Inferring Gender from the Content of Tweets: A Region Specific Example

Clay Fink,[a] Jonathon Kopecky,[a] Maksym Morawski[b]

[a]The Johns Hopkins University Applied Physics Laboratory, Laurel Maryland 20723

[b]The University of Maryland Baltimore County, Catonsville, Maryland 21250

clayton fink@jhuapl.edu

Abstract

There is growing interest in using social networking sites such as Twitter to gather real-time data on the reactions and opinions of a region's population, including locations in the developing world where social media has played an important role in recent events, such as the 2011 Arab Spring. However, many interesting and important opinions and reactions may differ significantly within a given region depending on the demographics of the subpopulation, including such categories as gender and ethnicity. This information may not be explicitly available in user content or metadata, however, and automated methods are required to infer such hidden attributes. In this paper we describe a method to infer the gender of Twitter users from only the content of their tweets. Looking at Twitter users from the West African nation of Nigeria, we applied supervised machine learning using features derived from the content of user tweets to train a classifier. Using unigram features alone, we obtained an accuracy of 80% for predicting gender, suggesting that content alone can be a good predictor of gender. An analysis of the highest weighted features shows some interesting distinctions between men and women both topically and emotionally. We argue that approaches such as the one described here can give us a clearer picture of who is utilizing social media when certain user attributes are unreliable or not available.

Introduction

Social media sites, such as Twitter, allow anyone with access to the Internet - whether from a desktop, laptop, or mobile device - to not only connect with friends, family, and colleagues but also share his or her opinions about the world and their reactions to events. There are numerous potential consumers who could utilize this public user-generated text, including market researchers, public opinion analysts, and epidemiologists, among others.

The population represented online via social media in a particular region may not be representative of the overall population, however, in terms of demography. One obvious example is that users of social media may be younger on average than the whole population in a region. Indeed, statistics available for Facebook usage illustrate this. For example, Nigerian Facebook users are predominately young and male, with over 70% of users between the ages of 18 and 35, and men outnumbering women two-to-one ("Nigeria Facebook Statistics," 2012). The demographic biases in these data have implications for inferences made from social media content and must be characterized and understood when making inferences from social media data about, for example, public sentiment or reactions to events. In some cases, reliable demographic information such as gender, ethnicity, or age is available for social media users. In many cases, however, these data are not present, unavailable, or unreliable. Therefore, techniques are needed to accurately infer such user attributes from content or metadata so that online populations can be accurately characterized.

Twitter profiles, unlike those for Facebook and Google+, do not provide a field for a person's gender, making gender a hidden attribute that must be inferred in some way. In most languages, a person's given name tends to be a good indicator of gender. On sites such as Facebook and Google+, it is expected that users have given their actual names and are who they say they are. Twitter, however, has different social norms; people can use any name they want, whether it is their actual name, a nickname, no name at all, or an alias. Because name data may be unreliable on Twitter, inferring gender from other available user information is necessary. In this paper we describe a technique to predict the gender of Twitter users from tweet content alone, providing the ability to automatically determine the distribution of this significant demographic attribute for a population of users.

We focus, in particular, on Twitter users in the West African country of Nigeria. Nigeria, the most populous country in Africa, has experienced rapid growth in the use of social media over the last few years, albeit not at the rate seen in some other major countries in the developing

world. As a subject of study, though, Nigeria is attractive in that much of the user-generated content from the country is in English, there is growing access to the Internet and wireless communications, and there are a number of large urban centers.

Using unigram features alone, we obtained an accuracy of 80% in predicting the gender of Twitter users, suggesting that content alone can be a good predictor of gender. The highest weighted unigram features also illustrate some interesting distinctions between men and women both topically and emotionally.

In the following sections we discuss related work, data collection and preparation, feature selection and machine-learning experiments, classification results, and conclusions and future work.

Related Work

Mislove et al. (2011), basing their work on a set of over one billion tweets collected between 2006 and 2009, looked at the geographic distribution, gender, and ethnicity of Twitter users in the United States. Gender was determined from matching the first name given in the user profile name field (i.e. the leftmost string in the name field) with a list of popular first names for babies born in the US from Social Security Administration data. They found that while the Twitter population was biased male, this bias lessened over the time period represented in the data.

Research into inferring the hidden attributes of social media users has had some limited representation in the literature in recent years. Chang et al. (2010) developed a hierarchical Bayesian model that used prior distributions of ethnicities over first and last names from U.S. census data to predict ethnicity from user names, using user's last names only, or first names and last names, for MySpace and Facebook users. Rao et al. (2010) developed models to predict the gender of Twitter users using Support Vector Machines (SVM), using as features emoticons, Web abbreviation features, and word unigrams and bigrams extracted from the concatenation of each user's tweets. Their classifier produced accuracies of 72%, 69%, and 72% for predicting gender using the sociolinguistic features, n-gram features, and both feature sets combined, respectively. Rao et al. (2011) developed a hierarchal Bayesian model for predicting the ethnicity and gender of Facebook users from Nigeria using letter n-grams from user names, as well as word n-grams from user content as features. They reported accuracies of 80% in predicting gender for both name features and name and content features together, suggesting that content gave little boost to prediction accuracy. Burger et al. (2011) based their work on Twitter users whose gender labels were derived by following the personal URL provided in a user's profile to blog sites that specified gender in the blog's profile. As features, they used letter and word n-grams derived from the concatenation of each user's tweets, screen name, name, and description present in the profile. They reported an accuracy of 84% using tweets, screen names, and descriptions, and 92% after adding name features. Only using tweets, they obtained 76% accuracy.

Data Collection and Preparation

To gather tweets from Nigeria, we used the Twitter Search API[1] and its *geocode* method, which accepts a position in latitude and longitude and a radius in miles. We ran searches for 45 Nigerian cities with populations over 100,000, using a radius of 40 miles. For users whose tweets were captured using this method, we used the Search API *search* method to obtain their other tweets. At the time the data used in this study were collected, the Search API returned only a user's screen name, not his or her name. To get user's names, we used the REST API to gather user profiles for the users found using the process described in the previous paragraph. From April 2010 to October 2011, 131 million tweets from 633,217 users were collected. Because the REST API is rate limited, we collected profiles for only 277,888 users.

For each tweet, the metadata returned using the *geocode* method includes the location from two sources: a field containing the location from the user's profile and a field containing a latitude/longitude pair populated using the optional geotagging feature supported by most mobile Twitter clients. We used these fields to calculate a user's location at the time of the tweet by matching the coordinates or location names against the Geonames gazetteer web service[2]. The geographic information returned by Geonames includes a place name, second administrative level (state or province), country code, latitude, and longitude.

Using this method, a total number of 398,534 users were found who had at least one reported location determined to be in Nigeria. Because the Twitter API's location assignments are noisy (approximately 18% of users had a resolved location outside Nigeria, and some locations inside Nigeria were reported for users that were verified to not have been in the country at the time), we restricted this study to the 117,155 users who had two or more resolved locations in Nigeria. This gave us a higher confidence that the user content used for training was actually from Nigeria. After merging these users with the available user profiles, we were left with 78,853 users.

[1] https://dev.twitter.com/docs
[2] http://www.geonames.org/export/
[3] http://developers.facebook.com/docs/reference/api/

Similar to Mislov (2011), where gender labels for US Twitter users were derived by matching the names in profiles against a list of popular first names from the US Social Security Administration, the gender labels for the Nigerian Twitter users in our study were obtained by matching first names with the first names taken from Nigerian Facebook users. Leveraging novel data sources – such as names from Facebook pages where gender labels are available – may be necessary when dealing with data from countries in the developing world since it may be difficult to obtain official name data analogous to that used by Mislov.

We collected posts and comments from a number of Nigerian Facebook pages for political figures and movements. For the authors of this content with public profiles, we collected name and gender pairs using the Facebook Graph API[3]. Looking only at first names that were associated solely with one gender, we obtained 6,680 female and 30,244 male names. After normalizing the Twitter user names by removing honorifics and titles, and eliminating Twitter accounts for organizations or businesses, we matched the leftmost string in the name with labeled first names from Facebook. Finally, we restricted our work to only those users for whom we had 50 or more tweets, so that only users who showed significant activity on the site were used for classification. This gave us 11,155 labeled users, with 4,034 females and 7,119 males, and reduced the total tweet count to 18.5 million tweets.

Prior to featurization, tweet text was normalized by expanding shortened words and, in many cases, correcting intentional misspellings of words. This also included translating a number of common words from Nigerian Pidgin English, an English Creole, to their Standard English equivalents. Web-specific acronyms were normalized to consensus representations. For example *OMG* and *OMFG* were normalized to *_omg_*, and variants of *LOL* were normalized to *_lol_* Emoticons and symbols from the Miscellaneous Symbols Unicode block were normalized to tokens such as *_happy_* and *_sad_*.

Feature Selection and Machine Learning Experiments

Using these 11,155 users for training and test data, and training instances, as in Rao (2010) and Berger, consisting of the concatenation of all of a given user's tweets, we generated features for word unigrams, hash tags, and psychometric properties derived from the Linguistic Inquiry and Word Count (LIWC) text analysis program (Pennebaker et al., 2007). LIWC maps the relative word frequency of certain high-frequency words to a set of psychological dimensions. Unigram and hash tag features were binary. The LIWC features took values between 0 and 1.

Feature Class	Feature Count
Unigrams	1,222,848
Hash Tags	9,008
LIWC	54
Total Features	1,231.910

Table 1 Feature Counts

Unigram features included all words, excepting stop words, and the normalized symbols for Web acronyms, emoticons, and Unicode symbols. Hash tag features were restricted on the most common of the 392,464 unique hash tags extracted from the almost 19 million tweets. Because the hash tag distribution is Zipfian, we used only those hash tags in the top 75% of the distribution, giving us a total of 9,008 hash tag features. LIWC was run on all user tweets, treating all tweets from a given user as a single document. The tweet text used for these features was normalized as described above for unigrams; however, stop words were retained because many LIWC properties are dependent on word counts for high frequency words such as function words and personal pronouns. LIWC generates 89 psychometric properties for each document analyzed. Thirty-five of these properties are summary properties. For this study we only used the 54 subordinate properties. The total counts for all features are given in Table 1.

Feature Class	Prec.	Rec.	F	Accu.
Hashtag	70.82	46.99	56.48	63.81
Hashtag, LIWC	71.51	47.17	56.83	64.18
Hashtag, LIWC, Unigram	**82.53**	**77.64**	**80.00**	**80.60**
Hashtag, Unigram	82.06	77.46	79.69	80.26
LIWC	70.76	75.07	72.85	72.02
LIWC, Unigram	82.22	77.79	79.93	80.48
Unigram	**82.50**	**77.47**	**79.90**	**80.51**

Table 2 – Classification Results

For our machine learning experiments we used the SVMLight (Joachims, 1999) Support Vector Machine implementation with a linear kernel. We performed ablation tests for the three feature classes, each test consisting of 10 Monte Carlo runs. For each run, we used balanced data, with male users under sampled, and an 80/20 training to test ratio. We allowed SVMLight to select the regularization parameter. The results for these runs are shown in Table 2 and report the mean values of all performance metrics (precision, recall, F-score, and accuracy).

Results

The combination of all features did marginally better than unigrams by themselves with a F-score of 80%, indicating that the hash tag and LIWC features added relatively little performance boost over simply using unigrams. The hash tag features did very poorly by themselves, giving low F-score results compared with the other feature sets alone. The LIWC features did much better, giving reasonable F-scores, although their performance alone was lower than that for unigrams.

	Female	Male
Highest Weighted Unigram Features	aww, hair, dear, soo, _omg_, miss, sad, yay, _lmao_, babes, darling, cute, _shaking_my_head_, boo, happy, hun, thank, birthday, dress, _happy_, husband, everyone, serious, crying, bed	brother, boss, oga, fuck, baba, play, dude, game, man, arsenal, nigga, guy, omo, album, team, united, pin, boys, die, playing, money, chelsea, far, vs, dull

Table 3 – Significant features for Unigram features

For the unigram features, we looked at the classification results and explored which features were the best discriminators for gender. To do this, we iterated across all of the Monte Carlo results for each feature set and computed the mean of the weighted sums of all support vectors for each feature. Because the positive class was female, features with positive weights were good discriminators for females and features with negative weights for males. In Table 3 we show the 25 highest weighted features for each class for the unigram feature set.

The unigram features that are significant for female users include emotive words or symbols such as *aww*, *mis*, *sad*, and the *_happy_* (i.e. *smiley*) emoticon. The features significant for males include profanity, and references to soccer such as game, *arsenal*, *united*, *playing*, and *chelsea*.

Conclusions and Future Work

This work demonstrates that it is possible to get robust estimates of hidden demographic characteristics of people represented in social media. We obtained good results in training a classifier for gender from the tweets authored by a population of Twitter users from Nigeria using simple Unigram features and the LIWC psychometric properties by using a Facebook-derived list of names and gender to label our training and test data. The classification results also showed an interesting set of discriminators for gender in the data. Future work will include applying this classification method to content from unlabeled Twitter users in Nigeria to look at the overall gender distribution in this social media channel. We argue that work like that described in this paper can help us get a clearer picture of just who uses social media, a necessary step when interpreting online opinion and reaction in different regions of the world.

Acknowledgements

This work was supported by Office of Naval Research grant N00014-10-1-0523.

References

Burger, J., Henderson, J., Kim, G., & Zarrella, G. (2011). Discriminating gender on Twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (pp. 1301.1309). Stroudsburg, PA: Association for Computational Linguistics.

Chang, J., Rosenn, I., Backstrom, L., & Marlow, C. (2010). epluribus: Ethnicity on social networks. In *Proceedings of the 4th International Conference in Weblogs and Social Media (pp. 1825)*. Menlo Park, CA: AAAI

Joachims, T. (1999). Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, & A. Smola (Eds.), Advances in Kernel Methods -Support Vector Learning. Cambridge, MA: MIT Press.

Mislove, A., Lehmann, S., Ahn, Y., & Onnela, J. (2011). Understanding the Demographics of Twitter Users. In *Proceedings of the 5th International Conference on Weblogs and Social Media*. Menlo Park, CA: AAAI.

Nigeria Facebook Statistics, Penetration, Demography - Socialbakers. (n.d.). Retrieved February 28, 2012, from http://www.socialbakers.com/facebook-statistics/nigeria

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (2007). The development and psychometric properties of LIWC 2007: Software manual.

Rao, D., Yarowsky, D., Shreevats, A., & Gupta, M. (2010). Classifying latent user attributes in Twitter. In Proceedings of the Second International Workshop on Search And Mining User-Generated Contents (p. 37). New York, NY: ACM Press.

Rao, D., Fink, C., & Oates, T. (2011). Hierarchical Bayesian Models for Latent Attribute Detection in Social Media. In In *Proceedings of the 5th International Conference in Weblogs and Social Media*. Menlo Park, CA: AAAI.