Using Word and Phrase Abbreviation Patterns to Extract Age From Twitter Microtexts

by

Nathaniel Moseley

A Thesis Proposal Submitted in Partial Fulllment of
the Requirements for the Degree of Master of Science in Computer Science

Supervised by

Dr. Manjeet Rege
Department of Computer Science
B. Thomas Golisano College of Computing and Information Sciences
Rochester Institute of Technology
Rochester, New York

Dr. Cecilia Ovesdotter Alm
Department of English
College of Liberal Arts
Rochester Institute of Technology
Rochester, New York

August 6, 2012

Approved by:

Dr. Manjeet Rege
Primary Advisor, Department of Computer Science

Dr. Cecilia Ovesdotter Alm
Reader, Department of English

Abstract

The wealth of texts available publicly online for analysis is ever increasing. Much work in computational linguistics focuses on syntactic, contextual, morphological and phonetic analysis on written documents, vocal recordings, or texts on the internet. Twitter messages present a unique challenge for computational linguistic analysis due to their constrained size. The constraint of 140 characters often prompts users to abbreviate words and phrases. Additionally, as an informal writing medium, messages are not expected to adhere to grammatically or orthographically standard English. As such, Twitter messages are noisy and do not necessarily conform to standard writing conventions of linguistic corpora, often requiring special pre-processing before advanced analysis can be done.

In the area of computational linguistics, there is an interest in determining latent attributes of an author. Attributes such as author gender can be determined with quite some success from many sources, using various methods, such as shallow linguistic patterns or topic analysis. Author age is more difficult to determine, but previous research has been fairly successful at classifying age as a binary, ternary, or even continuous variable using various techniques.

Twitter messages present a difficult problem for latent user attribute analysis, due to the pre-processing necessary for many computational linguistics analysis tasks. An added logistical challenge is that very few latent attributes are explicitly defined by users on Twitter. Twitter messages are a part of an enormous data set, but the data set must be independently annotated for latent writer attributes before any classification on such attributes not defined through the Twitter API can be done. The actual classification problem is a particular challenge due to the restrictions on tweets.

Previous work has shown that word and phrase abbreviation patterns used on Twitter can be indicative of some latent user attributes, such as geographic region or the Twitter client used to make posts. Language changes tend to be driven by youths, often in adolescence, who tend to gradually affect older authors. Older language users either adopt the changes or resist them. Twitter is a relatively new service. This study explores if there there are longitudinal patterns evident in 6 years of online interactions.

I propose the development of a large, growable data set annotated by Twitter users themselves for age and other useful attributes. I also propose an extension of prior work on Twitter abbreviation patterns to determine if these linguistic patterns are indicative of author age at time of posting and can be useful in determining an author's age, and to investigate what, if any, changes in these patterns occur over time. Lastly, I propose a time line for thesis completion with deliverables.

Contents

1 Introduction

Discovering latent data about authors of texts is a recognized topic in natural language processing. Historically, latent attribute gathering was done by researchers annotating texts or voice recordings. As computers came into existence, methodologies for these analyses developed into the field of computational linguistics. With the rising popularity of the internet, most texts and recordings that would be subjects of analysis are available online, and online texts themselves are an ever growing corpora. Computational linguistics provides methods to process and analyze these large language collections. Many corpora of internet language have received such attention, such as blog posts, online news, and scientific publications [19, 20].

One of the newest corpora developing and increasingly receiving attention is a set of texts linguists have termed *microblogs*. These include short, often character-limited messages, such as those found on Facebook update messages, SMS text messages, and Twitter messages. These present an interesting type of linguistic corpus, but often, particularly in the case of Twitter, the texts are noisy and difficult to work with because of nonstandard language use. Character restrictions prompt authors to develop and use word and phrase abbreviations to convey their messages in fewer characters [7]. Cook and Stevenson identified 12 types of abbreviations often used in SMS messages [2]. Usage of these abbreviations results in messages with significant portions that are out-of-vocabulary (OOV) for the language in which they are written. Such added linguistic sparsity can make linguistic analysis, such as for context, genre, and topic detection, more difficult to perform [15].

Part of the process of preparing Twitter messages for analysis involves mapping OOV word and phrase abbreviations to in-corpus equivalents. Gouws *et al.* identified 9 abbreviation patterns used in Twitter messages which account for over 90% of the transformations used. They used those patterns to identify the region from which English-writing Twitter users were posting (according to time zone data provided by Twitter) as well as the client (iPhone, android, Twitter website, *etc.*) used to post the message [5]. Based on the success of using deep syntactic patterns and shallow, token-level linguistic features to identify author age [20], it is reasonable to assume that such abbreviation patterns can be used to identify user age on Twitter.

Some prior studies have looked at identifying Twitter user age with relative success, but each had to develop a data set by hand, having researchers examine tweets for indicators of age, usually assigning users to binary or ternary categories, such as over or under 30 [18]. While the accuracy of such data is generally high, the granularity is low. Because of the required and costly manual annotation and time involved, the collected data sets were relatively small.

My thesis will present a novel data set developed to improve future Twitter research, in addition to reporting on analysis and processing methods and results of their application, as well as updated findings regarding user age classification using word and phrase abbreviations found in Twitter messages. Section 2 gives a background of computational linguistics as it pertains to my topic. Section 3 covers previous work pertaining to Twitter analysis in computational linguistics. Section 4 explains the details of my hypotheses as they relate to existing problems. Section 5 outlines the methodologies that I will use for my solution. Section 6 presents a timetable with deliverables required for completion of this thesis.

2 Background

Computational linguistics began with machine translation efforts in the 1950s, because researchers believed that computers would be able to produce effective translations more quickly than their human counterparts [9]. Today, computational linguistics has many other applications beyond translation. An understanding of language and meaning can allow a computer to more effectively interact with its human operators and can be used for text data analytics. This takes many forms: in advertising, there are products such as Google's Adsense show topic-centered ads based on page content [10]; in email, spam filters analyze messages for typical real usage and spam usage, then automatically flag messages that seem like spam; in human computer interaction, a computer is able to communicate with its user more effectively if it knows different ways to present information based on its user's age, education, emotional state, or other attributes.

Age is an acknowledged factor in language use, as noted by Wagner. A person's writing and speech patterns change over time as they learn and develop ('age grading') [21]. An individual goes through many stages of language use through childhood, adolescence, and adulthood. In childhood, language is acquired and understanding and conversational-interaction skills are developed. Adolescence marks a period of change in many respects, and a person trying to find their identity socially also explores their identity linguistically. Into adulthood, language use continues to change, often in response to changes in community language use ('generational change'). Youths often spur language changes, and adults respond by either matching these community changes, or adhering to their already learned linguistic standards [21].

Texts found on the internet present a gigantic collection for linguistic analysis of age grading. The simplicity and low cost of writing on the internet allows individuals to publish a prolific library of formal and informal texts. Many people keep blogs, write on message boards and newsgroups, or participate in social networking sites. The types of writing available range from long, scientific writing, which adheres to language standards with standard grammar, syntax, and orthography, to short, informal messages found on Twitter.

2.1 Twitter

Twitter is a relatively new service, made public in 2006, which allows users to post 140 character *updates*. They can follow other users, such as friends, celebrities, or companies to be alerted to those users' updates. Users can engage in public or private conversations with these short messages, forward messages they think their followers will be interested in by *retweeting*, or just post whatever they are doing, thinking, or want to write [14]. With 140 million active users and 340 million tweets per day, there is an incredible amount of information and text exchanged on Twitter [13].

In July 2011, Twitter crossed the one million mark for developer applications registered to use the Twitter API [11]. Twitter provides developers and researchers a robust API with which to interact with accounts and access user information and tweets. Every user defines a username, and optionally a real name, description, and location. Also available are the account's associated timezone and the account creation timestamp. Each tweet has several pieces of information available in addition to the message, such as its timestamp, the Twitter client it was posted from, if it was part of a conversation, a retweet, and count of people who retweeted it. However, Twitter does not elicit other data about the author that might be useful for latent attribute analysis, such as age or other demographic information.

3 Related Work

A variety of work has been published that focuses on linguistic analysis for author age, much of which focuses on high-level and contextual clues, such as analyzing topic and genre or n-gram patterns. Garera and Yarowsky found that sociolinguistic features (amount of speech in conversation, length of utterances, usage of passive tense, *etc.*) characterizing types of speech in conversation between partners improved age, gender, and native language binary attribute classification [3]. Nguyen *et al.* went a step beyond many other studies and classified age as a continuous variable. They found that stylistic analysis, unigram, and part of speech analysis were all indicative of author age [17]. Rosenthal and McKeown analyzed online behavior associated with blogs and found that behavior (number of friends, posts, time of posts, *etc.*) could effectively be used in binary age classifiers, in addition to linguistic analysis techniques similar to those mentioned above [19].

Similarly, many works investigating linguistic gender and age indicators focus on non-contextual and low-level indicators. Sarawgi *et al.* explored non-contextual syntactic patterns and morphological patterns to find if gender differences extended further than topic analysis and word usage could indicate. They found, by using probabilistic context-free grammars, token-based models, and character-level language models, that gender-specific patterns extend to the character-level, even in modern scientific papers [20].

Much of the analysis that has been done focuses on formal writing or conversation transcripts, which generally conform to standard English corpora and dialects, syntax, and orthography. Recently, more works have begun to look at new written and online texts which do not tend toward prescriptive standards, including SMS messages and social networking blurbs, such as Facebook and Twitter messages. There are various challenges when trying to analyze these typically noisy texts. Misspellings, unusual syntax, and word and phrase abbreviations are common in these texts, which most linguistic analysis tools can not deal with. Rao *et al.* found n-gram and sociolinguistic cues in unaltered Twitter messages could be used to determine age (binary: over or under 30), gender, region, and political orientation, similar to works that have focused on more formal writing [18]. Gimpel *et al.* developed a part-of-speech tagger designed to handle unique Twitter lexicon by extending the traditionally labeled parts of speech to include new types of text such as emoticons and special abbreviations [4].

Some research takes a different approach to noisy text, such as that found on Twitter. Before performing traditional text analysis, it is often first cleaned. There are various ways to approach the text normalization problem, such as treating it as a spell-checking problem, a machine translation problem, in which messages are translated from a noisy origin language to a target language, or as an automatic speech recognition (ASR) problem. ASR is often useful for texts such as SMS, since many of the OOV words are phoneme abbreviations using numbers [5]. Kaufmann and Kalita presented a system for normalizing Twitter messages into standard English. They observed that pre-processing tweets for orthographic modifications and twitter-specific elements (@-usernames and # hashtags) and then applying a machine translation system worked well [15].

Gouws *et al.* built on top of the techniques of Contractor *et al.* [1] using the pre-processing techniques of Kaufman and Kalita [15] to determine types of word and phrase transformations used to create OOV tokens in Twitter messages. Such transformations include phonemic character substitutions (too → 2; late → l8), dropping trailing characters or vowels (going → goin), and phrase abbreviations (laughing out loud → lol). They analyzed patterns in usage of these transformations compared to user time zone and Twitter client to see if there was a correlation. The analysis showed that variation in usage of these transformations were correlated with user region and Twitter client [5]. This thesis seeks to extend this work and analyze these transformations with respect to user age.

4 Hypothesis

As is, there are a few techniques to determine latent user attributes from general texts, but very few that specifically target Twitter messages and their unique corpus. Of those works that have focused on Twitter messages, they have two main types of shortcomings: (1) they focus on a small set of gathered data from hand-picked users, where latent attributes are determined and entered by researchers or volunteers, as opposed to by the Twitter users providing the information themselves; or (2) they use the full set of Twitter users and messages, but are limited to the latent attributes that are provided through the Twitter API. First, I propose to solve these issues through collection of a new, more robust data set where the tweeters themselves label their Twitter feds with demographic information.

Second, based on the work of Gouws *et al.*, I hypothesize that word and phrase abbreviation patterns used to write tweets are indicative of user age, as they are indicative of a user's region and Twitter client [5]. Third and lastly, I hypothesize that usage of these abbreviations changes as a user ages or spends more time using the Twitter service, like how language changes as a person ages and community language use evolves.

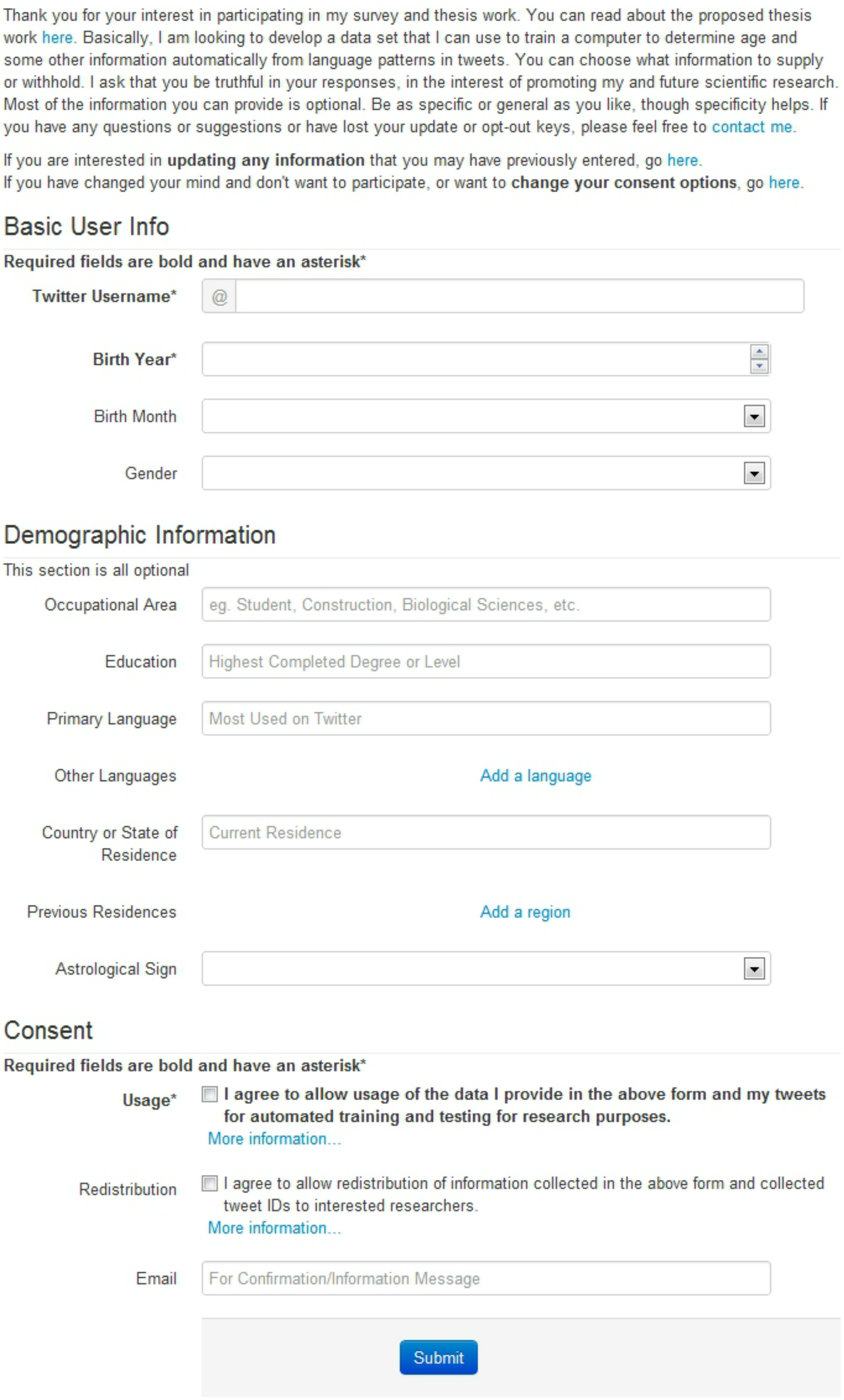
5 Solution Design and Implementation

The work of this thesis is comprised of three parts: collection of a Twitter data set, described in subsection 5.1; normalization of the collected user demographic information and association of users to tweets, described in subsection 5.2; and analysis of the abbreviation patterns within collected tweets for relation to author age, as well as longitudinal evolution, described in subsection 5.3.

5.1 Data Collection

First, I will develop a user-driven Twitter data set with high accuracy and granularity containing at a minimum a user's Twitter username and year of birth. The data set is populated via a user-friendly web form, shown in Figure 1, via Twitter users entering information to be associated with the account(s) they control. It is assumed that people submitting their information are being truthful, but it is possible that false information will be submitted. The majority of information collected should be truthful, and any false information may show as an outlier in some part of the analysis.

The web form explains basic requirements of the data collection and links to pages with more information. Users will be informed of the privacy of their data and given the opportunities to allow



Thank you for your interest in participating in my survey and thesis work. You can read about the proposed thesis work here. Basically, I am looking to develop a data set that I can use to train a computer to determine age and some other information automatically from language patterns in tweets. You can choose what information to supply or withhold. I ask that you be truthful in your responses, in the interest of promoting my and future scientific research. Most of the information you can provide is optional. Be as specific or general as you like, though specificity helps. If you have any questions or suggestions or have lost your update or opt-out keys, please feel free to contact me.

If you are interested in **updating any information** that you may have previously entered, go here.
If you have changed your mind and don't want to participate, or want to **change your consent options**, go here.

Basic User Info

Required fields are bold and have an asterisk*

Twitter Username*	@
Birth Year*	
Birth Month	▾
Gender	▾

Demographic Information

This section is all optional

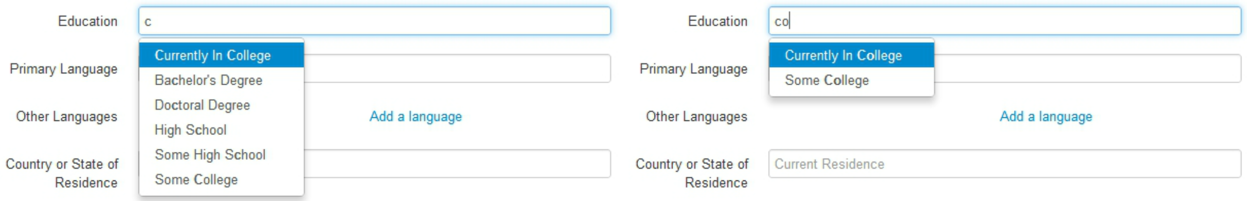
Occupational Area	eg. Student, Construction, Biological Sciences, etc.
Education	Highest Completed Degree or Level
Primary Language	Most Used on Twitter
Other Languages	Add a language
Country or State of Residence	Current Residence
Previous Residences	Add a region
Astrological Sign	▾

Consent

Required fields are bold and have an asterisk*

Usage*	□ **I agree to allow usage of the data I provide in the above form and my tweets for automated training and testing for research purposes.** More information...
Redistribution	□ I agree to allow redistribution of information collected in the above form and collected tweet IDs to interested researchers. More information...
Email	For Confirmation/Information Message

Figure 1: The web form for Twitter users to submit their information.



(a) A user types ‘c’

(b) A user types ‘o’ after typing ‘c’

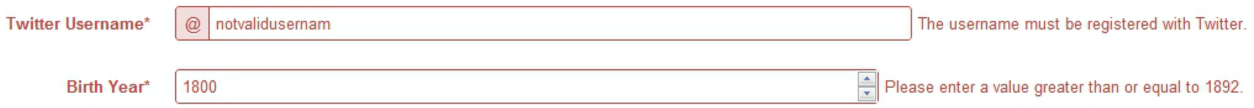
Figure 2: The web form suggests options as the user types, sorted alphabetically. The top option is a user-generated option that has been used more than a threshold number of times. The other options are predefined.

their data to be redistributed to future researchers, update the information they provide, or opt out entirely from the research. The minimum amount of information that will be collected is Twitter username and year of birth. Twitter users can also supply 8 additional attributes: (1) month of birth, (2) gender, (3) occupational area, (4) highest education level, (5) languages used, (6) regions of residence, (7) astrological sign, and (8) email. Most users appear willing to include some if not all of this additional information. In the interest of future research, non-identifying demographic information is collected in addition to the age information that will be used for this thesis. Astrological sign is suggested as a control variable, since there has been no support of a scientific link between astrology and personal characteristics [8].

Email is collected for part of the form processing and future automated notifications. On completion of the form, users are shown acknowledgment and links to opt out, or update their information. If they provided an email, they are emailed the same collection of information. Lastly, a Twitter account for this thesis (@NMoseleyThesis) follows the user. In the event that the user's tweets are protected (only authorized users may view their tweets), this allows them to be read for later analysis.

The web form populates a database backend which in turn offers suggestions for the form inputs. As shown in Figure 2, a set of predefined values for each attribute, as well as the most frequently used user-created values are suggested as a user types in the form, via JavaScript. Ideally, the suggestions will help increase initial data coherence. Additionally, each form field only allows certain characters to be entered, according to the type of data that is expected.

Two checkboxes are included in the form which have brief sentences indicating user consent and links for a more in-depth explanation, which also show the same explanation on mouseover, if the user has JavaScript enabled. The first is required for submission of the form. It explains that the data collected will be used for autonomous analysis and will not be sold or redistributed except in cases where the user checks the second consent box. Checking the second checkbox acknowledges



Twitter Username* @ notvalidusernam The username must be registered with Twitter.

Birth Year* 1800 Please enter a value greater than or equal to 1892.

Figure 3: The web form highlights input errors and displays an associated message, requiring the user to correct errors before the form will submit.

that the user is willing to allow the data collected to be redistributed to interested researchers in the future. It defines collected data to include all entries in the web form, as well as collected tweet identifiers. The tweets themselves cannot be redistributed, but the identifiers produced by Twitter can be, as per the Twitter data use policy [12].

The web form is cross-browser compatible, including on mobile platforms, and performs the same with or without JavaScript enabled. Some added functionality, such as the above input suggestions and consent information mouseover is not available without JavaScript. All value checking is still done server-side, whether or not value checking is done with JavaScript. However, when JavaScript is enabled, users will be required to enter correct values and informed of their errors, as shown in Figure 3.

For development purposes, the database may be populated with public data about public personas (celebrities, politicians, *etc.*) if there is not enough initial data. Data will be solicited from a range of sources on the internet, including on Twitter itself. Pages linked to from the web form offer social networking buttons, allowing users to suggest to their contacts they should also participate. Soliciting participation from public figures with many followers could dramatically boost the enrollment and referral rates.

5.2 Data Normalization and Tweet Loading

Before the collected user data can be used in analysis, each of the 10 values (section 5.1) will need to be normalized to ensure that data values are consistent and useful. Username, birth year, month, gender, astrological sign, and email should not need any additional cleaning. The web form and database backends ensure that each field has a valid value. The username must be registered with Twitter, the birth year must be within the last 120 years, the email must be a valid format, and the rest must be empty or contain one of the relevant pre-specified values.

Occupation, education, languages, and regions will need special pre-processing, since users are allowed to write in any value that conforms to the character level constraints. Unique values are most frequently included in education and region where users write phrases or sentences explaining their educational history or history of where they have lived. Additionally, each of these values may need to be generalized in order to be most useful. Some people may include cities in their submitted regions, for example, and these would need to be generalized to the respective state or country region.

I will be developing a code base to download all the tweets for a user and do linguistic analysis on them. The methods used for analysis will be discussed in subsection 5.3. I plan to use the Twitter4J library in Java to authenticate and download tweets from the users that have volunteered for this research. Normalization of tweet texts will be done as part of the analysis described in 5.3.

5.3 Age Categorization

I plan to analyze the collected tweets as an extension of the work of Gouws *et al.* [5]. For much of the basic linguistic analysis, I plan to use the Natural Language Toolkit (NLTK) [16], a Python library. Gouws *et al.* used a naive context free analysis for part of their linguistic analysis, and they used the NLTK throughout their work [5].

For the second part of their analysis, Gouws *et al.* first pre-processed all messages according to the work of Kaufmann and Kalita [15]. Kaufmann and Kalita observed that similar processes can be much improved by some simple pre-processing, specific to Twitter messages. This involves replacing all @-username tokens with a placeholder "*USR*", all URL tokens with "*URL*", and normalizing # hashtags. If a hashtag token is within a sentence, it is assumed to be part of the sentence and the '#' is removed. If it is at the end of a message, it is determined to be extraneous and is removed completely. Next, the message is tokenized and compared to a standard English corpus developed from the LA Times. For all OOV tokens, an in-corpus equivalent is determined, and the substitution is recorded [5].

The bulk of the thesis will be analysis of the collected tweets and recorded word and phrase substitutions, then training and testing a classifier to predict user age from the recorded abbreviation patterns. I plan to use the WEKA toolkit's machine learning algorithms [6] classifiers to classify the tweets, using supervised learning algorithms such as the Bayesian network classifier. I will report using standard metrics, such as accuracy, precision, recall, and F-1. I would also like to look at unsupervised learning through clustering algorithms to determine if patterns could be observed that way.

For evaluation, I will begin by looking at the results of the tweet normalization processes to see how my results compare to those of Gouws *et al.* To determine the efficacy of using word and phrase transformation patterns to determine age, I will have to compare the results of those experiments to relative naive baselines. I will likely experiment with binning ages into ten-year groups with separate runs for bins ending on five-year and ten-year marks and with age as a continuous variable. For each experiment, I will be comparing to naive baseline results. Additionally, I can run equivalent experiments to see if my test patterns are indicative of collected astrology information. I would expect the results for astrology tests to be pure noise. This can also serve as a sanity check to make sure my experiment is designed well and I am not introducing biases anywhere.

Given a reasonable success rate over baselines for determining user age and sufficient time, I would like to do longitudinal analysis of the transformation patterns to see if any communal change or other consistent longitudinal patterns exist in the time span of collected tweets. There are only six years of Tweets to draw on at the present time, but that may be enough to see some evolution in Twitter language use.

6 Roadmap

I have already developed the data collection web interface and backend described in subsection 5.1. I currently have 65 participants signed up. Going forward, I will continue to solicit new participants and seek to improve the referral rates, and work to complete the deliverables below within the following timetable.

6.1 Deliverables

Upon completion of this thesis, I plan to deliver:

1. A thesis document, explaining the experiments and results in depth

2. Source code that I develop or am allowed to redistribute and relevent open source licensing and release notes

3. A summary of the current state of the collected data set

4. Details of how to gain access to the data set for future research and licensing and release notes

6.2 Timetable

Date(s)	Objectives
May 2012 – June 2012	✓ Develop and release data gathering web page ✓ Write preproposal and seek approval for it
July 2012	• Write proposal and seek approval for it • Solicit Twitter users and public figures to participate • Fill database with public data if needed
August 2012	• Develop tweet retrieval and basic linguistic analysis software • Extend Gouws' word and phrase transformation determining software and apply to collected tweets
September 2012 – December 2012	• Conduct experiments on collected data • Refine and re-run experiments as necessary • Create useful graphics from collected results • Write thesis • Defend thesis

References

[1] Danish Contractor, Tanveer A. Faruquie, and L. Venkata Subramaniam. Unsupervised Cleansing of Noisy Text. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 189–196, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[2] Paul Cook and Suzanne Stevenson. An Unsupervised Model for Text Message Normalization. In *Proceedings of the Workshop on Computational Approaches to Linguistic Creativity*, CALC '09, pages 71–78, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[3] Nikesh Garera and David Yarowsky. Modeling Latent Biographic Attributes in Conversational Genres. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 710–718, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[4] Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments. In *ACL (Short Papers)*, pages 42–47. The Association for Computer Linguistics, 2011.

[5] 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.

[6] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1):10–18, November 2009.

[7] Bo Han and Timothy Baldwin. Lexical Normalisation of Short Text Messages: Makn Sens a #twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 368–378, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[8] Peter Hartmann, Martin Reuter, and Helmuth Nyborg. The Relationship Between Date of Birth and Individual Differences in Personality and General Intelligence: A Large-scale Study. *Personality and Individual Differences*, 40(7):1349–1362, May 2006.

[9] John Hutchins. Retrospect and Prospect in Computer-Based Translation. In *Machine Translation Summit VII, 13th-17th September 1999, Kent Ridge Labs, Singapore. Proceedings of MT Summit VII "MT in The Great Translation Era"*, pages 30–34, Tokyo, September 1999. Asia-Pacific Association for Machine Translation.

[10] Google Inc. How does Google target ads to my website? - AdSense Help. [Online]. Available: `http://support.google.com/adsense/bin/answer.py?hl=en&answer=9713`, 2012.

[11] Twitter Inc. Twitter Blog: One Million Registered Twitter Apps. [Online]. Available: `http://blog.twitter.com/2011/07/one-million-registered-twitter-apps.html`, July 2011.

[12] Twitter Inc. Developer Rules of the Road. [Online]. Available: `https://dev.twitter.com/terms/api-terms`, May 2012.

[13] Twitter Inc. Twitter Blog: Twitter Turns Six. [Online]. Available: `http://blog.twitter.com/2012/03/twitter-turns-six.html`, March 2012.

[14] Twitter Inc. Twitter Help Center — Get to Know Twitter: New User FAQ. [Online]. Available: `https://support.twitter.com/groups/31-twitter-basics/topics/104-welcome-to-twitter-support/articles/13920-get-to-know-twitter-new-user-faq`, 2012.

[15] Max Kaufmann and Jugal Kalita. Syntactic Normalization of Twitter Messages. In *Proceedings of ICON-2010: 8th International Conference on Natural Language Processing*, pages 149–158, Karagpur, India, 2010. Macmillan Publishers, India.

[16] Edward Loper and Steven Bird. NLTK: the Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, volume 1 of *ETMTNLP '02*, pages 63–70, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[17] 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.

[18] 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.

[19] Sara Rosenthal and Kathleen McKeown. Age Prediction in Blogs: A Study of Style, Content, and Online Behavior in Pre- and Post-Social Media Generations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 763–772, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[20] 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.

[21] Suzanne Evans Wagner. Age Grading in Sociolinguistic Theory. *Language and Linguistics Compass*, 6:371–382, June 2012.