**User Review Sites as a Resource for Large-Scale Sociolinguistic Studies**

**Introduction**

Sociolinguistics is concerned with language in social and cultural context, especially how people with different social identities (e.g. gender, age, race, ethnicity, class) speak and how their speech changes in different situations.

**Issues with traditional studies**

* Small sample size
* Study relied on individual interview, questionnaires and manual transcripts
* the resulting corpora lacked statistical power to establish relationships between language use and socio-economic variables.

**Proposed Solution**

* Web data (large amount of text data)
* Lot of web data is personalized text, such as blogs and other social media texts i.e., they represent actual language use.
* User review sites
* these sites contain extra-linguistic information
* Contains upto millions of user reviews.
* The language in the reviews, not as informal as e.g., Twitter, is much less canonical than newswire, and thus likely to reflect the socio-economic background of the user.

**Trust Pilot**

* Publishes reviews for online businesses.
* Reliable information about the time the review was posted.
* Users often supply information about their location, age, and gender.
* Collected data spans over a period of seven years.

**Data Augmentation**

The retrieved data set is augmented in two ways, by adding

1. Gender information based on first names, and

2. Geotagging information (latitude and longitude).

Problems

Lack of gender info: handled by propagating existing data with a gender ratio of at least 0.95 and a minimum frequency of 3. Gender of certain names is country-specific.

Canonical Towns: For geographical information, the Geonames database is used to attach latitude and longitude information to the user. Latitude and longitude values allow us to precisely place the user review on a map. Various heuristics are applied to get interpret abbreviated location names.

**Representativeness**

* Consider data where the users’ age ranges from 16 to 80.
* The age distributions for both genders follow a reasonable distribution.
* The median age in our data is typically close to the country’s median value.
* Within each country, the gender distributions look similar, but there are always more male than female users.
* Average number of reviews per user is around 4.

**Emoticons, age, and gender**

* Eyes ( : ; ) Nose ( - or none) Mouth ( ( , ) , [ ,\* etc)
* women use emoticons almost twice as often as men do
* for all ages, the use of a nose is highly anti correlated with age

**Ratings, categories, gender, and age**

* men tend to vote slightly more negative than women
* people in the younger group are more likely to use negative ratings than people in the older group

**Denmark**

* missing distinction between the reflexive possessive pronouns - sin/sit (“his/her own”) and non-reflexives - hans/hendes/dens/dets (“his”). i.e., there is no distinction between “He met his (own) wife” and “He met his (=someone else’s) wife”.
* record the frequency of sin/sit (his/her own) and the joint frequency of all possessive pronouns(his). Then compute the ratio of the former in all pronouns.

**German**

* Replacement : β with ss
* dass/daβ, “that", and the modal mussen/muβen, “to must”
* older speakers retain the traditional spelling they acquired in their youth to a much greater extent .

**Conclusion**

* Traditional sociolinguistic studies often lack statistical power to draw valid conclusions, while big-data approaches to language studies mostly lack extra-linguistic information that would enable sociolinguistic studies.
* This paper clearly presents user-review sites as a possible solution to this dilemma. The data provides a combination of non-canonical textual information with meta-information about the authors, including age, gender, and location, as well as time-stamps.

**IDENTIFYING SARCASM IN TWITTER**

**Introduction**

* Sarcasm transforms the polarity of an apparently positive or negative utterance
* into its opposite
* Need for a paper on sarcasm?
* Automatic detection of sarcasm is still in its infancy
* Identify the different type of features(lexical, pragmatic, contextual) that needs to be
* analysed
* Why identifying sarcasm is tough?
* Absence of accurately-labelled naturally occurring utterances that can be used to train
* machine learning systems

**Methodology**

* To build our corpus of sarcastic (S), positive (P) and negative (N) tweets, author

relied on the annotations that tweeters assign to their own tweets using hashtags

Ex: Sarcasm - #sarcasm #sarcastic #happy #joy #angry #frustrated

* Assumption is that the best judge of whether a tweet is intended to be sarcastic is
* the author of the tweet. This is the gold standard

**Data Collection**

* Author used a Twitter API to collect tweets that include hashtags that express sarcasm

(#sarcasm, #sarcastic), direct positive sentiment (e.g., #happy, #joy, #lucky), and direct

negative sentiment (e.g., #sadness, #angry, #frustrated), respectively

* Automatic filters are applied to remove retweets, duplicates, quotes, spam and tweets

written in language other than English

* Automatically filtered all tweets where the hashtags of interest were not located at the

very end of the message

* A manual review of the filtered tweets is done to double check that the remaining end

hashtags were not part of the message

**Data processing**

* From the data, tweets which has the hashtags but not necessarily implying

sarcasm, happy or positive are removed

* Ex: “I really love #sarcasm” is removed from the corpus of tweets
* Final corpus consists of 900 tweets in each of the three categories, sarcastic,

positive and negative

* Emoticons are also used to categorise positive and negative tweets

**Lexical Factors**

* Unigrams
* Dictionary Based
* Use three emotion (Positive, Negative and ToUser)
* We have a very good coverage of 85% between the words in combined dictionary and the words

in the tweets.

**Machine Learning vs Human Performance**

* Humans do not perform better than simple automatic classification methods
* Main issues for the judges
* Lack of context and the brevity of the messages
* Sometimes it was necessary to call on world knowledge such as recent events in order to make

judgments about

* Accurate automatic identification of sarcasm on Twitter requires information about interaction between the tweeters

**Conclusion**

* Accuracy of humans and machine learning classifiers are the same but they are still low
* Signifies the difficulty of sarcasm classification in both humans and machine learning methods
* The length of tweets as well as the lack of explicit context makes this classification task quite

difficult

* In future work, author plans to investigate the impact of contextual features such as common

ground