RIT Department of Computer Science
MSc Thesis Pre-Proposal:
Using Morpheme Modification Patterns to Extract Age From Twitter Microtexts

Nathaniel Moseley

May 9, 2012

1 Problem

Extracting latent user data from online, unstructured texts is a problem in the area of Computational Linguistics that has been gaining popularity. These texts often include blogs, articles, and status updates, such as found on Facebook and Twitter. The texts found on Twitter present a unique problem for linguistic analysis, as users are restricted to a 140 character maximum when submitting updates, which prompts users to develop various linguistic transformations to fit within the character restrictions.

Prior research has looked at patterns within these transformations to distinguish users' regions and Twitter clients [1], as well as analyzing basic shallow linguistic patterns to determine user gender and other information [4]. Other research on determining age from online or published texts and telephone conversations had relatively high levels of success using different linguistic classifiers [2]. Many texts use a binary classification, such as over or under 30, and achieve high levels of accuracy compared to relative naive baselines. One prior work managed to achieve standard mean error ranges of 4.1 to 6.8 years on a continuous age scale [2].

A few organizations already provide services that extract user data and statistics from various web services. This is of interest in several areas, such as marketing research for companies and in linguistic research. Many online services, including Twitter, do not have age as a defined element of a user's profile. Accurately determining age can be quite useful in various research areas.

I propose to determine a discrete age category for a person by analyzing their tweets for linguistic word transformation patterns. I will also experiment with other social user attributes collected and other shallow linguistic patterns.

2 Methodology

2.1 Data Gathering

The first contribution will consist of data gathering. I will develop a simple web page which Twitter users can use to self-label their social attributes. In order to develop a more useful data set, it will ask users for at least year of birth and their Twitter user name, and they will also have the option to include data such as their month of birth, geographical location, industry (student, unemployed, or part of the work force), gender, household situation, education level, etc. As a control, users will also have the option to input their astrological sign or number of pets. Limited additional data will be gathered from Twitter through the API as well, such as the time zone and age of the user's account. The page will also give users the option to solicit their contacts for participation through social networking sites. As user meta data is gathered by the web page, automated processes will subscribe to the users' tweets, download, and preprocess them.

In order to shield the user's real identity, I will only gather non-personally identifiable user information. Additionally, the page will present two simple elements of informed consent, briefly noting the nature of research and autonomous tweet analysis. The page will present the option to allow users' data to be used for this research and also to opt out of having it redistributed to future researchers. Additionally, users will have an option to back out at a later date, as well.

In the event that not enough data can be gathered in this way, there are minimally sized collections of annotated Twitter data that already contain age information, and I will have to contact authors of prior research [3] or data gathering, such as TREC.

2.2 Data Analysis

For the second, main contribution, I will do preprocessing of basic linguistic features with resources such as the Natural Language Toolkit. For the main linguistic feature of the experiment, I will implement or reuse the word transformation analysis tools used for prior Twitter social context research [1].

Once a data set is collected, I will separate it into development and testing sets, then I will use resources, such as Weka, in supervised learning experiments with the age target labels and extracted features. Additionally, time allowing, I will attempt unsupervised learning, and generated clusters compared to the various labels for cluster purity statistics, as well as looking longitudinally for changes in user behavior.

3 Evaluation

The process should result in at least a discrete age category analysis per user, such as 5 or 10 year buckets, which I will compare to the gathered age data to get accuracy, as well as some manner of mean errors. This metric could be exchanged for a continuous age value extraction, which is generally more complex to get. Time and storage are other metrics, but they are not what this approach is testing. Accuracy should be determined from several runs of several different supervised and unsupervised classifiers.

There is limited statistical data about Twitter's user base which suggests 55% are female, and most are in the age range 26-34, followed by 35-44. A naive guess in these ranges will have the highest accuracy, and serve as a naive baseline. An additional baseline using a naive Bayesian classifier may offer a higher baseline. Provided that a level of accuracy reasonably above baseline values for age prediction can be achieved, the hypothesis will be verified. If a satisfactory level of accuracy can not be achieved, the hypothesis will be shown incorrect.

References

[1] Stephan Gouws, Donald Metzler, Congxing Cai, and Eduard Hovy. Contextual bearing on linguistic variation in social media. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 20–29, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[2] Dong Nguyen, Noah A. Smith, and Carolyn P. Rosé. Author age prediction from text using linear regression. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, LaTeCH '11, pages 115–123, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[3] Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international Workshop on Search and Mining User-Generated Contents*, SMUC '10, pages 37–44, New York, NY, USA, 2010. ACM.

[4] Ruchita Sarawgi, Kailash Gajulapalli, and Yejin Choi. Gender attribution: Tracing stylometric evidence beyond topic and genre. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, CoNLL '11, pages 78–86, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.