Votter Corpus: A Corpus of Social Polling Language

Nathan David Green, Septina Dian Larasati
Globeotter
http://globeotter.com/
ndgreen@globeotter.com, larasati@globeotter.com

Abstract

The Votter Corpus is a new annotated corpus of social polling questions and answers. The Votter Corpus is novel in its use of the mobile application format and novel in its coverage of specific demographics. With over 26,000 polls and close to 1 millions votes, the Votter Corpus covers everyday question and answer language, primarily for users who are female and between the ages of 13-24. The corpus is annotated by topic and by popularity of particular answers. The corpus contains many unique characteristics such as emoticons, common mobile misspellings, and images associated with many of the questions. The corpus is a collection of questions and answers from The Votter App on the Android operating system. Data is created solely on this mobile platform which differs from most social media corpora. The Votter Corpus is being made available online in XML format for research and non-commercial use. The Votter android app can be downloaded for free in most android app stores.

Keywords: Social Media, Corpora, Question and Answer, Annotation

1. Introduction

Social media has changed the way a generation communicates, as well as how the language is being used. From SMS text messages on phones, to Twitter (`http://twitter.com`) and Facebook (`http://facebook.com`) updates, all the way to your classroom essay, the language itself has been fundamentally changed. Addressing language changes in Natural Language Processing (NLP), under most situations, means statistically based techniques, which require a massive amount of data from a variety of data sets.

Traditional data sets in NLP typically have been oriented around news, finance, and some sports. It is not news or shocking to anyone in the field that these data sets are inadequate for handling new media types. The languages used in newspapers is far different than that used on mobile phones, social networks, or even in day to day communication.

While there are some corpora for modern media, they typically fall into three categories: Wikipedia (`http://wikipedia.org`), Facebook, or Twitter. Votter (`http://globeotter.com`) differs by these in its use of the medium. The Votter Corpus comes from mobile phones only and is almost entirely question and answer based, often solely based on opinion.

Votter is unique for its short question and answer format. Most questions are opinion based, so it contains different types of information compared to Yahoo Answers (`http://answers.yahoo.com/`) or Quora (`http://quora.com`). Additionally, the system allows images to be associated with the questions, opening up further research avenues for image processing as well. Demographically, Votter gives NLP researches access to a very specific group of mostly female users ranging from age 13-24.

2. Background

The Twitter Corpus out of Edinburgh (Petrović et al., 2010) is a corpus with a similar goal to the Votter Corpus, to capture current language being used. With approximately 100 million tweets, it covers its domain rather well. Votter differs in a few critical aspects. First, Votter is not character limited like Twitter. This means the language is not abbreviated as often. If a word is abbreviated in the Votter Corpus, there is a greater chance that the abbreviation has become a standard. Second, Twitter covers multiple genres of statements and questions. Votter is a very specific domain in questions and answers. The French Social Media Bank (Seddah et al., 2012) is a similar effort to the Twitter corpus. Additionally, they include other user generated texts such as Facebook messages. It is currently being used as a test for parsing and part of speech (POS) accuracy on noisy data. Parsing and POS tagging is a rather early stage process in NLP, much of the intended use of social media for NLP can be seen in higher level applications such as sentiment analysis (Habernal et al., 2013), which has been applied for Czech social media, and opinion mining (Martínez-Cámara et al., 2013). In both of these areas we believe Votter's opinion based question and answer data will be of use.

3. Votter App

The Votter app is a social polling app for the Android operating system and is currently available for all Android phones and tablets. The app allows users to post polls in a number of categories and have other Android users answer those polls. Polls come in a few varieties:

- **Text Poll:** A text question with up to ten text based answers e.g
  - **Q:** "What is your favorite animal?"
  - **A:** 1) Dog 2) Cat 3) Fish ...
- **One Image Poll:** Contains a text question and up to ten text based answers but also includes an image e.g.
  - **Q:** "Do you think my dog is cute?" (an image is included in the poll)
  - **A:** 1) yes 2) no

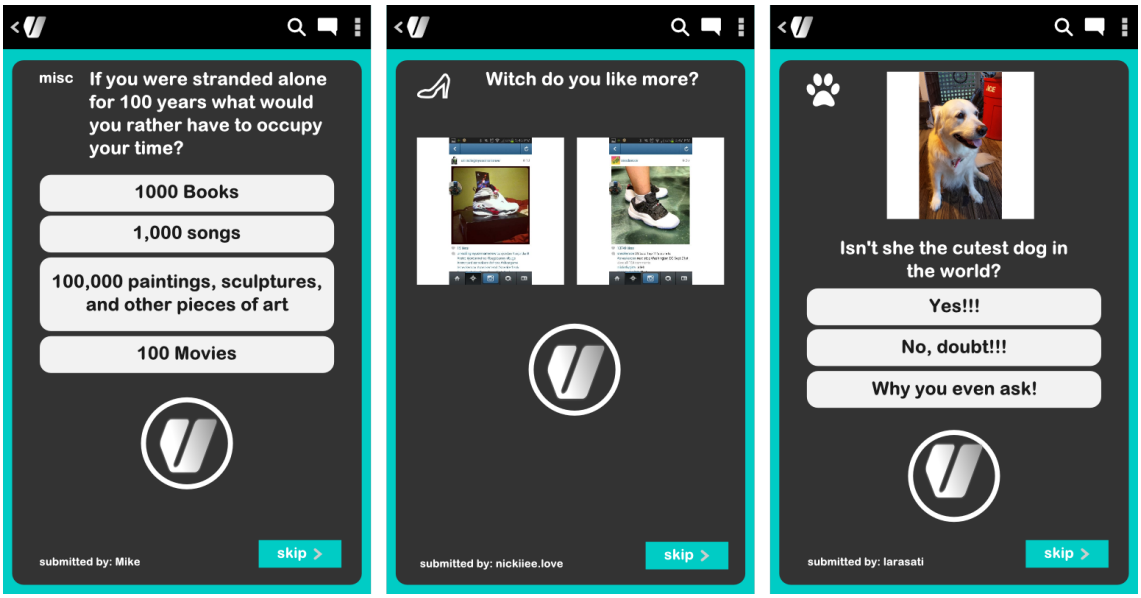

Figure 1: Screen shot of a 3 Typical Votter polls. A text poll on the left, a 2 image poll in the middle, and a 1 image poll on the right

- **Two Image Poll:** Contains a text question. The answers are 2 possible images and user selects one of the images as their answer. This does not contain any text answers e.g.
  - **Q:** "Which dress would be best for prom?"
  - **A:** Two images are included to choose from.

These poll types can be seen in Figure 1 as actual screenshots from the app. One can easily see how the three types allow different analysis from text to images. Errors are intentionally left in, such as the common misspelling of "which" and "witch" seen in the middle figure.
The votes of each user are tracked and tallied for a final result. Everyone that has voted can see the result. The average number of votes per user is 47 across the entire Votter user base. We have 26,000 polls and just shy of 1 million votes at this time.
The Votter users can login via Votter's registration page or by Facebook Login. The users are overwhelmingly younger and female. The demographics for the Facebook users can be seen in Figure 2. This shows that most of Votter's users are females between the ages 13-24. This demographic is represented by the category distribution seen in Table 1 with users mostly concentrating on "Fashion" and "Am I Pretty" photos of themselves. We see far less activity in the "Travel" and "Politics" categories, where users of this age might not be of age to vote or travel independently.

4. Data

The Votter Corpus is a collection of poll questions and their corresponding answers to the polls. The poll questions are entirely created by Votter users and are answered by other Votter users. The corpus not only consists of polls and answers that Votter users created, but also we add other information such as the poll results, categories and timestamps.

Category	Count
Am I Pretty?	9726
Dating	3727
Fashion	2847
Entertainment	2390
Miscellaneous	1868
Follow Me	1365
Pets and Animals	606
Sports	478
Foodie	369
Technology	309
Fitness	289
Movies	254
Travel	238
Politics	200
Total	**24666**

Table 1: Poll categories and their corresponding number of polls

The corpus is available in XML format that can be found at (`http://globeotter.com/vottercorpus`). The Votter corpus is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License, which allow people to share and adapt the material under Attribution, NonCommercial, and ShareAlike terms.

4.1. Annotation

User Annotation:
Some parts of the Votter Corpus are annotated by the users as they create the poll. The parts of the corpus that are created by the Votter users are as follows:

1. **Question**: The user supplies a question, which is open to all users for voting. These questions may be seeking an opinion or the correct answer to a question.

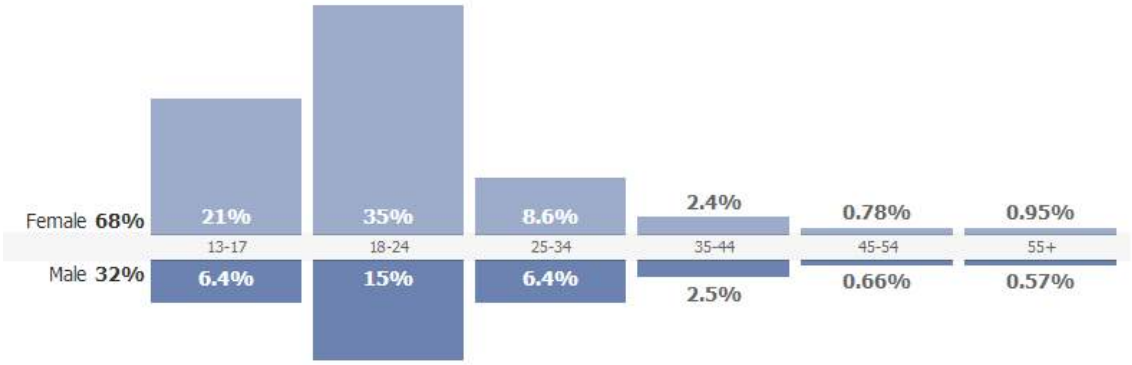

Figure 2: Demographics of Votter's Facebook Users

2. **Set of possible answers**: This is a set of possible answers to the given question for other Votter users to vote on. Interestingly, there are users who submit emoticons or certain expressions as their possible answers instead of standard answers as seen in Table 3. We allow up to 10 different possible answers for each question.

3. **Images**: When creating a poll, Votter users can use images in two ways. First, the user may submit an image to add to any textual poll questions with a set of textual possible answers. Second, we allow users to submit two images as a set of possible image answers to a poll with a textual question. In this case, the images are used as the possible answers for the question. For instance, a poll question *"Which photo should I use as my profile picture ?"* can have two image possible answers and no textual possible answers.

4. **Category**: Each poll submitted has a category assigned to it as selected by the user from a list of categories. These categories are predefined by the Votter development team. If a category is deemed to be incorrect, users can report polls that fall into the wrong category and they can be corrected on the backend.

System Annotation:
Additional automatic data is recorded for each question including timestamps and an anonymous user id. The Votter Corpus also gives the voting results for the set of possible answers for a given poll.

4.2. Data Format

Figure 3 shows one poll snippet from the Votter Corpus. The Votter Corpus is stored in a simple XML format. One poll in the Votter Corpus consists of seven main XML tags. Those tags are:

- **poll**: One poll in Votter. The poll has an *id_poll* attribute that is a unique id number assigned to the poll.
- **creator**: A Votter user id that created the poll.
- **category**: A category assigned to a poll. There are 14 possible categories (see Table 1).
- **date**: the poll submission date.
- **question**: the poll question.
- **answergroup**: the set of possible answers.
- **answer**: one of the possible answers. The answer has *id_answer* and *count* as its attribute. The *id_answer* attribute is a unique id assigned to the answer. It is a combination of the *id_poll* and the sequence number for the possible answers in the poll. The *count* attribute is the number of vote count for a particular possible answer.

```
<poll id_poll="15474">
  <creator>20627</creator>
  <category>Sports</category>
  <date>2013-05-10 17:32:11</date>
  <question>Who is going to win the NBA finals?
  </question>
  <answergroup>
    <answer id_answer="15474_0" count="12">
      Miami heat</answer>
    <answer id_answer="15474_1" count="7">
      San Antonio spurs</answer>
    <answer id_answer="15474_2" count="3">
      Oklahoma City thunder</answer>
    <answer id_answer="15474_3" count="7">
      Chicago Bulls</answer>
    <answer id_answer="15474_4" count="4">
      Golden state warriors</answer>
    <answer id_answer="15474_5" count="4">
      Memphis grizzlies</answer>
  </answergroup>
 </poll>
```

Figure 3: Votter Corpus Snippet

4.3. Characteristics

Figure 4 and Figure 5 show the top 10 unigrams that appear in poll questions and in each poll's possible answers respectively. The results show typical usage of questions and answers with a slight twist of "pretty" and "ugly" being in the top counts.
Table 2 shows the total and unique n-gram counts for up to $n = 3$ for poll questions, poll answers, and a mix of both. It is interesting to note the use of repeated language in the counts. For each dataset the unigrams are roughly 5% unique, bigrams 24% unique, and trigrams are 48% unique. Another feature of the corpus is that the questions seems to have different language style than the answer data set. This

N-gram	Question		Answer		Both	
	Total	Unique	Total	Unique	Total	Unique
1-gram	242892	10969	310054	14800	552946	20583
2-gram	218226	53865	239144	57794	457370	100928
3-gram	193560	93094	168234	80758	361794	168270

Table 2: N-Gram counts for the Votter Corpus

can be seen in the "Both" calculations. Out of 25,769 total unique unigrams from both questions and answers, only 5,186 overlapped.

Table 1 shows the poll categories and their corresponding number of polls. While Votter was initially intended for political discussion and debate, "Am I Pretty", "Fashion" and "Dating" clearly are the most popular categories. This is a very positive outcome, as it gives researchers a very focused corpus given the demographics.

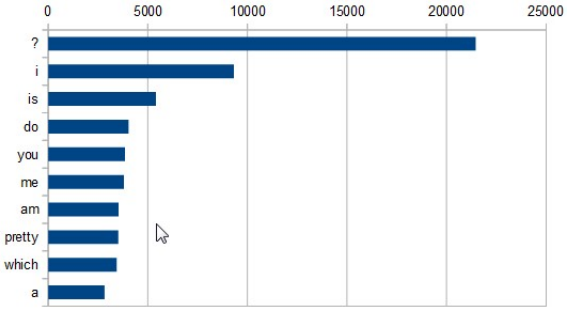

Figure 4: Top 10 Unigrams in the poll questions

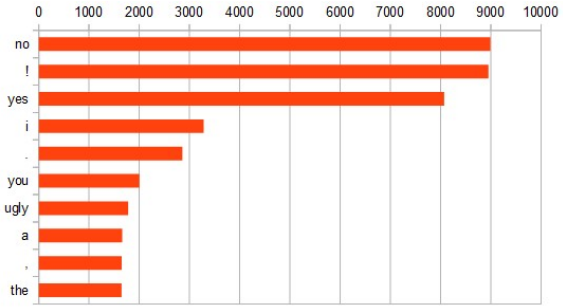

Figure 5: Top 10 Unigrams in the poll set of possible answers

Votter users use many textual emoticons to portray their emotions on the polls questions and answers. Table 3 shows the usage of some frequent emoticons. These may be used for sentiment analysis of both text and possibly their related images.

5. Conclusion

We have shown and released a new data set covering a new social media format, social polling. The contents of the corpus are opinionated questions and answers that differ greatly from other question and answer sites. Additionally, photos are available for some corresponding questions.

The corpus contains over 552,946 unigrams, many differing from typical n-grams. With very unique demographics,

Emoticon	Counts	
	Question	Answer
:)	310	568
(:	295	352
:(	38	230
:/	25	188
:D	26	89
:-)	35	41
C:	5	57
c:	13	31
:3	14	23
):	5	25
:C	0	28
:o	10	17
:*	6	20
/:	6	19
D:	2	23
:P	4	21
:p	9	13

Table 3: Several emoticon examples and their corresponding occurrence counts in the data

The Votter corpus should be useful as a new data set for NLP researchers dealing with question/answer systems and modern social language amongst other NLP tasks.

6. Future Work

We currently provide the user with a means to discuss their poll and the results of the poll. We associate a discussion board to each poll. The discussion board is powered by Disqus (`http://disqus.com`), a blog comment hosting service. In the future we can harvest further opinions given by the users related to each poll. We plan to apply named entity recognition into this work, to capture named entities and relate them to the sentiment that the users are giving through their opinion in the poll and discussion. We plan to add more information about the poll creators and the voters, such as their age group, location, gender, etc., to provide a better understanding about the language style used in the questions and answers, or the trends of the votes in the Votter Corpus.

7. Acknowledgments

Votter users, Android technology, selfie pictures, and trending pop culture. You guys rocks!

8. References

Habernal, I., Ptáček, T., and Steinberger, J. (2013). Sentiment analysis in czech social media using supervised

machine learning. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 65–74, Atlanta, Georgia, June. Association for Computational Linguistics.

Martínez-Cámara, E., Martín-Valdivia, M. T., Molina-González, M. D., and Ureña López, L. A. (2013). Bilingual experiments on an opinion comparable corpus. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 87–93, Atlanta, Georgia, June. Association for Computational Linguistics.

Petrović, S., Osborne, M., and Lavrenko, V. (2010). The edinburgh twitter corpus. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, pages 25–26, Los Angeles, California, USA, June. Association for Computational Linguistics.

Seddah, D., Sagot, B., Candito, M., Mouilleron, V., and Combet, V. (2012). The French Social Media Bank: a treebank of noisy user generated content. In *Proceedings of COLING 2012*, pages 2441–2458, Mumbai, India, December. The COLING 2012 Organizing Committee.