Suresh Gyan Vihar University, Jaipur
International Journal of Converging Technologies and Management (IJCTM)
Volume 1, Issue 1, 2015
ISSN : 2455-7528

User Demographics and Language in an Implicit Social Network

Deepika Kumari
Information and Technology
Suresh Gyan Vihar University
Jaipur, India
deepika07.Kilania@gmail.com

***Abstract*- we** regard as the task of predict the gender of the YouTub1 users and distinction two information sources: the comments they leave and the social environment induce from the connection graph of users and videos. We spread gender information all the way through the videos and show that a user's gender can be predicted from her social environment with the precision 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 link more strongly with the gender main in the user's environment than with the sex of the person as reported in the profile. We also examine how the two views (linguistic and social) can be combined and analyse how forecast accurateness change.

Keyword:-*Social Network , Language, demographics .*

I. 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 (Facebook2, Google+3, Twitter4, etc.) has been steadily growing, many users communicate online on a daily basis we simply demonstrate that a classifier trained to predict the predominant gender in the user's social network, as approximated by the YouTube graph of users and videos, achieves higher accuracy for both genders than one trained to predict the user's inborn gender they also investigate ways of how the language-based and the social views can be mutual to improve prophecy truthfulness Finally, we look at three age groups teenagers, people in their twenties and people over thirty – and show that gender characteristics is more marked 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 these paper is organized as follows: we first evaluation allied work on the language of social media and user demographics (Sec. 2) and detailed on the goals of our research (Sec. 3). Then we depict our data (Sec. 4), introduce the demographics dissemination experiments (Sec. 5) and the experiments on supervise culture gender from language (Sec. 6).

II.RELATED WORK

Prior study on tongue and demographics which look date online data can be notable with respect to their aims.(1) Studies coming from the sociolinguistic the people aim empirically confirm hypothesis, such as that female speaker Use more pronouns, or that males tend to use longerwords. (2) A standard goal of an NLP study is to build an repeated system which truthfully solves a given task which in the case of demographics is predicting user age, gender or country of foundation. In this section we start by review the first kind of studies, which are about data analysis and hypothesis checking. These are appropriate for our choice of features. Then we briefly recap a selection of studies on demographics prediction to better position.

2.1 Language and demographics analysis:

Previous sociolinguistic studies mostly confirm edhypotheses formulate before the pervasive use of the Internet, such as that women use hedge more habitually (Lakoff, 1973) or that men use more negation (Mulec et al., 2000), or looked at specific words or word classes. Newman et al. (2008) provide a wide-ranging review of such work and a narrative of the non-web corpora used therein. Some of those hypothesis were deep-rooted by observed substantiation, some not. For example, Herring & Paolillo (2006) analyse masculinity- and genre-specific use of language in online communication on a sample of about 130 blogentries. Looking at a number of stylistic features

Which had previously been claim to be farsighted 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 interactions and find that they are gendered. Similarly, Huffaker & Calvert (2005) inspect the question of identity of youngster bloggers (e.g., age, gender, sexuality) and find language features investigative 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 draw a parallel with the age and run an experiment with the goal of predict whether the blog author is over 18.

2.2 Demographics prediction from language

The studies we review here used supervise contraption learning to obtain models for predicting gender or age. Other demographic attributes, like location, traditions, 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 subaseline 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).

III Motivation for the present study

The same to previous NLP studies, our early goal is to predict the self-reported user gender. The first brightness of our research is that in doing so we discrepancy two source of information: the user's social environment and the text she has printed. indisputably, 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 atmosphere. The data analysis studies (Sec. 2.1) verified hypotheses as regards the reliance between a language attribute (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 loneliness. An exception to this is Garera & Yarowsky (2009) who showed that, for gender prediction in a discourse, 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 abrupt interlocutors and in a much broader context than a dialogue. Language is a social happening, and it is this fact that motivate all the sociolinguistic research. Many if not most language individuality are not hard-wired or intuitive but can be explained by looking at who the person interact most with. Since every language narrator can be seen as a member of multiple overlapping communities (e.g., computer scientists,

French, males, runners), the language of the person may echo her attachment in different community to various degrees. Repeated exchanges with other language speakers authority the way the person speaks (Baxter et al., 2006; Bybee, 2010), and he control is observable on all the levels of the talking representation (Croft, 2000). For example, it has been shown that the more a person is integrated in certain community and the tighter the tie of the social network are, the more high-flying are the commissioner qualities of that community in the language of the person (Milroy & Milroy, 1992; Labov, 1994). In our study we approve a similar view and analyse the implication it has for gender calculation. Given its social nature, does the language return the norms of a neighborhood the user belongs to or the actual value of a demographic erratic? In our study we address this issue with a particular rmodeling technique: we assume that the observed online behavior satisfactorily reflect the offline life of a user (more on this in Sec. 4 and 5) and based on this postulation 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 inherited 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. Linking to gender predictions crosswise different age groups, how does classifier performance change? Judging from the online communication, do adolescents signal their gender identity more than older people? In terms of classifier precision, is it easier to predict a teenager's gender than the gender of an immature? The final freshness of our study is that we are the first to exhibit how YouTube can be used as a valuable resource for sociolinguistic research. In the following section we underscore the points which make YouTube interesting and distinctive. IV.Data

Most social networks endeavor to care for user privacy and by default do not interpretation profile information or expose 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 remarks, are widely available. YouTube is an interesting domain for sociolinguistic research for a number of reasons

High diversity: it is not classified 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 cartoon strips to academics watching lectures on idea.

Spontaneous speech: the user comments are arguably more spur-of-the-moment than blogs which are more likely to obey the rules to the norms of written language. At the same time they are less top secret than tweets written under the length constriction which encourages highly compacted 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).

V.Gender propagation

We first think about the user's social environment to Se whether there is any parallel between the gender of a user and the gender allotment in her vicinity, independent of the language. We use a simple dissemination course of action 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 fling the gender in sequence (female, male or unknown) to all the videos the user has comment on. This way for every video we attain a multinomial allotment over three classes.

2. We send the gender distributions from each video back to all the users who commented on it and typical over all the videos the user is connected with However, in doing so we fine-tune the distribution for every user so that her own demographics is disqualified. The influence we have a fair location the original gender of the user is never included in what she gets back from the joined videos. Thus, the gender of a user contributes to the locality 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 inadvertently or purposely showing to how other speakers use the language. Similar to Baxter et al. (2006), we assume that these encounters authority the way the person speaks. Note that if the user's gender has no control on her choice of videos, then, on average, we would look forward to every video to have the same delivery 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 of the distribution mass.

VI .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 supervise draw near 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 extract from the user profile, like name, which would make the gender prediction task noticeably easier (Burger et al., 2011). Finally, we do not pull out any features from the videos the user has commented on because our goal here is to see the sights the language as a individual source of information. Here we simply want to investigate the level to which the language of the user is investigative of her gender which is found in the profile and which, ignoring the noise, correspond 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 instruction and create a training occurrence for every user who made her gender perceptible on the profile page (4.9M). The enduring 20% of the data are used for testing (1.2M). We use the following three groups of features: (1) character-based: familiar comment length, ratio of capital letters to the total number of letters, ratio of punctuation to the total number of characters; (2) tokenbased: average comment length in words, ratio of extraordinary 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.

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 greater part class baseline (all male) which guarantees70% exactness. For the supervised classifiers we report accuracy and per-gender precision, recall and measure. Table 3 presents the results for the starting classifier qualified to predict profile gender. In order to investigate the relationship between the social environment of a person, her gender and the language, we crack the users from the test set into two groups: those whose profile gender matched the gender propagate from the environs and those for whom there was a variance.

VII .Conclusions

In our study we addressed the gender calculation task from two perspectives: (1) the social one where we looked at a membership graph of users and videos and propagated gender information between users, and (2) the language one where we qualified a classifier on features which have been claim to be investigative of gender. We confirmed 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 superior by the social part, compare the accuracy of predictions transversely different age groups.

VII. References

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

[2] 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.

[3] 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. [4] Burger, J. D., J. Henderson, G. Kim & G. Zarrella (2011). Discriminating gender on Twitter. In Proc. of EMNLP-11, pp. 1301–1309.

[5] 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.

[6] Bybee, J. (2010). Language, Usage and Cognition.

Mulac, A., D. R. Seibold & J. R. Farris (2000). Female and male managers' and professionals' criticism giving:

Differences in language use and ef