Twitter N-Gram Corpus with Demographic Metadata

Amaç Herdağdelen*

January 1, 2012

Abstract

Social media is a natural laboratory for linguistic and sociological purposes. In micro-blogging platforms such as Twitter, people share hundreds of millions of short messages about their lives and experiences on a daily basis. These messages, coupled with the metadata about their authors, provide an opportunity to understand a wide variety of phenomena ranging from political polarization to geographic and demographic lexical variation. Lack of publicly available micro-blogging datasets has been a hindrance to replicable research. In this paper, I introduce Rovereto Twitter N-Gram Corpus, a publicly available n-gram dataset of Twitter messages, which contains gender-of-the-author and time-of-posting tags associated with the n-grams. I compare this dataset to a more traditional web-based corpus and present a case study which shows the potential of combining an n-gram corpus with demographic metadata.

DRAFT

1 Introduction

Social media and networking platforms have become pervasive in our daily lives (Madden and Zickuhr, 2011). As a result, a large collection of social media posts became publicly available. Massive, user-generated data present new research opportunities in social science and computational linguistics (Lazer et al, 2009; Thelwall et al, 2010; Argamon et al, 2007). The availability of user-generated data also presents an abundant opportunity to tap the personal and informal aspect of communication and daily routines of people, otherwise very hard – if not impossible – to observe in more controlled settings. In this paper, I present Rovereto Twitter N-Gram Corpus (RTC), an n-gram dataset of almost 75 million short, personal, social media posts in English, along with the information on the gender of the authors of the posts and the time of the posting[1].

*Crimson Hexagon, Boston, USA (work carried out at CIMeC, University of Trento, Italy)

[1] `http://clic.cimec.unitn.it/amac/twitter_ngram`

Twitter[2], a micro-blogging service, lets its users post short messages (*tweets*). Everyday hundreds of millions of tweets are shared publicly. The content of tweets may include, but is not limited to, comments on recent events, what the users have been doing, or small talk – in other words, anything that people may find worth to share with their friends and other people (Kelly, 2009; Naaman et al, 2010). Twitter-based data have already been used to study a variety of social and computational problems such as commonsense data collection and gender modeling (Herdağdelen and Baroni, 2011), news propagation analysis (Lerman and Ghosh, 2010), fast earthquake detection (Sakaki et al, 2010), influenza outbreaks Culotta (2010), daily mood changes (Golder and Macy, 2011), linguistic change (Cunha et al, 2011), community detection and political polarization (Conover et al, 2011), demographic and geographic lexical variation (O'Connor et al, 2010), and prediction of election results (Tumasjan et al, 2010).

Twitter data is publicly available via an application programming interface (API)[3]. However, it has not been possible to construct and share a publicly available dataset of tweets because of legal limitations on redistribution of raw tweets. The lack of a commonly available dataset is an impediment to replicability – researchers cannot evaluate their algorithms and methods on a shared dataset – and it results in duplication of effort – almost every researcher who wishes to utilize Twitter data is faced with the burden of collecting a new dataset.

This study describes an extensive, publicly available dataset of tweets. Put briefly, I built and shared a Twitter-based dataset using n-grams; thereby, overcoming the limitations on the redistribution of raw tweets. This is similar to the strategy employed in Google Books N-Gram Dataset which used n-grams because of the limitations imposed by copyright issues (Michel et al, 2011). Also, using n-grams gives an additional advantage for providing aggregate statistics about the demographics of the users who used the n-grams in their tweets, without creating privacy issues. Combination of demographic data and the time of posting with the actual content of tweets presents previously unavailable opportunities for research in computational linguistics, sociolinguistics, and related areas. Here, I focus on gender and posting time of the tweets. The gender label allows to compare the differences in the content or linguistic patterns between the tweets of the two genders. The time label enables to identify which topics the users are more interested in different times of the day or week, and study how the engagement pattern with social media changes throughout time. Together the two labels provide a rich metadata layer on the n-gram corpus. The corpus is distributed under a Creative Commons Attribution-NonCommercial-ShareAlike license[4].

The rest of this paper is structured as follows. In Section 2, I discuss previous work in building web-based or user-generated corpora. In Section 3, I explain the data collection procedure and the technical details such as tokenization, language identification, gender guessing and extracting the local time of posting.

[2] `http://www.twitter.com`
[3] `https://dev.twitter.com/`
[4] `http://creativecommons.org/licenses/by-nc-sa/3.0/`

In Section 4, I compare the newly constructed corpus to a web-based corpus. In Section 5, I show how the corpus with demographic metadata can be used to extract stereotypical gender expectations from Twitter. Finally, I conclude with a general discussion and future directions in Section 6.

2 Relevant work

One of the largest Twitter-based corpora that was released to the public was the Edinburgh Twitter Corpus (Petrović et al, 2010) which consisted of approximately 100 million tweets and which had to be retracted following a request coming from Twitter (Osborne, ????). Another similar example is the 427-million-tweet dataset released by Yang and Leskovec (2011) which was again taken down[5] (Yang and Leskovec, 2011). Note that in both cases, the datasets consisted of full tweets. Even during their brief period open to the public, these datasets enjoyed considerable interest among the research community (Naveed et al, 2011; Herdağdelen and Baroni, 2011; Bifet and Frank, 2010; Zhao et al, 2011; Liu et al, 2011).

Tweets2011 is a dataset of 16 million tweet identifiers (but not the content or metadata) released as part of TREC 2011 Microblog track[6]. The aim is to agree on a shared set of tweets on which researchers may work, without actually releasing the content, and letting researchers download the tweets from the Twitter API[7] themselves. Unfortunately, with regular API rates, a user can only download 350 tweets per hour or has to resort to screen scraping – both of which result in very long fetching times, and the latter option has the unfortunate drawback of being illegal in most cases.

Another source of personal textual data is weblogs (blogs). Even though blogs are not deeply embedded in the daily routine of people as micro-blogging platforms are, they provide a rich dataset for opinion modeling. In 2009, Yano et al (2009) released a dataset which consisted of several million posts and associated comments collected from five political blogs between 2007 and 2008. As part of the ICWSM 2011 Data Challenge, a dataset consisting of over 133 million blog posts and 231 million social media posts collected between January 13th and February 14th 2011 was released to the public (Burton et al, 2011).

There is a growing interest in constructing web-based corpora by crawling web pages. In this paper, I will use ukWaC, a billion-word English corpus built by crawling web pages in the .uk domain, as a reference corpus to show the differences between user-generated corpus and "traditional" web-based corpora (Baroni et al, 2009). Interested readers can refer to Baroni et al (2009) and Hundt et al (2007) for further examples and discussion of using the web as a corpus.

In addition to the studies cited in the previous section, which employ user-generated content in a variety of tasks, social media and Twitter in particular

[5] `http://snap.stanford.edu/data/twitter7.html`
[6] `http://trec.nist.gov/data/tweets/`
[7] `https://github.com/lintool/twitter-corpus-tools`

are also used for linguistic tasks. As an example, Pak and Paroubek (2010) use sentiment analysis in Twitter to help adjective disambiguation in Chinese. There is also an increasing effort in identifying sarcasm and verbal irony in short texts (González-Ibáñez et al, 2011; Davidov et al, 2010).

Before moving on to the details of text processing and data extraction, it is worth to discuss briefly the content of tweets. Earlier studies showed that more than half of the sampled tweets were about "self at present" or anecdotal or personal (Naaman et al, 2010; Kelly, 2009). In addition, the temporal references in tweets are primarily about the immediate present or near future (Ritter et al, 2011). These findings suggest that people use Twitter in an embedded manner to their daily lives, and it is reasonable to assume that people talk (i.e. *tweet*) about their daily experiences and what matters to them.

3 Data Collection, Preprocessing and Distribution

In this section, I will explain the data collection, preprocessing steps such as language identification, spam removal and tokenization, and metadata generation heuristics to identify the gender of the authors and posting time of the tweets.

Fetching—Twitter Streaming API[8] provides a continuous stream of roughly 1% of all public tweets. Among the data fields that were received, the important ones for our purposes were tweet content (up to 140 characters), tweet posting time in UTC, author's name, author's screen name and user id, and author's local timezone. The fetching process continued for seven months from December 5, 2010 to June 25, 2011, and resulted in 239 million tweets.

Language Identification—Twitter is a multi-lingual environment with a significant number of non-English tweets (Hong et al, 2011). We want an English-only corpus; however, filtering out non-English tweets is not a trivial problem because of the short text length and informal nature of tweets. Abbreviations, typos and acronyms are commonly observed in Twitter. Therefore, language models based on more traditional texts are inadequate. Here, I used a combination of an off-the-shelf language model and bootstrapping techniques. First, a language model trained on more traditional text was used to create a large Twitter-based training set, and then, a custom-built language model was trained on this training set to identify the language of the tweets. The details are below.

I first used the language detection library (which I refer to as LDL) of Nakatani Shuyo [9], which contains character n-gram models of 53 languages trained on Wikipedia abstracts, to create a, possibly noisy, but large and training set of English and non-English text based on tweets. The content of each tweet was first lowercased and then passed through a whitespace tokenizer. A

[8] `https://dev.twitter.com/docs/streaming-api`
[9] `http://code.google.com/p/language-detection/`

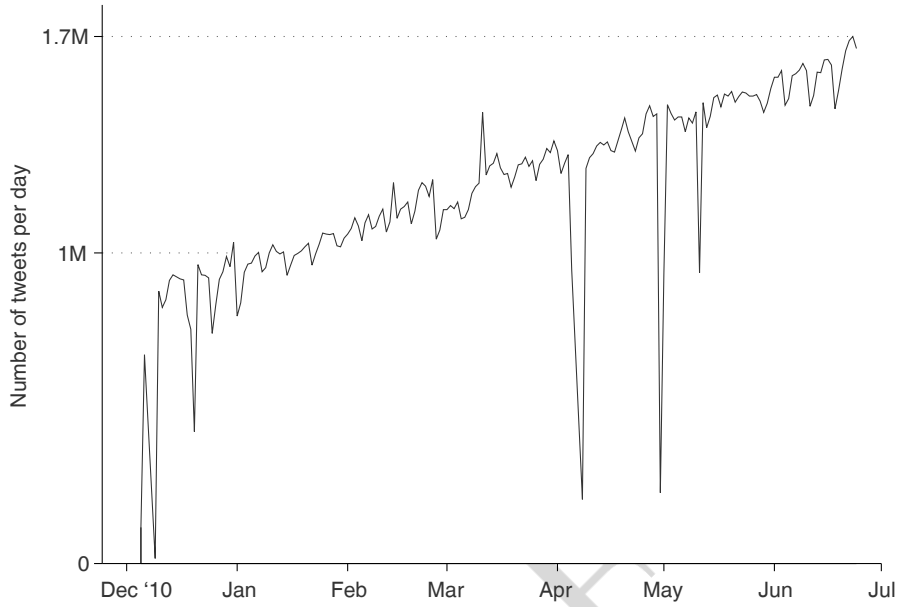

Figure 1: Number of tweets sampled per day from the Twitter Streaming API. Sharp drops correspond to periods of outages in the network or Twitter API, during which limited data were collected.

DRAFT

token was removed from the content if it satisfied at least one of the following conditions:

- is longer than 50 characters,
- is a *hashtag* or a *user mention* (starts with "#" or "@" respectively),
- is a URL (starts with "http://" or "https://"),
- contains more than four consecutive repetitions of the same character,
- is the special token "rt" which denotes a *retweet*.

If the remaining content was longer than 100 characters (remember that a tweet can have at most 140-characters) then its language was identified by LDL (based on the reduced content). This procedure makes sure that only long tweets with linguistically meaningful content are labeled with a language. I used 15 languages[10] and used the "other" label for all other languages as a catch-all category. This step resulted in approximately 13 million English and 12 million non-English tweets.

[10]English, Italian, German, Portuguese, Spanish, French, Turkish, Indonesian, Japanese, Dutch, Korean, Arabic, Russian, and Thai

In the second stage of the bootstrapping, I used LingPipe (Alias-i, 2008), a natural language processing library, to train character n-gram models for different languages with Witten-Bell smoothing (Witten and Bell, 1991) (with $n \leq 4$). Details of the implementation can be found on LingPipe's Web page[11] and in Carpenter (2005). I used all of the tweets that were language-labeled in the previous step with the exception of English, Indonesian, Spanish and Portuguese tweets for which the training sets were limited to the first 1,000,000 characters for each language, due to memory limitations. Note that during the training phase, the entire content of the tweets are used – not just the reduced contents that were utilized in the first step. This way, stylistic and non-standard usages specific to each language is also employed in the training procedure. After the second step of the bootstrapping procedure, I evaluated all 239 million tweets by the trained models and labeled each tweet with the most-likely language according to the n-gram models. This last step resulted in 87 million English tweets. Note that, even though the language models are trained on lowercased text to reduce noise, the tweets are kept in their original casing.

In order to evaluate the performance of language identification, I manually labeled 250 randomly chosen tweets as English or non-English (without knowing their assigned labels). Multi-lingual or ambiguous tweets were labeled as non-English. The precision and recall values of bootstrapped language identifier are given in Table 1 along with the performance measures of LDL as baseline. Bootstrapping improved both the recall (by 8.74 percentage points) and the precision (by 3.22 percentage points). The two classifiers disagreed on 18 tweets, of which 15 were correctly labeled by the bootstrapped classifier ($p = 0.008$ for a two-tailed binomial test with equal proportions).

Measure	LDL	Bootstrapped
True Positive	91	100
True Negative	143	146
False Positive	4	1
False Negative	12	3
Precision	95.79% (89.67%–98.35%)	99.01% (94.60%–99.83%)
Recall	88.35% (80.73%–93.21%)	97.09% (91.78%–99.00%)

Table 1: Performance measures of LDL and bootstrapped classifier in identifying English tweets. Ranges in parentheses are 95% confidence intervals (Wilson score intervals).

Post-fetch cleaning—We want our corpus to consist of the tweets of real people. To achieve this aim, spam and bot accounts should be identified and removed from the dataset. In Twitter, spam is a serious concern and spam filtering is not a trivial task (Yardi et al, 2010; F. Benevenuto et al, 2010; Kyumin

[11] `http://alias-i.com/lingpipe/`

Lee et al, 2010). Twitter itself is continuously fighting with spam[12] and removing accounts which post unsolicited URLs or engage in spamming activity. In order to remove spam from the corpus, I opted to make use of Twitter's efforts: Three and a half months after the fetching process was completed, I re-queried the Twitter API to see which accounts in the dataset were still active. I removed all accounts which were deleted in the meantime. The trade-off that needs to be addressed here is between precision and recall. Discarding all deleted accounts will possibly result in removing valid tweets (i.e., because of the accounts which were deleted for reasons other than being spam). Given the virtually unlimited supply of tweets (the real limitation is not the supply of tweets, but our ability to handle them), current technique favors precision over recall.

The second issue is bots – Twitter accounts managed by software systems that post tweets automatically and regularly. They do not necessarily engage in spamming or malicious activities (e.g., a bot may simply tweet humorous texts taken from an outside source several times a day, or the weather condition every hour), but still, they may distort the corpus content because of the sheer number of the tweets they post. Identification of bots in Twitter is an active area of research (Chu et al, 2010). Here, I used a tweet-count heuristic to identify the – inhumanly – prolific accounts and discard their tweets. As already explained, the fetching process continued for approximately 200 days and it consisted of a 1% sample of all tweets. A user who tweeted 100 times a day, for every day during the 200-day sampling period, would have posted 20,000 tweets during the period, and the expected number of tweets we would see in our sample would be 200. Thus, I set the threshold to 200, and all accounts with more than 200 tweets in the entire sample were removed from the dataset.

Spam and bot removal heuristics filtered 1,797,699 accounts. After filtering, I was left with 74,579,965 English tweets posted by 11,434,388 unique users which constitute the basis of rest of the analysis in this study.

Tokenization—Tweets contain non-standard lexical entities that we want to protect in the tokenization process, such as hashtags (i.e., keywords prefixed by '#', acting as user-provided tags for the tweet) and user name mentions (e.g., *@username*). In order to preserve the original content as much as possible, I employed a minimalistic tokenizer which detects and preserves URLs, abbreviations, emails, HTML entities (e.g., "), hashtags, user name mentions and emoticons (e.g., ":-)", ":["). All other punctuation marks are considered as single tokens.

An example content and its tokenization is given below:

original *RT @justinbieber: and that's for those that dont know...they've great records like this. #GREATMUSIC: http://www.youtube.com/watch?v=cF*

tokenized RT @justinbieber : and that ' s for those that dont know . . . they ' ve great records like this . #GREATMUSIC : http://www.youtube.com/watch?v=cF

Extracting gender and timezone—We want to tag each n-gram with the gender of the author who wrote the tweet that contained the n-gram and the time of posting. In this section, I will explain and evaluate the algorithms that

[12] `http://blog.twitter.com/2010/03/state-of-twitter-spam.html`

are used to extract this information. The statistics of the extracted data are given in the next section. Gender and local timezone were extracted based on the profile information of a user which might have changed during the fetching period. I re-computed these tags for each tweet separately. Thus, if a user changed his timezone during the fetching period, the dataset captures this change. The user-based statistics are based on the first computed tag for the user.

Twitter does not collect or disclose the gender of its users. However, user-provided first names can be used to guess the gender of users – at least, for the ones that provide one. For this purpose, I used census statistics including name frequencies for both genders, released by U.S. Census Bureau and U.S. Social Security Administration[13]. If a name is frequently associated with males (females) all users having this first name are classified as male (female). The scripts used are available publicly[14].

It is possible to evaluate the method's performance on a sample of people with known genders. Unfortunately, we lack a gold standard dataset for Twitter users. However, Freebase[15], a large, collaborative and structured database, contains demographics (e.g., full names, birthdays, and genders) of more than two million people – who are, in one way or another, distinguished individuals and whose data are collected from various sources on the web and contributed by Freebase's users. In order to create a gold standard, I sampled all U.S.-born people whose genders were specified in the database and who were born after Jan. 1, 1900. This resulted in a set of 91,520 people. The first-name based heuristic could not provide a guess for the gender of 9762 people (10.70%). Among the remaining people, the accuracy was 97.85%, with approximately symmetric error rates for either gender: 2.94% of the females and 2.35% of the males were misclassified.

This sample of users is not a perfect substitute for a gold standard collected from Twitter. Nevertheless, previous studies report additional sanity checks which suggest the method works quite reasonably (Mislove et al, 2011; Herdağdelen and Baroni, 2011), and the above results provide further assurance that name-based heuristics have a very high accuracy with reasonable coverage.

The posting time of tweets are given in Coordinated Universal Time (UTC) which need to be converted to the local time relative to the user. 65% of the users (accounting for 78% of all English tweets) identified a timezone in their profile settings, and the local time of the tweets were computed by correcting the time in UTC. This process may introduce some noise for reasons such as inaccurate timezones provided by people or user mobility. Nevertheless, previous studies suggest that user-provided timezones are accurate enough to compute local posting times in more detailed analyses (Golder and Macy, 2011).

By using the local time of tweets, I computed two tags: the day of the week (monday, ..., sunday), and hour of the day (0, ..., 23) for given a tweet in the

[13] Note that the gender information of individual users are intentionally lost during the n-gram aggregation process; thus the predicted genders of users are not disclosed in the dataset.

[14] `https://github.com/amacinho/Name-Gender-Guesser`

[15] `http://www.freebase.com/`

posting user's local timezone. The n-grams occurring in the tweet are tagged with the corresponding day and hour labels. The n-grams of the users without a timezone are labeled with *unknown-timezone*.

Aggregation and Distribution Format—Two kinds of statistics are reported for the n-grams – frequency and user count. Frequency is the total number of times the n-gram is mentioned in the corpus; user count is the total number of users who mentioned the n-gram at least once. Each tokenized tweet in the dataset was assigned to a slot which corresponds to a tuple consisting of gender, day of the week, and hour of the day, depending on the gender of the user and the local time of posting. The tweets in the same slots were aggregated and the number of times each n-gram was observed was counted along with the number of unique users who mention the n-gram for a given slot. In order to keep the corpus size at manageable levels, n-grams which were observed less than 3 times in the entire dataset were removed from the dataset.

There are seven days a week, 24 hours a day, two genders, and two statistics (frequency and user count). Adding the unknown tag (which is used when the timezone or gender is not specified), this translates into $(7+1)*(24+1)*(2+1)*2 = 1200$ slots. There is a small redundancy in this representation, which I decided to keep for consistency: if the day of the week is not known – because the corresponding users do not have a timezone specified in their profiles – also the hour of the day is not known.

For practical purposes, the corpus is released as six separate files, one for each n value of the n-grams (from 1 to 6). Each n-gram is represented on a separate line. Each line is a tab-separated list of columns where the first column is the n-gram (with space between the tokens). The remaining 1200 columns on each line are the frequency and user-count statistics of the n-gram for every combination of gender, day-of-week, and hour-of-day slots.

A truncated example is given in Table 2. In this table, each column corresponds to a tab-separated field in the corpus files. The slots iterate over frequency, user count, gender, hour, and day in the given order with the "unknown" tags placed at the end of the iteration for each variable. Details are given in the corpus homepage[16]

n-gram	monday-0-female-freq	monday-0-female-user	monday-0-male-freq	...
car	164	159	174	...

Table 2: Example line from a corpus file. Each data field corresponds to a (day, hour, gender) tuple and for each tuple two numbers are reported: total number of frequency and number of unique users. The values are aggregated over all tweets which fall under the given slot.

[16] `http://clic.cimec.unitn.it/amac/twitter_ngram`

4 Analysis and Comparison

In this section, I will first report the gender and time-based analysis of Rovereto Twitter N-Gram Corpus, and then compare it to a web-based corpus.

Gender—The gender guessing heuristic, when applied to the 11,434,388 users represented in RTC, labels 28.79% of the users as female and 26.78% as male (for the remaining 44.43% no gender is guessed and they are labeled as *unknown*. The tweets of these users are labeled with the gender of their authors, accordingly. The details of the gender tagging are given in Table 3.

	Female	Male	Unknown	Total
Users	3,291,849	3,062,113	5,080,426	11,434,388
Tweets	21,179,637	17,190,618	36,209,710	74,579,965
Tokens	337,225,305	268,143,434	569,226,268	1,174,595,007

Table 3: Size of English tweets dataset, broken down by gender.

	Female	Male	Unknown	Total
1-grams	3,505,120	3,305,252	4,344,364	4,874,312
2-grams	14,389,583	13,896,057	18,157,254	19,746,857
3-grams	26,234,119	24,479,613	32,464,761	35,377,485
4-grams	26,669,440	23,485,753	32,420,228	35,836,687
5-grams	20,635,630	17,240,092	25,008,601	27,991,244
6-grams	7,931,888	6,605,643	8,810,292	9,193,536

Table 4: Number of types (surface forms), broken down by gender.

Time—In Figure 2, we can see the hourly change in Twitter over the course of week – aggregated over all weeks in the dataset – in terms of total count of unigrams written per hour. There is a clear daily pattern: In the first four working days (from mondays to thursdays) the activity is at its lowest at around 04:00 in the morning. Then, we see an uninterrupted increase in the number of tokens until 20:00 in the evenings. Fridays start similar to the other working days but friday evenings diverge from the regular pattern. We see that the friday evening peak is much less pronounced. Saturdays are marked with a generally reduced activity compared to all other days. On sundays, the regular pattern resumes with an almost identical shape to those of the working days.

We can also cross-tabulate the gender and time tags to compare the relative frequency of male and female activity on Twitter for different times during the week. In Figure 3, we can see the fluctuations in female and male activity in Twitter, in terms of percentage points deviated from the expected baseline based on total number of tokens written by females and males (which is 56% for females and 44% for males). In the figure, this baseline is represented as the 0% line. For example, right after 04:00 o'clock on sunday mornings, the number

DRAFT

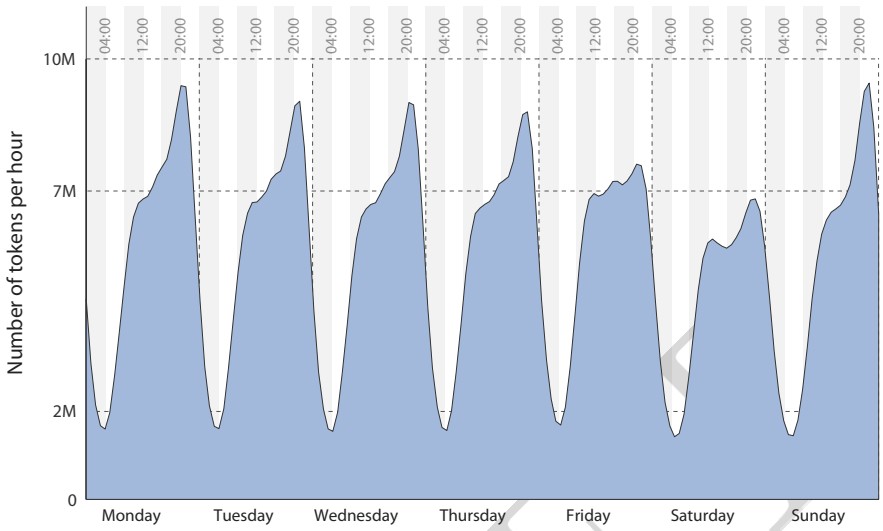

Figure 2: Total number of tokens observed in RTC, broken down by the hour the week. Each hourly data point is computed by aggregating all tweets posted in the corresponding hour and day of the week, collected during the 7-month fetching period. Only the tweets of users whose local times are given in their profiles are used.

of tokens written by males deviate from the baseline more than 5 percentage points, which means the percentage of tokens written by males is appproximately 49% at that time. The deviations may seem small, but there is a remarkably consistent, bi-modal, daily pattern for all days of the week: increased male activity in the early morning and evening, and increased female activity during midnight and midday. In addition to this pattern, during the weekends, the overall male activity increases substantially.

Comparison with ukWaC—A comparison of Rovereto Twitter N-Gram Corpus (RTC) with a web-based corpus like ukWaC will help to better understand the particular nature of Twitter-based content. In this section, I will treat ukWaC as a baseline corpus and report the differences of RTC from this baseline. ukWaC is a web-based corpus that contains approximately two billion tokens extracted from web pages in the uk domain. It is comparable to RTC in size. Its web-based nature makes ukWaC an ideal candidate as a baseline corpus because, just like social media, the web provides an almost unlimited amount of textual data and it is important to understand the differences between the web and the social media as two sources for language-related data.

As a first step, I computed the contingency table for the unigrams occurring in at least one corpus. To simplify the matters, I lowercased the unigrams, aggregated the frequencies of duplicate unigrams (occurring because of lowercasing), and limited the analyses to tokens which contain only letters from the

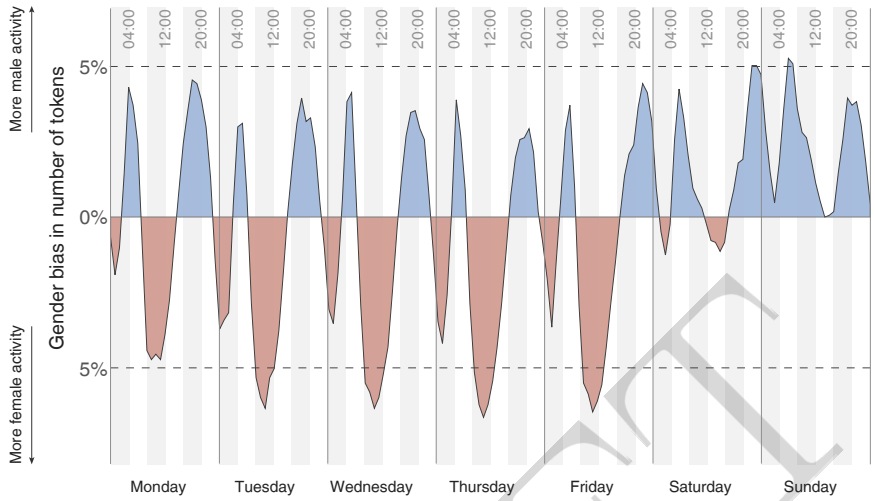

Figure 3: The ratio of tokens written by females and males, deviating from the overall gender ratio. Each data point is computed by comparing the total number of tokens written by either gender during the corresponding hour and day of the week, aggregated for all tweets collected in the 7-month fetching period. Only the tweets of users whose local times and genders are guessed are used.

English alphabet (without any lemmatization). The resulting set is used also for the rest of the analyses in this section. In Table 5, we see the 3 by 3 contingency table for the unigrams occurring in at least one corpus. The unigrams are divided into three sets for each corpus according to their frequency: a word is either not occurring in a corpus, occurring at least once but less than 30 times in total, or greater than or equal to 30 times. The threshold 30 is picked arbitrarily, but should provide a more detailed view than simply dividing the tokens into two sets as observed or not observed.

		ukWaC		
		$= 0$	$0 < < 30$	≥ 30
RTC	$= 0$	-	322,482	35,942
	$0 < < 30$	1,147,185	156,487	44,673
	≥ 30	28,654	38,143	78,798

Table 5: Contingency table for the unigrams occurring in at least one corpus.

We see that the overlap between the two corpora is rather limited. Out of the 226,610 unigrams which occur at least 30 times in RTC or ukWaC, 161,614 (71%)

DRAFT

are observed in both corpora. Furthermore, more than one million unigrams which are observed in RTC are not observed in ukWaC at all.

In Table 6, we see the most frequent 20 unigrams in RTC and ukWaC, which are not observed in the other corpus. The Twitter-specific unigrams are mostly online services or Twitter-related terms (retweet, twitpic, tweeting, tweeps, tweeted, unfollow, twitcon, tumblr), artists or artist-related slang (belieber, minaj), or interjections/acronyms that are frequently found in the so called textspeak or other short messagaging conventions (lmaoo, ctfu, bouta, shawty, yhu, lmfaoo, lmfaooo, smdh, hahaa). The ukWaC-specific terms include acronyms (mostly governmental departments or organizations in the U.K.), and legal terms such as appelants, rateable, pursuance, and senatus.

Corpus	Unigrams
RTC	retweet, tumblr, lmaoo, tweeted, tweeps, ctfu, beliebers, unfollow, bouta, shawty, yhu, minaj, belieber, twitcon, lmfaoo, lmfaooo, smdh, tweetin, hahaa, twitpic
ukWaC	dfes, odpm, hmso, coddan, deweyclass, hmso, ltsn, usdaw, ukl, rateable, appellants, dfee, eevl, ctbt, tripos, oftel, kiania, ahds, pursuance, senatus, lgps

Table 6: Most frequent 20 unigrams in RTC and ukWaC, which are not observed in the other corpus respectively.

In Figure 4, the relative frequency of personal pronouns is visualized. Note that the counting is done over the surface forms without any part-of-speech (POS) tagging. The first-person pronouns "I", "me" and "my" are clearly over-represented in RTC – approximately 90% of all occurrences of these three pronouns are observed in RTC as opposed to ukWaC. This reflects the first-person narrative style of the tweets. People are predominantly talking about themselves.

I also computed the relative frequency of all unigrams in RTC and ukWaC, and identified the unigrams that are most over-represented in either corpora (i.e., the unigrams that are disproportionately used more often in one of the corpora). For illustrative purposes, I grouped the unigrams according to their most-assigned POS tags in ukWaC. In Table 7, we see the most-overrepresented 10 nouns and 10 verbs in RTC and ukWaC. In RTC, we see nouns and verbs that express emotions and tastes such as love, hate, smile, excited, and hurt. In addition, temporal references such as "tonight" and "tomorrow" reflect the immediate nature of Twitter posts. It is harder to categorize the over-represented unigrams in ukWaC (though, it looks like they come from more formal documents). However, a striking pair of unigrams is "october" and "november", two months that fell outside of the fetching period of the RTC corpus. This suggests that one must be careful about the temporal idiosyncrasies that may be present in RTC and similar corpora.

DRAFT

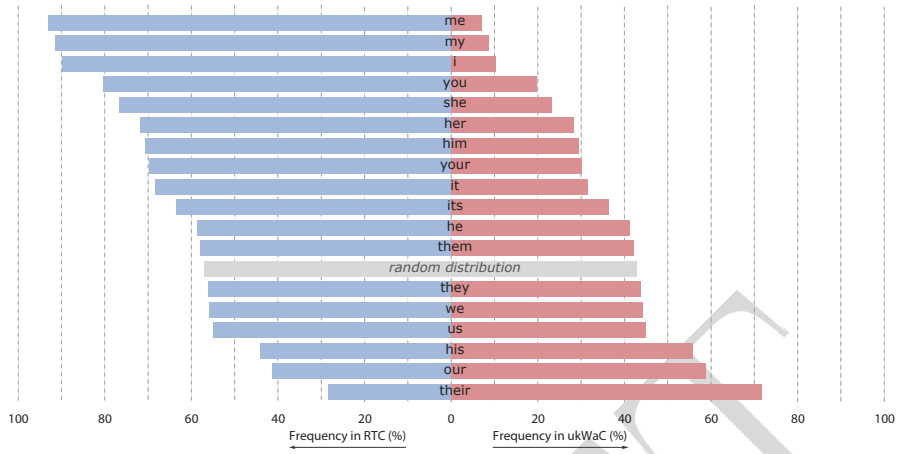

Figure 4: Relative frequencies of personal pronouns in Twitter-based RTC and web-based ukWaC. For example, more than 90% of all occurences of the pronoun "me" are observed in RTC, more than 70% of all occurrences of "their" are observed in ukWaC. The random distribution – which corresponds to the no-bias case – is based on the total token counts in the two corpora.

5 Case Study: Stereotypical Gender Expectations

In this section, I describe a previous study which used Edinburgh Twitter Corpus – a Twitter-based corpus with a comparable size to our new corpus, but not available anymore – to predict the stereotypical gender expectations of actions, and replicate the findings of the original study with the new Rovereto Twitter N-Gram Corpus (RTC).

Stereotypical gender expectations play an important role in our daily and social lives whether we agree with or are simply aware of them. In Herdağdelen and Baroni (2011), we argued that in artificial intelligence (AI) applications, such knowledge should be explicitly marked and handled accordingly. We stated "whether it's right or wrong, an AI should know that (we expect that) *women like shopping* and *men like football*." (emphasis in original). If stereotypical knowledge is made explicit in knowledge repositories – which AI systems rely on – AI developers will have the opportunity to disregard such knowledge, or make use of them in order to make AI systems relate to human beings.

In Herdağdelen and Baroni (2011), we focused on daily activities of people and propose to use corpus analysis to predict the stereotypical expectations about these activities. We used two corpora, Edinburgh Twitter Corpus and ukWaC, to compute the *gender bias* of verb phrases. In the Twitter-based corpus, we guessed the gender of each user by using the same tools I used in this study, and computed the number of times each phrase was mentioned by

DRAFT

Nouns		Verbs	
RTC	ukWaC	RTC	ukWaC
dont (97.92%)	information (95.01%)	don (99.48%)	provide (94.40%)
tonight (97.81%)	council (94.36%)	hate (97.96%)	published (93.00%)
tomorrow (96.49%)	environment (93.67%)	watching (94.86%)	required (92.77%)
girl (95.46%)	range (93.39%)	excited (94.72%)	include (92.53%)
sleep (95.37%)	development (93.13%)	ok (94.41%)	develop (92.50%)
love (95.32%)	centre (93.07%)	follow (94.35%)	including (91.98%)
hell (95.32%)	october (92.77%)	hurt (93.99%)	includes (91.21%)
birthday (94.64%)	department (92.34%)	wait (93.74%)	included (90.79%)
smile (94.40%)	november (92.24%)	guess (93.59%)	designed (89.50%)
somebody (93.18%)	research (92.21%)	loves (93.02%)	shown (88.32%)

Table 7: Most typical (i.e., over-represented) 10 nouns and 10 verbs in Twitter-based RTC and web-based ukWaC (see text for POS tagging). The percentages in parentheses are the proportion of the occurrences of each unigram in the corresponding corpus. For example, 97.92% of all occurrences of "dont" are observed in RTC, and 95.01% of all occurrences of information are observed in ukWaC. Analysis is limited to unigrams which occur at least 10,000 times in both corpora.

a male or female user. In ukWaC, there is no metadata about gender. Instead a pronoun-based heuristic was employed, which looked at the first pronoun or proper name preceding a verb phrase and used its gender to count the male and female occurrences of a verb phrase.

In the Twitter-based gender bias computation, the assumption is that if a certain action is mentioned more often by females it is a more salient action for females (vice versa for males). The ukWaC-based gender bias quantifies which gender is more frequently reported to carry out a given action (e.g., "she became pregnant" would count as a female occurrence). In other words, the RTC provides a measure of "who says what" whereas (web-based) ukWaC provides a measure of "who is reported to be doing what". Below I briefly describe the dataset and methodology in the original study and compare the results with those obtained on RTC with the same methods. Details pertaining to the original study can be found in the original paper (Herdağdelen and Baroni, 2011).

A publicly available commonsense knowledge repository, OMCS (Speer, 2007; Havasi et al, 2009), was used to create a random sample of verb phrases (e.g., become nurse, enjoy power, impress people, feel cold, etc.). The sampled phrases were rated by human judges on a 5-point scale (i.e., typically/slightly feminine/masculine or neutral). Each phrase was rated by at least five judges. The average of the scores of a given phrase was used as its *gender score*. The sampling and annotation process resulted in 441 actions with assigned gender scores. For example, among the phrases the judges deemed the most feminine were "become nurse", "make doll", and "freshen up". The most "masculine" phrases were

DRAFT

"want woman", "catch football", and "see pretty girl". This dataset has been made public[17].

The computation of the gender bias of a given phrase starts with determining the number of its male and female occurrences (for Twitter-based scores this is the frequency of the phrase in male and female users' tweets; for ukWaC-based scores this is the number of times the phrase is preceded by a male or female pronoun or proper name). While looking for the occurrences, both the target verb phrases and the corpora are lemmatized, and at most one intermittent token is allowed between two consecutive lemmas of the phrase (e.g., "become a nurse" is counted as an occurrence of the phrase "become nurse"). Let f and m denote the number female and male occurrences of a verb phrase, respectively. If we assume the phrase is not biased towards either gender, the ratio of male (and female) occurrences should follow a binomial distribution with $n = f + m$ and $p = M/(M + F)$ where F and M are the total female and male occurrences of all verb phrases (i.e., $F = \sum f$ and $M = \sum m$). Significant deviations from the expected value can be attributed to an inherent gender bias. In Herdağdelen and Baroni (2011), we used the standard score (by normal approximation) of male occurrence ratio as the gender score: $(m - pn)/\sigma$ where $\sigma = \sqrt{np(1-p)}$. This notation arbitrarily assigns positive scores to "masculine" phrases and negative scores to "feminine" phrases. A score of 0 corresponds to the gender-neutral case.

I passed all of the n-grams in RTC through TreeTagger for lemmatization[18] (Schmid, 1995) The corpus already contains the number of male and female occurrences of n-grams. I used the same phrase detection heuristics to count the male and female occurrences of the verb phrases occurring in the n-grams. To ensure that an occurrence is not double counted because of overlapping n-grams, I applied an additional constraint that the phrase must fully span the n-gram to be counted as an occurrence (i.e., the n-gram starts and ends with the first and last lemmas of the phrase respectively). Once the female and male occurrence counts were computed, the rest of the gender score computation was carried out as in the original study, described above.

In the original study, Spearman correlations between the computed gender scores and human-based gold standard are reported along with the accuracy of the gender scores in predicting the polarity of the gender bias (whether it is masculine or feminine). In Table 8, the Spearman correlations reported in the original study and the correlations obtained with the new RTC are given. In "matching" cases the Spearman correlation is computed for the verb phrases for which both ukWaC and the Twitter-based corpora have the same gender polarity. The results based on Edinburgh Twitter Corpus – which contains the full texts of the tweets – and RTC are virtually indistinguishable, and the coverage of RTC is almost as high as the former. Note that combining ukWaC-based scores and Twitter-based scores substantially increases the correlation, in both cases – with a penalty of coverage.

[17] `https://github.com/amacinho/Gender-Expectations`

[18] The limitation of n-grams instead of full sentences may introduce errors in lemmatization

DRAFT

Method	Spearman	Coverage
ukWaC[1]	0.27	98% (433)
ETC[1]	0.28	100% (441)
matching ukWaC and ETC[1]	0.47	52% (231)
RTC[2]	0.26	99% (438)
matching ukWaC and RTC[2]	0.46	49% (217)

Table 8: Spearman correlations between predicted gender scores and human annotations, obtained in 1) Herdağdelen and Baroni (2011) and 2) this study. Coverage is the percentage and number of phrases observed in the corresponding method and corpora. All reported correlations are statistically significant for $p < 0.001$.

In Table 9, area under the ROC curve (AUC) and accuracy values are reported. In this evaluation, neutral verb phrases (those with a gender bias of 0) are left out. The results based on the binary classification of the directionality of the gender bias are consistent with the correlation-based results. Again, results based on Rovereto Twitter N-Gram Corpus are almost the same with the original results and the coverage is scarcely affected because of the n-gram representation.

Method	AUC	Accuracy	Coverage
ukWaC (375)	0.64	0.61	99% (375)
ETC (380)	0.65	0.61	100% (380)
matching ukWaC & ETC (205)	0.76	0.70	54% (205)
RTC	0.66	0.58	99% (378)
matching ukWaC and RTC	0.76	0.69	51%(194)

Table 9: Accuracy and area under the ROC curves for predicting the directionality of the gender bias. Both for AUC and the accuracy, the random baseline is 0.50.

DRAFT

6 Conclusions and Future Work

In this study, a large, publicly available dataset of English tweets with the gender-of-the-author and time-of-posting tags is introduced. The methods employed for spam removal, gender guessing, local time extraction, and language detection are described and relevant code is shared. The descriptive statistics and a preliminary analysis of gender- and time-based activity in Twitter reveals interesting patterns that deserve to be studied in detail. A previous

process, mainly because of the ambiguities in part of speech.

gender-based research based on Twitter text is replicated with the new data with almost identical results.

The corpus contains only n-gram statistics as opposed to full text of the tweets. This limitation is not by choice, but a necessity because of the restrictions on redistributing Twitter-based data. The corpus lacks lemmatization and POS tags. Lemmatization and POS tagging are challenging tasks on Twitter-based text because of the informal and noisy nature of the data. However, there are promising attempts to train Twitter-specific POS taggers (Gimpel et al, 2011), and in the future, I plan to evaluate and apply a suitable part-of-speech tagging method on the corpus. Also, releasing a multi-lingual version of the corpus is in future plans.

I believe Rovereto Twitter N-Gram Corpus will be useful for researchers, freeing them from the burden of having to curate a new dataset from scratch, and for the research community in general, providing them a shared dataset on which different methods and techniques can be compared to each other.

7 Acknowledgments

I thank Marco Baroni and Eser Aygün for their valuable comments on a draft of this manuscript.

References

Alias-i (2008) LingPipe 4.1.0. URL `http://alias-i.com/lingpipe`

Argamon S, Koppel M, Pennebaker J, Schler J (2007) Mining the blogosphere: age, gender, and the varieties of self-expression. First Monday 12(9)

Baroni M, Bernardini S, Ferraresi A, Zanchetta E (2009) The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. Language Resources and Evaluation 43(3):209–226

Bifet A, Frank E (2010) Sentiment Knowledge Discovery in Twitter Streaming Data. In: Pfahringer B, Holmes G, Hoffmann A (eds) Discovery Science, Lecture Notes in Computer Science, vol 6332, Springer Berlin Heidelberg, Berlin, Heidelberg, pp 1–15, DOI 10.1007/978-3-642-16184-1, URL `http://www.springerlink.com/content/m778071255jx3223/`

Burton K, Kasch N, Soboroff I (2011) The ICWSM 2011 Spinn3r Dataset. In: Proceedings of the Fifth Annual Conference on Weblogs and Social Media (ICWSM 2011)

Carpenter B (2005) Scaling high-order character language models to gigabytes. In: ACL Software Workshop, Association for Computational Linguistics, pp 86–99, DOI 10.3115/1626315.1626322, URL `http://portal.acm.org/citation.cfm?id=1626322&dl=GUIDE,`

Chu Z, Gianvecchio S, Wang H, Jajodia S (2010) Who is tweeting on Twitter: human, bot, or cyborg? In: Proceedings of the 26th Annual Computer Security Applications Conference, ACM, Austin, Texas, pp 21—-30

Conover MD, Ratkiewicz J, Francisco M, Gonçalves B, Flammini A, Menczer F (2011) Political Polarization on Twitter. In: Proc. 5th Intl. Conference on Weblogs and Social Media

Culotta A (2010) Towards detecting influenza epidemics by analyzing Twitter messages. KDD Workshop on Social Media Analytics pp 115–122, DOI 10.1145/1964858.1964874, URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.167.5375&rep=rep1&type=pdf

Cunha E, Magno G, Comarela G, Almeida V, Gonçalves MA, Benevenuto F (2011) Analyzing the Dynamic Evolution of Hashtags on Twitter: a Language-Based Approach. ACL HLT 2011 p 58

Davidov D, Tsur O, Rappoport A (2010) Semi-Supervised Recognition of Sarcastic Sentences in Twitter and Amazon. In: CoNLL '10 Proceedings of the Fourteenth Conference on Computational Natural Language Learning, Association for Computational Linguistics, July, pp 107–116, URL http://portal.acm.org/citation.cfm?id=1870582

F Benevenuto, G Magno, T Rodrigues, V Almeida (2010) Detecting Spammers on Twitter. In: Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)

Gimpel K, Schneider N, O'Connor B, Das D, Mills D, Eisenstein J, Heilman M, Yogatama D, Flanigan J, Smith NA (2011) Part-of-speech tagging for Twitter: annotation, features, and experiments. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics, Portland, pp 42–47, URL http://dl.acm.org/citation.cfm?id=2002736.2002747

Golder SA, Macy MW (2011) Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures. Science 333(6051):1878–1881, DOI 10.1126/science.1202775, URL http://www.sciencemag.org/content/333/6051/1878.abstract

González-Ibáñez R, Muresan S, Wacholder N (2011) Identifying sarcasm in Twitter: a closer look. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Portland, Oregon, USA, pp 581–586, URL http://dl.acm.org/citation.cfm?id=2002736.2002850

Havasi C, Speer R, Pustejovsky J, Lieberman H (2009) Digital intuition: Applying common sense using dimensionality reduction. IEEE Intelligent Systems 24(4):24–35

DRAFT

Herdağdelen A, Baroni M (2011) Stereotypical gender actions can be extracted from web text. Journal of the American Society for Information Science and Technology 62(9):1741–1749, DOI 10.1002/asi.21579, URL http://doi.wiley.com/10.1002/asi.21579http://onlinelibrary.wiley.com/doi/10.1002/asi.21579/full

Hong L, Convertino G, Chi EH (2011) Language Matters in Twitter : A Large Scale Study Characterizing the Top Languages in Twitter Characterizing Differences across Languages Including URLs and Hashtags. Artificial Intelligence (1):518–521

Hundt M, Nesselhauf N, Biewer C (2007) Corpus Linguistics and the Web. Rodopi

Kelly R (2009) Twitter Study. URL http://www.pearanalytics.com/blog/wp-content/uploads/2010/05/Twitter-Study-August-2009.pdf

Kyumin Lee, James Caverlee, Steve Webb (2010) Uncovering social spammers: social honeypots + machine learning. In: Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval, ACM, Geneva, Switzerland, pp 435—-442, URL http://doi.acm.org/10.1145/1835449.1835522

Lazer D, Pentland AS, Adamic L, Aral S, Barabasi AL, Brewer D, Christakis N, Contractor N, Fowler J, Gutmann M, Others (2009) Life in the network: the coming age of computational social science. Science (New York, NY) 323(5915):721

Lerman K, Ghosh R (2010) Information contagion: An empirical study of the spread of news on Digg and Twitter social networks. In: Proceedings of 4th International Conference on Weblogs and Social Media (ICWSM)

Liu F, Weng F, Wang B, Liu Y (2011) Insertion, deletion, or substitution?: normalizing text messages without pre-categorization nor supervision. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Portland, Oregon, USA, pp 71–76, URL http://dl.acm.org/citation.cfm?id=2002736.2002753

Madden M, Zickuhr K (2011) 65 Tech. rep., Pew Research Center, URL http://pewinternet.org/Reports/2011/Social-Networking-Sites.aspx

Michel JB, Shen YK, Aiden AP, Veres A, Gray MK, Pickett JP, Hoiberg D, Clancy D, Norvig P, Orwant J, Others (2011) Quantitative analysis of culture using millions of digitized books. science 331(6014):176

Mislove A, Lehmann S, Ahn YY, Onnela JP, Rosenquist JN (2011) Understanding the Demographics of Twitter Users. In: Fifth International AAAI Conference on Weblogs and Social Media, AAAI, pp 554–557, URL http://www.ccs.neu.edu/home/amislove/publications/Twitter-ICWSM.pdf

DRAFT

Naaman M, Boase J, Lai CH (2010) Is it really about me?: message content in social awareness streams. In: Proceedings of the 2010 ACM conference on Computer supported cooperative work, ACM, New York, NY, USA, CSCW '10, pp 189–192, DOI http://doi.acm.org/10.1145/1718918.1718953, URL `http://doi.acm.org/10.1145/1718918.1718953`

Naveed N, Gottron T, Kunegis J, Alhadi AC (2011) Bad News Travel Fast: A Content-based Analysis of Interestingness on Twitter. In: ACM WebSci'11,, Koblenz, Germany, pp 1–7, URL `http://journal.webscience.org/435/2/Fig_1.png`

O'Connor B, Eisenstein J, Xing EP, Smith NA (2010) A mixture model of demographic lexical variation. In: Proceedings of NIPS Workshop on Machine Learning in Computational Social Science

Osborne M (????)

Pak A, Paroubek P (2010) Twitter based system: Using Twitter for disambiguating sentiment ambiguous adjectives. In: SemEval '10 Proceedings of the 5th International Workshop on Semantic Evaluation, Association for Computational Linguistics, pp 436–439, URL `http://dl.acm.org/citation.cfm?id=1859664.1859761`

Petrović S, Osborne M, Lavrenko V (2010) The Edinburgh Twitter corpus. In: Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media, Association for Computational Linguistics, pp 25–26

Ritter A, Clark S, Mausam, Etzioni O (2011) Extracting a Calendar From Twitter. In: In Submission

Sakaki T, Okazaki M, Matsuo Y (2010) Earthquake shakes Twitter users: real-time event detection by social sensors. In: Proceedings of the 19th international conference on World wide web, ACM, pp 851–860

Schmid H (1995) Improvements in part-of-speech tagging with an application to German. In: Proceedings of the EACL-SIGDAT Workshop, Dublin, Ireland

Speer R (2007) Open Mind Commons: An inquisitive approach to learning common sense. In: Proceedings of the Workshop on Common Sense and Intelligent User Interfaces, Honolulu, HI

Thelwall M, Buckley K, Paltoglou G (2010) Sentiment in Twitter events. Journal of the American Society for Information Science and Technology

Tumasjan A, Sprenger TO, Sandner PG, Welpe IM (2010) Predicting elections with twitter: What 140 characters reveal about political sentiment. In: Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, pp 178–185

DRAFT

Witten I, Bell T (1991) The zero-frequency problem: estimating the probabilities of novel events in adaptive text compression. IEEE Transactions on Information Theory 37(4):1085–1094, DOI 10.1109/18.87000, URL http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=87000

Yang J, Leskovec J (2011) Patterns of temporal variation in online media. In: Proceedings of the fourth ACM international conference on Web search and data mining - WSDM '11, ACM Press, New York, New York, USA, p 177, DOI 10.1145/1935826.1935863, URL http://dl.acm.org/citation.cfm?id=1935826.1935863

Yano T, Cohen WW, Smith NA (2009) Predicting response to political blog posts with topic models. In: Proceedings of Human Language Technologies The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics on NAACL 09, Association for Computational Linguistics, June, p 477, DOI 10.3115/1620754.1620824, URL http://portal.acm.org/citation.cfm?doid=1620754.1620824

Yardi S, Romero D, Schoenebeck G, Boyd D (2010) Detecting spam in a Twitter network. First Monday 15(1):1–13, URL http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2793/2431

Zhao WX, Jiang J, Weng J, He J, Lim EP, Yan H, Li X (2011) Comparing twitter and traditional media using topic models. In: ECIR'11 Proceedings of the 33rd European conference on Advances in information retrieval, pp 338–349, URL http://dl.acm.org/citation.cfm?id=1996889.1996934

DRAFT