User Demographics and Language in an Implicit Social Network

Katja Filippova
Google Inc.
Brandschenkestr. 110
Zürich, 8004 Switzerland
katjaf@google.com

Abstract

We consider the task of predicting the gender of the YouTube[1] users and contrast two information sources: the comments they leave and the social environment induced from the affiliation graph of users and videos. We propagate gender information through the videos and show that a user's gender can be predicted from her social environment with the accuracy above 90%. We also show that the gender can be predicted from language alone (89%). A surprising result of our study is that the latter predictions correlate more strongly with the gender predominant in the user's environment than with the sex of the person as reported in the profile. We also investigate how the two views (linguistic and social) can be combined and analyse how prediction accuracy changes over different age groups.

1 Introduction

Over the past decade the web has become more and more social. The number of people having an identity on one of the Internet social networks (Facebook[2], Google+[3], Twitter[4], etc.) has been steadily growing, many users communicate online on a daily basis. Their interactions open new possibilities for social sciences, and linguistics is no exception. For example, with the development and growth of Web 2.0, it has become possible to get access to masses of text data labeled with respect to different social parameters such as country, age, gender, profession or religion. The study of language varieties between groups separated by a certain social variable belongs to the field of sociolinguistics which more generally investigates the effect of society on how language is used (Coulmas, 1998). Historically, sociolinguistics is connected to dialectology whose focus has been primarily on the phonetic aspect of the regional dialects but was later extended to sociolects (Chambers & Trudgill, 1998). A usual study would involve sampling speakers from a population, interviewing them and analyzing the linguistic items with respect to social variables (Hudson, 1980).

The last decade has seen several studies investigating the relationship between the language and the demographics of the users of blogs or Twitter (see Sec. 2 for references). Most of those studies used social network sites to collect labeled data– samples of text together with the demographics variable. However, they did not analyse how social environment affects language, although very similar questions have been recently posed (but not yet answered) by Ellist (2009). In our work we attempt to address precisely this issue. In particular, we consider the task of user gender prediction on YouTube and contrast two information sources: (1) the comments written by the user and (2) her social neighborhood as defined by the bipartite user-video graph. We use the comments to train a gender classifier on a variety of linguistic features. We also introduce a simple gender propagation procedure to predict person's gender from the user-video graph.

In what follows we will argue that although language does provide us with signals indicative of the

[1] www.youtube.com

[2] www.facebook.com

[3] www.plus.google.com

[4] www.twitter.com

Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 1478–1488, Jeju Island, Korea, 12–14 July 2012. ©2012 Association for Computational Linguistics

user's gender[5] (as reported in the user's profile), it is in fact more indicative of a socially defined gender. Leaving aside the debate on the intricate relationship between language and gender (see Eckert & McConnell-Ginet (2003) for a thorough discussion of the subject), we simply demonstrate that a classifier trained to predict the predominant gender in the user's social environment, as approximated by the YouTube graph of users and videos, achieves higher accuracy for both genders than the one trained to predict the user's inborn gender. We also investigate ways of how the language-based and the social views can be combined to improve prediction accuracy. Finally, we look at three age groups – teenagers, people in their twenties and people over thirty – and show that gender identity is more evident in the language of younger people but also that there is a higher correlation between their inborn gender and the predominant gender in their social environment.

The paper is organized as follows: we first review related work on the language of social media and user demographics (Sec. 2) and elaborate on the goals of our research (Sec. 3). Then we describe our data (Sec. 4), introduce the demographics propagation experiments (Sec. 5) and the experiments on supervised learning gender from language (Sec. 6).

2 Related work

Previous studies on language and demographics which looked at online data can be distinguished with respect to their aims. (1) Studies coming from the sociolinguistic community aim at empirically confirming hypotheses, such as that female speakers use more pronouns, or that males tend to use longer words. (2) A standard goal of an NLP study is to build an automatic system which accurately solves a given task which in the case of demographics is predicting user age, gender or country of origin. In this section we start by reviewing the first kind of studies, which are about data analysis and hypotheses checking. These are relevant for our choice of features. Then we briefly summarize a selection of studies on demographics prediction to better situate and motivate our approach.

2.1 Language and demographics analysis

Previous sociolinguistic studies mostly checked hypotheses formulated before the widespread use of the Internet, such as that women use hedges more often (Lakoff, 1973) or that men use more negations (Mulac et al., 2000), or looked at specific words or word classes. Newman et al. (2008) provide a comprehensive review of such work and a description of the non-web corpora used therein. Some of those hypotheses were confirmed by empirical evidence, some not.

For example, Herring & Paolillo (2006) analyse gender- and genre-specific use of language in online communication on a sample of about 130 blog entries. Looking at a number of stylistic features which had previously been claimed to be predictive of gender (Argamon et al., 2003; Koppel et al., 2004), such as personal pronouns, determiners and other function words, they find no gender effect. Unlike them, Kapidzic & Herring (2011) analyse recent chat communications and find that they are gendered. Similarly, Huffaker & Calvert (2005) investigate the question of identity of teenager bloggers (e.g., age, gender, sexuality) and find language features indicative of gender (e.g., use of emoticons by males). Burger & Henderson (2006) consider the relationship between different linguistic (e.g., text length, use of capital and punctuation letters) and non-linguistic (e.g., interests, mood) features and blogger's age and location. They find that many features correlate with the age and run an experiment with the goal of predicting whether the blog author is over 18.

2.2 Demographics prediction from language

The studies we review here used supervised machine learning to obtain models for predicting gender or age. Other demographic attributes, like location, ethnicity, or educational level, have also been predicted automatically (Gillick, 2010; Rao & Yarowsky, 2011, inter alia). Also, generative approaches have been applied to discover associations between language and demographics of social media users (Eisenstein et al., 2011, inter alia) but these are of less direct relevance for the present work. For su-

[5] Although it might be more correct to talk about the user's sex in place of gender (Eckert & McConnell-Ginet, 2003), we stick to the terminology adopted in previous NLP research on gender prediction.

pervised approaches, major feature sources are the text the user has written and also her profile which may list the name, interests, friends, etc. There have also been studies which did not look at the language at all but considered the social environment only. For example, MacKinnon & Warren (2006) aim at predicting the age and the location of the LiveJournal[6] users. What they found is that there is a remarkable correlation between the age and the location of the user and those of her friends, although there are interesting exceptions.

Burger et al. (2011) train a gender classifier on tweets with word and character-based ngram features achieving accuracy of 75.5%. Adding the full name feature alone gives a boost to 89.1%, further features like self-written description and screen name further help to get 92%. Also, a self-training method exploring unlabeled data is described but its performance is worse. Other kinds of sociolinguistic features and a different classifier have been applied to gender prediction on tweets by Rao & Yarowsky (2010).

Nowson & Oberlander (2006) achieve 92% accuracy on the gender prediction task using ngram features only. Their corpus consist of 1,400/450 posts written by 47 females and 24 males, respectively. However, the ngram features were preselected based on whether they occurred with significant relative frequency in the language of one gender over the other. Since the complete dataset was used to preselect features, the results are inconclusive.

Yan & Yan (2006) train a Naive Bayes classifier to predict the gender of a blog entry author. In total they looked at 75,000 individual blog entries authored by 3,000 bloggers, all of them posted their genders on the profile page. They measure precision and recall w.r.t. the minority class (males) and get the best f-measure of 0.64 (precision and recall are 65% and 71%, respectively).

Rosenthal & McKeown (2011) predict the age of a blogger, most features they use are extracted from the blog posts, other features include blogger's interests, the number of friends, the usual time of posting, etc. Similarly to Schler et al. (2006), they run a classification experiment with three age classes removing intermediate ages and use the majority-class baseline for comparison. In their other experiment they experiment with a binary classifier for age distinguishing between the pre- and post-social media generations and using the years from 1975-1988 as a boundary. The prediction accuracy increases as later years are taken.

Interestingly, it has been shown that demographics can be predicted in more restricted genres than the personal blog or tweets and from text fragments even shorter than tweets (Otterbacher, 2010; Popescu & Grefenstette, 2010).

3 Motivation for the present study

Similarly to previous NLP studies, our starting goal is to predict the self-reported user gender. The first novelty of our research is that in doing so we contrast two sources of information: the user's social environment and the text she has written. Indeed, a topic which has not yet been investigated much in the reviewed studies on language and user demographics is the relationship between the language of the user and her social environment. The data analysis studies (Sec. 2.1) verified hypotheses concerning the dependency between a language trait (e.g., average sentence length) and a demographic parameter (e.g., gender). The demographics prediction studies (Sec. 2.2) mostly relied on language and user profile features and considered users in isolation. An exception to this is Garera & Yarowsky (2009) who showed that, for gender prediction in a dialogue, it helps to know the interlocutor's gender. However, we aim at investigating the impact of the social environment in a much broader sense than the immediate interlocutors and in a much broader context than a conversation.

Language is a social phenomenon, and it is this fact that motivates all the sociolinguistic research. Many if not most language traits are not hard-wired or inborn but can be explained by looking at who the person interacts most with. Since every language speaker can be seen as a member of multiple overlapping communities (e.g., computer scientists, French, males, runners), the language of the person may reflect her membership in different communities to various degrees. Repeated interactions with other language speakers influence the way the person speaks (Baxter et al., 2006; Bybee, 2010), and

[6] `www.livejournal.com`

the influence is observable on all the levels of the language representation (Croft, 2000). For example, it has been shown that the more a person is integrated in a certain community and the tighter the ties of the social network are, the more prominent are the representative traits of that community in the language of the person (Milroy & Milroy, 1992; Labov, 1994). In our study we adopt a similar view and analyse the implications it has for gender prediction. Given its social nature, does the language reflect the norms of a community the user belongs to or the actual value of a demographic variable?

In our study we address this issue with a particular modeling technique: we assume that the observed online behavior adequately reflects the offline life of a user (more on this in Sec. 4 and 5) and based on this assumption make inferences about the user's social environment. We use language-based features and a supervised approach to gender prediction to analyse the relationship between the language and the variable to be predicted. To our knowledge, we are the first to question whether it is really the inborn gender that language-based classifiers learn to predict. More concrete questions we are going to suggest answers to are as follows:

1. Previous studies which looked at online data relied on self-reported demographics. The profile data are known to be noisy, although it is hard to estimate the proportion of false profiles (Burger et al., 2011). Concerning the prediction task, how can we make use of what we know about the user's social environment to reduce the effect of noise? How can we benefit from the language samples from the users whose gender we do not know at all?

2. When analyzing the language of a user, how much are its gender-specific traits due to the user's inborn gender and to which extent can they be explained by her social environment? Using our modeling technique and a language-based gender classifier, how is its performance affected by what we know about the online social environment of the user?

3. Concerning gender predictions across different age groups, how does classifier performance change? Judging from the online communication, do teenagers signal their gender identity more than older people? In terms of classifier accuracy, is it easier to predict a teenager's gender than the gender of an adult?

The final novelty of our study is that we are the first to demonstrate how YouTube can be used as a valuable resource for sociolinguistic research. In the following section we highlight the points which make YouTube interesting and unique.

4 Data

Most social networks strive to protect user privacy and by default do not expose profile information or reveal user activity (e.g., posts, comments, votes, etc.). To obtain data for our experiments we use YouTube, a video sharing site. Most of the YouTube registered users list their gender, age and location on their profile pages which, like their comments, are publicly available. YouTube is an interesting domain for sociolinguistic research for several reasons:

High diversity: it is not restricted to any particular topic (e.g., like political blogs) but covers a vast variety of topics attracting a very broad audience, from children interested in cartoons to academics watching lectures on philosophy[7].

Spontaneous speech: the user comments are arguably more spontaneous than blogs which are more likely to conform to the norms of written language. At the same time they are less restricted than tweets written under the length constraint which encourages highly compressed utterances.

Data availability: all the comments are publicly available, so we have do not get a biased subset of what a user has written for the public. Moreover, we observe users' interactions in different environments because every video targets particular groups of people who may share origin (e.g., *elections in Greece*) or possession (e.g., *how to unlock iPhone*) or any other property. Some videos attract a well-defined group of people (e.g., *the family of a newborn child*), whereas some videos appeal to a very broad audience (e.g., *a kitten video*).

[7] For more information and statistics see the official YouTube demographics on `http://www.youtube.com/yt/advertise/affinities.html`.

female	male	nn
26%	62%	12%

Table 1: Gender distribution for the extracted 6.9M users.

From the users, videos and the comment relationship we build an affiliation graph (Easley & Kleinberg, 2010): a user and a video are connected if the user commented on the video (Fig. 1(a)). Our graph is unweighted although the number of comments could be used to weight edges. The co-comment graph is a stricter version of a more popular co-view graph used in, e.g., video recommendation studies (Baluja et al., 2008, inter alia).

We obtained a random sample of videos by considering all the videos whose YouTube ID has a specific prefix[8]. From those, we collected the profiles of the users whose commented on the videos. In total, we extracted about 6.9M profiles of users who have written at least 20 comments, not more than 30 comments were collected for every user. The threshold on the minimum number of comments is set in order to reduce the proportion of users who have used YouTube only a few times and possibly followed the suggestions of the site in their video choice. The users' gender distributions is presented in Table 1. Although females, in particular teenagers, have been reported to be more likely to blog than males (Herring et al., 2004), males are predominant in our dataset. A random sample from a pool of users without the 20-comments threshold showed that there are more male commenters overall, although the difference is less remarkable for teenagers: 58% of the teenagers with known gender are male as opposed to 74% and 79% for the age groups 20-29 and 30+. Teenagers are also more numerous accounting for about 35% in our data.

Although we did not filter users based on their location or mother tongue as many users comment in multiple languages, the comment set is overwhelmingly English.

5 Gender propagation

We first consider the user's social environment to see whether there is any correlation between the gender of a user and the gender distribution in her vicinity, independent of the language. We use a simple propagation procedure to reach the closest neighbors of a user, that is, other users "affiliated" with the same videos. Specifically, we perform the following two steps:

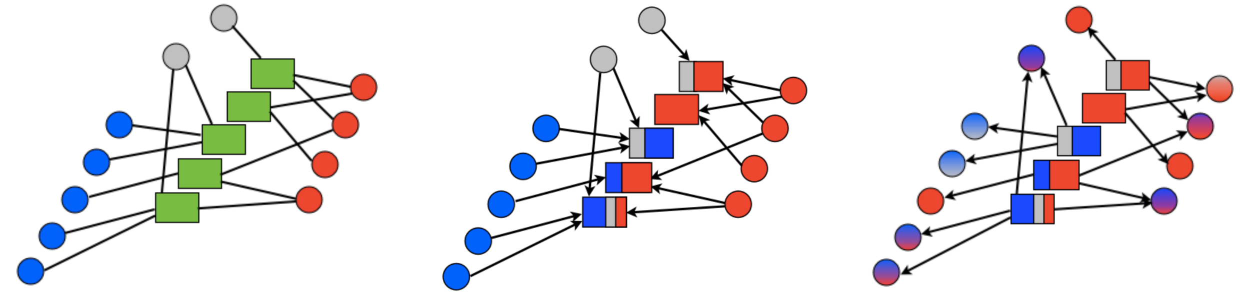
1. We send the gender information (female, male or unknown) to all the videos the user has commented on. This way for every video we obtain a multinomial distribution over three classes (see Fig. 1(b)).

2. We send the gender distributions from every video back to all the users who commented on it and average over all the videos the user is connected with (see Fig. 1(c)). However, in doing so we adjust the distribution for every user so that her own demographics is excluded. This way we have a fair setting where the original gender of the user is never included in what she gets back from the connected videos. Thus, the gender of a user contributes to the vicinity distributions of all the neighbors but not to her own final gender distribution.

In line with our motivation and modeling technique, we chose such a simple method (and not, say, classification) in order to approximate the offline encounters of the user: does she more often meet women or men? The way we think of the videos is that they correspond to places (e.g., a cinema, a cosmetic shop, a pub) visited by the user where she is unintentionally or deliberately exposed to how other speakers use the language. Similar to Baxter et al. (2006), we assume that these encounters influence the way the person speaks. Note that if the user's gender has no influence on her choice of videos, then, on average, we would expect every video to have the same distribution as in our data overall: 62% male, 26% female and 12% unknown (Table 1).

To obtain a single gender prediction from the propagated distribution, for a given user we select the gender class (female or male) which got more

[8] The YouTube API (`http://code.google.com/apis/youtube/getting_started.html`) can be used to retrieve user profiles and video metadata as well as the comments.



(a) Color represents gender information: *blue*=male, *red*=female, *grey*=unknown.

(b) Propagating gender from users to videos.

(c) Propagating gender distribution from videos to users.

Figure 1: Affiliation graph of users (circles) and videos (rectangles).

of the distribution mass. The exact procedure is as follows: given user *u* connected with videos $V_u = \{v_1, ..., v_m\}$, there are *m* gender distributions sent to *u*: $P_V(u) = \{p(g|v_i) : 1 \leq i \leq m, g \in \{f, m, n\}\}$. A single distribution is obtained from $P_V(u)$: $\hat{p}(g|u) = \sum_i p(g|v_i)/m$.

To address the skewness in the data, i.e., the fact that 70% of our users $(62/(26 + 62))$ with known gender are male, we select the female gender if (a) it got more than zero mass and at least as much mass as male: $\hat{p}(f) > 0 \wedge \hat{p}(f) \geq \hat{p}(m)$, or (b) it got at least τ of the mass: $\hat{p}(f) \geq \tau$. We set $\tau = 0.26$ initially because it corresponds to the expected proportion of females (26%) but further experimented with different τ values in the range of 0.25-0.4. We obtained best accuracy and f-measures with the threshold of 0.33, the difference in accuracy from the initial threshold of 0.26 being less than 2%. The fact that the optimal τ value is different from the overall proportion of females (26%) is not surprising given that we aggregate per video distributions and not raw user counts.

The predictions obtained with the described propagation method are remarkably accurate, reaching 90% accuracy (Table 2). The baseline of assigning all the users the majority class (*all male*) provides us with the accuracy of 70% – the proportion of males among the users with known gender.

Although the purpose of this section is not to present a gender prediction method, we find it worth emphasizing that 90% accuracy is remarkable given that we only look at the immediate user vicinity. In the following section we are going to investigate how this social view on demographics can help us in predicting gender from language.

	Acc%	P%	R%	F1
Baseline	70	-	-	-
all	90	-	-	-
fem	-	84.3	80.8	83
male	-	92.2	93.8	93

Table 2: Precision and recall for propagated gender.

6 Supervised learning of gender

In this section we start by describing our first gender prediction experiment and several extensions to it and then turn to the results.

6.1 Experiments

Similar to previous studies on demographics prediction, we start with a supervised approach and only look at the text (comments) written by the user. We do not rely on any information from the social environment of the user and do not use any features extracted from the user profile, like name, which would make the gender prediction task considerably easier (Burger et al., 2011). Finally, we do not extract any features from the videos the user has commented on because our goal here is to explore the language as a sole source of information. Here we simply want to investigate the extent to which the language of the user is indicative of her gender which is found in the profile and which, ignoring the noise, corresponds to the inborn gender.

In our experiments we use a distributed implementation of the maximum entropy learner (Berger et al., 1996; McDonald et al., 2010) which outputs a distribution over the classes, the final prediction is the class with the greater probability. We take 80% of the users for training and generate a training instance for every user who made her gender visible on the profile page (4.9M). The remaining 20% of the data are used for testing (1.2M). We use the following three groups of features: (1) *character-based:* average comment length, ratio of capital letters to the total number of letters, ratio of punctuation to the total number of characters; (2) *token-based:* average comment length in words, ratio of unique words to the total tokens, lowercase unigrams with total count over all the comments (10K most frequent unigrams were used, the frequencies were computed on a separate comment set), use of pronouns, determiners, function words; (3) *sentence-based:* average comment length in sentences, average sentence length in words.

Enhancing the training set. The first question we consider is how the affiliation graph and propagated gender can be used to enhance our data for the supervised experiments. One possibility would be to train a classifier on a refined set of users by eliminating all those whose reported gender did not match the gender predicted by the neighborhood. This would presumably reduce the amount of noise by discarding the users who intentionally provided false information on their profiles. Another possibility would be to extend the training set with the users who did not make their gender visible to the public but whose gender we can predict from their vicinity. The idea here is similar to co-training where one has two independent views on the data. In this case a social graph view would be combined with the language-based view.

Profile vs. vicinity gender prediction. The next question posed in the motivation section is as follows: Does the fact that language is a social phenomenon and that it is being shaped by the social environment of the speaker impact our gender classifier? If there are truly gender-specific language traits and they are reflected in our features, then we should not observe any significant difference between the prediction results on the users whose gender matches the gender propagated from the vicinity and those whose gender does not match. A contrary hypothesis would be that what the classifier actually learns to predict is not as much the inborn but a social gender. In this case, the classifier trained on the propagated gender labels should be more accurate than the one trained on the labels extracted from the profiles.

To address these questions we contrast two classifiers: (1) the one described in the beginning of the section which is trained on the gender labels collected from the user profiles; (2) a classifier trained on the vicinity gender, that is the dominating gender of the environment of a speaker as obtained with the procedure described in Section 5.

Age groups and gender prediction. Finally, we look at how gender predictions change with age and train three age-specific models to predict gender for teenagers (*13-19*), people in their twenties (*20-29*) and people over thirty (*30+*), the age is also extracted from the profiles. These groups are identified in order to check whether teenagers tend to signalize their gender identity more than older people, a hypothesis investigated earlier on a sample of blog posts (Huffaker & Calvert, 2005).

6.2 Results

We report the results of the supervised experiments for all the settings described above. As an estimate of the lowest bound we also give the results of the majority class baseline (*all male*) which guarantees 70% accuracy. For the supervised classifiers we report accuracy and per-gender precision, recall and f-measure. Table 3 presents the results for the starting classifier trained to predict profile gender.

	Acc%	P%	R%	F1	Total
Baseline	70	-	-	-	619K
all	89	-	-	-	619K
fem	-	83	78	80	182K
male	-	91	94	93	437K

Table 3: Results on the test set.

In order to investigate the relationship between the social environment of a person, her gender and the language, we split the users from the test set into two groups: those whose profile gender matched the gender propagated from the vicinity and those for whom there was a mismatch. Thus Table 4 presents the same results as Table 3 but separated for these two groups of users. It also gives user counts w.r.t. the profile gender.

	Acc%	P%	R%	F1	Total
all (same)	94	-	-	-	557K
fem (same)	-	89	87	88	147K
male (same)	-	95	96	96	410K
all (diff)	47	-	-	-	62K
fem (diff)	-	54	39	45	35K
male (diff)	-	42	56	48	27K

Table 4: Results for users whose profile gender matches/differs from the vicinity gender.

Enhanced training set. In the next experiment we refined the training set by removing all the users whose vicinity gender did not match the gender reported in the profile. The evaluation was done on the unmodified set (Table 5). The predictions made by the model trained on a refined set of users turned out to be slightly less accurate than those made by the model trained on the full training set (Table 3). The refined model performed slightly ($< 1\%$) better than the starting one on the users whose vicinity and the profile genders matched but got very poor results on the users with a gender mismatch, the accuracy being as low as 37%. The accuracy of the starting model on those users is 47% (Table 4).

	Acc%	P%	R%	F1
all	88	-	-	-
fem	-	83	76	79
male	-	90	94	92

Table 5: Results of the models trained on the refined training set.

In another experiment we extended the training data with the users whose gender was unknown but was predicted with the propagation method. However, a larger training set makes a difference only if there is a substantial performance gain over the increasing size of the training set. We observed only a minor gain in performance ($< 1\%$) when the training data size was increased by an order of magnitude. Given that, it is not surprising that adding 12% did not affect the results.

Language, the vicinity and the profile genders. The gap in accuracies of predictions for the two user groups in Table 4 is remarkable: 47% vs. 94%. If we extrapolate what we observe in the affiliation graph to other online and offline life, then this result may suggest that gender traits are more prominent in the language of people spending more time with the people of their gender than in that of the people who spend more time with the people of the opposite gender. Given the remarkable difference, a further question arises whether the classifier actually learns to predict a kind of socially rather than the profile gender. To investigate this, we looked at the results of the model which knew nothing about the profile gender but was trained to predict the vicinity gender instead (Table 6). This model relied on the exact same set of features but both for training and testing it used the gender labels obtained from the propagation procedure described in Section 5.

	Acc%	P%	R%	F1
all	91	-	-	-
fem	-	86	80	83
male	-	92	95	94

Table 6: Results of the models trained and tested on the propagated gender.

According to all the evaluation metrics, for both genders the performance of the classifier trained and tested on the propagated gender is higher (cf. Table 3): the differences in f-measure for female and male are four and two points respectively, both statistically significant. This indicates that it is the predominant *environment* gender that a language-based classifier is better at learning rather than the inborn gender.

Predictions across age groups. Finally, to address the question of whether gender differences are more prominent and thus easer to identify in the language of younger people, we looked at the accuracy of gender predictions across three age groups. Table 7 summarizes the results and gives the accuracy of the all male baseline as well as of the propagation procedure (*Prop-acc*). Although the overall accuracy over the three groups does not degrade much, from 89% to 87%, both precision and recall do decrease significantly for females. This is not directly reflected in the accuracy because the number of females drops dramatically from 42% among teenagers to 26% and then 21% in the latter groups. For a comparison, the accuracy of the propagated gender (*Prop-acc*) also decreases from younger to older age groups although it is slightly higher than that of language-based predictions. One conclusion we can make at this point is that a teenager's gender is easier to predict from the language which is in line with the hypothesis that younger people signalize their gender identities more than older people. Another observation is that, as the person gets older, we can be less sure about her gender by looking at her social environment. This in turn might be an explanation of why there are less gender signals in the language of a person: the environment becomes more mixed, and the influence of both genders becomes more balanced.

	13-19	20-29	30+	Overall
Base-acc%	58	74	79	70
Prop-acc%	91	90	88	90
Accuracy%	89	89	87	89
Fem-P%	87	81	74	83
Fem-R%	87	76	62	78
Fem-F1	87	78	68	80
Male-P%	90	92	90	91
Male-R%	90	94	94	94
Male-F1	90	93	92	93

Table 7: Results across the age groups.

7 Conclusions

In our study we addressed the gender prediction task from two perspectives: (1) the social one where we looked at an affiliation graph of users and videos and propagated gender information between users, and (2) the language one where we trained a classifier on features which have been claimed to be indicative of gender. We demonstrated that both perspectives provide us with comparably accurate predictions (around 90%) but that they are far from being independent. We also investigated a few ways of how the performance of a language-based classifier can be enhanced by the social aspect, compared the accuracy of predictions across different age groups and found support for hypotheses made in earlier sociolinguistic studies.

We are not the first to predict gender from language features with online data. However, to our knowledge, we are the first to contrast the two views, social and language-based, using online data and to question whether there is a clear understanding of what gender classifiers actually learn to predict from language. Our results indicate that from the standard language cues we are better at predicting a social gender, that is the gender defined by the environment of a person, rather than the inborn gender.

The theoretical significance of this result is that it provides support for the usage-based view on language (Bybee, 2010), namely that the person's language is largely shaped by the interactions with her social environment. On the practical side, it may have implications for targeted advertisement as it enriches the understanding of what gender classifiers predict.

Acknowledgements: I am thankful to Enrique Alfonseca, Keith Hall and the EMNLP reviewers for their feedback.

References

Argamon, S., M. Koppel, J. Fine & A. R. Shimoni (2003). Gender, genre, and writing style in formal written texts. *Text*, 23(3).

Baluja, S., R. Seth, D. Sivakumar, Y. Jing, J. Yagnik, S. Kumar, D. Ravichandran & M. Aly (2008). Video suggestion and discovery for YouTube: Taking random walks through the view graph. In *Proc. of WWW-08*, pp. 895–904.

Baxter, G. J., R. A. Blythe, W. Croft & A. J. McKane (2006). Utterance selection model of language change. *Physical Review*, E73.046118.

Berger, A., S. A. Della Pietra & V. J. Della Pietra (1996). A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Burger, J. D., J. Henderson, G. Kim & G. Zarrella (2011). Discriminating gender on Twitter. In *Proc. of EMNLP-11*, pp. 1301–1309.

Burger, J. D. & J. C. Henderson (2006). An exploration of observable features related to blogger age. In *Proceedings of the AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs,* Stanford, CA, 27-29 March 2006, pp. 15–20.

Bybee, J. (2010). *Language, Usage and Cognition.* Cambridge University Press.

Chambers, J. & P. Trudgill (1998). *Dialectology.* Cambridge University Press.

Coulmas, F. (Ed.) (1998). *The Handbook of Sociolinguistics.* Blackwell.

Croft, W. (2000). *Explaining Language Change: An Evolutionary Approach.* London: Longman.

Easley, D. & J. Kleinberg (2010). *Network, crowds, and markets: Reasoning about a highly connected world.* Cambridge University Press.

Eckert, P. & S. McConnell-Ginet (2003). *Language and Gender.* Cambridge University Press.

Eisenstein, J., N. A. Smith & E. P. Xing (2011). Discovering sociolinguistic associations with structured sparsity. In *Proc. of ACL-11*, pp. 1365–1374.

Ellist, D. (2009). Social (distributed) language modeling, clustering and dialectometry. In *Proc. of TextGraphs at ACL-IJCNLP-09*, pp. 1–4.

Garera, N. & D. Yarowsky (2009). Modeling latent biographic attributes in conversational genres. In *Proc. of ACL-IJCNLP-09*, pp. 710–718.

Gillick, D. (2010). Can conversational word usage be used to predict speaker demographics? In *Proceedings of Interspeech,* Makuhari, Japan, 26-30 September 2010.

Herring, S. C. & J. C. Paolillo (2006). Gender and genre variation in weblogs. *Journal of Sociolinguistics*, 10(4):439–459.

Herring, S. C., L. A. Scheidt, S. Bonus & E. Wright (2004). Bridging the gap: A genre analysis of weblogs. In *HICSS-04*.

Hudson, R. A. (1980). *Sociolinguistics.* Cambridge University Press.

Huffaker, D. A. & S. L. Calvert (2005). Gender, identity and language use in teenager blogs. *Journal of Computer-Mediated Communication*, 10(2).

Kapidzic, S. & S. C. Herring (2011). Gender, communication, and self-presentation in teen chatrooms revisited: Have patterns changed? *Journal of Computer-Mediated Communication*, 17(1):39–59.

Koppel, M., S. Argamon & A. R. Shimoni (2004). Automatically categorizing written text by author gender. *Literary and Linguistic Computing*, 17(4):401–412.

Labov, W. (1994). *Principles of Linguistic Change: Internal Factors.* Blackwell.

Lakoff, R. (1973). Language and woman's place. *Language in Society*, 2(1):45–80.

MacKinnon, I. & R. Warren (2006). Age and geographic inferences of the LiveJournal social network. In *Statistical Network Analysis: Models, Issues, and New Directions Workshop at ICML-2006,* Pittsburgh, PA, 29 June, 2006.

McDonald, R., K. Hall & G. Mann (2010). Distributed training strategies for the structured perceptron. In *Proc. of NAACL-HLT-10*, pp. 456–464.

Milroy, L. & J. Milroy (1992). Social network and social class: Toward an integrated sociolinguistic model. *Language in Society*, 21:1–26.

Mulac, A., D. R. Seibold & J. R. Farris (2000). Female and male managers' and professionals' criticism giving: Differences in language use and effects. *Journal of Language and Social Psychology*, 19(4):389–415.

Newman, M. L., C. J. Groom, L. D. Handelman & J. W. Pennebaker (2008). Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, 45:211–236.

Nowson, S. & J. Oberlander (2006). The identity of bloggers: Openness and gender in personal weblogs. In *Proceedings of the AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs,* Stanford, CA, 27-29 March 2006, pp. 163–167.

Otterbacher, J. (2010). Inferring gender of movie reviewers: Exploiting writing style, content and metadata. In *Proc. of CIKM-10*.

Popescu, A. & G. Grefenstette (2010). Mining user home location and gender from Flickr tags. In *Proc. of ICWSM-10*, pp. 1873–1876.

Rao, D. & D. Yarowsky (2010). Detecting latent user properties in social media. In *Proc. of the NIPS MLSN Workshop*.

Rao, D. & D. Yarowsky (2011). Typed graph models for semi-supervised learning of name ethnicity. In *Proc. of ACL-11*, pp. 514–518.

Rosenthal, S. & K. McKeown (2011). Age prediction in blogs: A study of style, content, and online behavior in pre- and post-social media generations. In *Proc. of ACL-11*, pp. 763–772.

Schler, J., M. Koppel, S. Argamon & J. Pennebaker (2006). Effects of age and gender on blogging. In *Proceedings of the AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs,* Stanford, CA, 27-29 March 2006, pp. 199–205.

Yan, X. & L. Yan (2006). Gender classification of weblogs authors. In *Proceedings of the AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs,* Stanford, CA, 27-29 March 2006, pp. 228–230.