguistic features are not exactly consistent with each other, causing poor performance on Japanese using the same training features. As seen the results from Table 9, exploiting the same linguistic features on different language dataset, Turkish outperformed other languages with overall accuracy 87%. However, the accuracy lowered down to 63% when applying the method to Japanese. Therefore, it gives an indication that we might need to explore more linguistic features to make this method more robust or we need to develop language-specific classification models.

As for the visual features, we notice that face image detection method can perform very well, but it also suffers many challenges. Inspired by the fact presented previously that if we just use first names to predict gender, we obtain 82% accuracy, however, it rises to reach 92% accuracy when we combine the the features from first names and face image. Thus, we could argue that the classification model could perform better while combining several types of features than just relying on purely visual features.

After observing the results from various behavioral features, it is surprising to find out that it seems to exist an upper bound on gender prediction accuracy since all papers lead to similar accuracy around 80% yet using different classification methods and various behavioral features. Of course, we could challenge this assumption by combining behavioral features and other types of features to see whether the upper limit always exists.

To sum up, if we want to develop a better gender classification model, we might need to explore more features and the combinations of different features.

Table 9. The accuracy of the SVM-based classifier on each of the language datasets

Language	Male	Female	Overall (F1)
French	79%	73%	76%
Indonesian	87%	80%	83%
Turkish	89%	85%	87%
Japanese(t)	50%	76%	63%
Japanese(u)	58%	68%	63%

References

1. Y. Zhong, N. J. Yuan, W. Zhong, F. Zhang, and X. Xie, “You are where you go: Inferring demographic attributes from location check-ins,” in *Proceedings of the eighth ACM international conference on web search and data mining.* ACM, 2015, pp. 295–304.
2. D. Duong, H. Tan, and S. Pham, “Customer gender prediction based on e-commerce data,” in *Knowledge and Systems Engineering (KSE), 2016 Eighth International Conference on.* IEEE, 2016, pp. 91–95.
3. E. Zheleva and L. Getoor, “To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles,” in *Proceedings of the 18th international conference on World wide web.* ACM, 2009, pp. 531–540.
4. T. M. Phuong *et al.*, “Gender prediction using browsing history,” in *Knowledge and Systems Engineering.* Springer, 2014, pp. 271–283.
5. M. Merler, L. Cao, and J. R. Smith, “You are what you tweet pic! gender prediction based on semantic analysis of social media images,” in *Multimedia and Expo (ICME), 2015 IEEE International Conference on.* IEEE, 2015, pp. 1–6.
6. J. Hu, H.-J. Zeng, H. Li, C. Niu, and Z. Chen, “Demographic prediction based on user’s browsing behavior,” in *Proceedings of the 16th international conference on World Wide Web.* ACM, 2007, pp. 151–160.
7. W. Deitrick, Z. Miller, B. Valyou, B. Dickinson, T. Munson, and W. Hu, “Author gender prediction in an email stream using neural networks,” *Journal of Intelligent Learning Systems and Applications*, vol. 4, no. 03, p. 169, 2012.
8. D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta, “Classifying latent user attributes in twitter,” in *Proceedings of the 2nd international workshop on Search and mining user-generated contents.* ACM, 2010, pp. 37–44.
9. S. C. Herring, L. A. Scheidt, S. Bonus, and E. Wright, “Bridging the gap: A genre analysis of weblogs,” in *System sciences, 2004. proceedings of the 37th annual Hawaii international conference on.* IEEE, 2004, pp. 11–pp.
10. X. Yan and L. Yan, “Gender classification of weblog authors.” in *AAAI spring symposium: computational approaches to analyzing weblogs.* Palo Alto, CA, 2006, pp. 228–230.
11. A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. N. Rosenquist, “Understanding the demographics of twitter users.” *ICWSM*, vol. 11, no. 5th, p. 25, 2011.
12. F. Karimi, C. Wagner, F. Lemmerich, M. Jadidi, and M. Strohmaier, “Inferring gender from names on the web: A comparative evaluation of gender detection methods,” in *Proceedings of the 25th International Conference Companion on World Wide Web.* International World Wide Web Conferences Steering Committee, 2016, pp. 53–54.
13. D. Gillick, “Can conversational word usage be used to predict speaker demographics?” in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.
14. G. Antipov, S.-A. Berrani, and J.-L. Dugelay, “Minimalistic cnn-based ensemble model for gender prediction from face images,” *Pattern recognition letters*, vol. 70, pp. 59–65, 2016.
15. Q. You, S. Bhatia, T. Sun, and J. Luo, “The eyes of the beholder: Gender prediction using images posted in online social networks,” in *Data Mining Workshop (ICDMW), 2014 IEEE International Conference on.* IEEE, 2014, pp. 1026–1030.

16. B. Bi, M. Shokouhi, M. Kosinski, and T. Graepel, “Inferring the demographics of search users: Social data meets search queries,” in *Proceedings of the 22nd international conference on World Wide Web*. ACM, 2013, pp. 131–140.
17. D. Jinyu, “Study on gender differences in language under the sociolinguistics,” *Canadian Social Science*, vol. 10, no. 3, pp. 92–96, 2014.
18. J. A. Suykens and J. Vandewalle, “Least squares support vector machine classifiers,” *Neural processing letters*, vol. 9, no. 3, pp. 293–300, 1999.
19. E. Vasilev, “Inferring gender of reddit users,” Master’s thesis, University of Koblenz and Landau, 2018.
20. P. Wang, J. Guo, Y. Lan, J. Xu, and X. Cheng, “Your cart tells you: Inferring demographic attributes from purchase data,” in *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. ACM, 2016, pp. 173–182.
21. S. Kabbur, E.-H. Han, and G. Karypis, “Content-based methods for predicting web-site demographic attributes,” in *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. IEEE, 2010, pp. 863–868.
22. M. Ciot, M. Sonderegger, and D. Ruths, “Gender inference of twitter users in non-english contexts,” in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 1136–1145.
23. S. Volkova and D. Yarowsky, “Improving gender prediction of social media users via weighted annotator rationales,” in *NIPS 2014 Workshop on Personalization*, 2014.