Political Discourse in Polish Internet – Corpus of Highly Emotive Internet Discussions

Antoni Sobkowicz
National Information Processing Institute
E-mail: antoni.sobkowicz@opi.org.pl

Abstract

In this work, we present description and initial statistical analysis on a corpus of comments from most popular polish news-related website, Onet.pl. Presented corpus contains highly informal texts, politically polarized texts, with highly emotive content. We gathered corpus containing 4,829,076 texts and 1,826,906 unique tokens total during 9 month time period which held several important political events in Poland. Presented corpus is freely available, and we intend to update it regularly, with additional texts being currently retrieved.

Keywords: politically related texts, polish language corpus, social media, computer-mediated communications

1. Introduction

Discussion about politics on the internet, especially in such politically polarized country as Poland – with supporters of two dominating parties being very vocal and active - are often very emotive. People tend to not only express their feelings about events but resort to personal insults or insults directed at politicians. This makes these discussions very interesting for analysis.

We have gathered over 4.8 million comments from largest polish news-oriented website, Onet.pl, choosing only comments under political related news, over the 9 months that were very intensive in terms of political events in the country (presidential and parliamentary elections where party ruling for last 8 years lost, Constitutional Tribunal crisis, changes in public media). We have analyzed basic properties of this set and we encourage researchers in text analysis related fields to use it. Collected dataset is freely available, and we intend to update it every three months with new content.

Dataset described in this paper was previously used in several works, although it was not publicly available.

2. Related Work

Corpora regarding political text are widely available, with examples being a multilingual corpus of annotated political programs (Merz et al., 2016), the corpus of political speeches with annotated audience reactions (Guerini et al., 2013) or political speech corpus of Bulgarian (Osenova & Simov, 2012). This corpus however only touches text with a higher degree of formalization.

Corpora build on less formal text sources are also available – based on Tweets (Longhi & Wigham, 2015), blogs (Eisenstein & Xing, 2010) and other sources. These are more similar to corpus described in this paper because of the informality of those sources.

Work on similar kind of dataset – politically related comments in the Polish language, also done on Onet.pl data - was done by Sobkowicz and Sobkowicz (2012), although dataset was highly limited.

3. Corpus Source Description

Corpus was scraped from Onet.pl website, one of the largest and most commented news related websites in Polish internet. Onet.pl is a news website, covering topics from politics to sport and entertainment, with complex comment section under each news piece, news tagging.

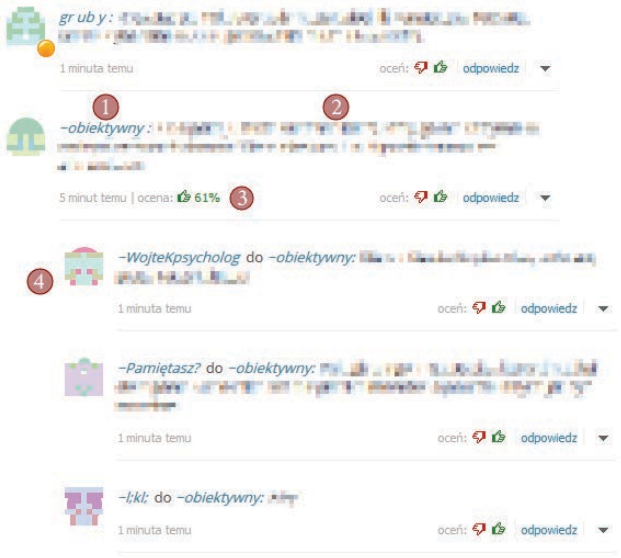

Figure 1: A chunk of typical discussion in the comment section on Onet.pl – source for corpus described in this paper. Elements are as follow: 1 - Comment poster name; 2 - Comment text; 3 - Comment score; 4 - Replies to comment, nested, with information who replied to which post. Texts were blurred out because they may be offensive.

The comment section is tree based, meaning comments that reply to other comments are displayed below with indentation. The user can rate comments, and the average rating is displayed near each comment, along date and time of posting and name of the original poster. An example of such tree and data are is shown in figure 1.

3.1 Discussions on Onet.pl

As Onet.pl does not enforce registrations, the user can post under any nickname. This seems to encourage more heated discussions, with lots of insults – token based analysis of the dataset using only known, heavily emotive negative tokens shown that around 5% of all messages can be considered as directly insulting (insulting other users or other parties connected to the topic of the discussion). Manual sentiment analysis on small randomly selected subset of data shows that purely neutral texts are only 15% of data (146 texts out of 950 assessed). In data analyzed by Sobkowicz and Sobkowicz (2012), neutral texts were 56% of all texts, however, authors used different neutrality and emotiveness measure.
Each posted news piece tends to have several hundred texts, the more dividing in opinion the topic is, the more posts are written by users.

4. Corpus Description

We have gathered comments under articles (along with article text). Data was gathered in three periods, from May – August 2015, September – December 2015 and January – March 2016, with fourth part being currently downloaded.

4.1 Comment Data Description

Comments are scraped from the website while preserving their tree structure, along time of posting and user handle. This information can be used to retrieve back user network if needed. Comments themselves are stored in their raw form, without any alteration to their text. We decided against storing only extracted and processed tokens, as we believe that preserving additional data (such as discussion tree) is very important, and extracting/lemmatizing/stemming can be done when needed.

4.2 Basic Corpus Properties

Corpus contains 4,829,076 texts, with average length of 179 characters and length distribution shown in figure 2. Average length in tokens is 33, with distribution shown in figure 2. Both of distributions seem to follow lognormal distribution as expected from human produced texts (Sobkowicz et al. 2013).
Distribution of unique tokens to a number of texts in the corpus is shown in figure 3 and 4 – non-unique tokens and unique tokens only respectively. Corpus itself contains over 160 million tokens, with 1,826,906 unique tokens (as we do no extract lemmas from the words, this number in inflated by different conjugations).
We do not provide sentiment annotation for the corpus, because given it's size and lack of good sentiment analysis tools for the Polish language, we believe we cannot give accurate or semi-accurate sentiment information for the corpus. This is the case also for lemmatization and POS tagging.

4.3 Anonymization

We believe that given the fact that source website does not require users to register and does not provide any other information about the user beyond their username anonymization is not required for this dataset.

5. Toolset Description

Data was gathered using specialized tools written in Python using scraps library. Scrappers parsed all comment pages, going from first to last, and saving all data to JSON files. These files were then parsed and saved into an SQLite database for easier use.

6. Availability

Corpus is available for free, but taking into consideration the fact that the collected corpus is relatively large in size – around 6GB, we currently do not provide direct download link – instead, we encourage to contact us to prepare data for transfer via selected service.

7. Conclusions and Future Work

We have built new corpus containing politically related comments from under news pieces on largest polish news related site. We gathered over 4.8 million texts spanning 9 months period and calculated basic corpus statistics. In near future, we plan to finish downloading fourth part of data (spanning the time from April to June 2016) and keep corpus up-to-date for foreseeable future.
We encourage researchers to use this corpus and analyze it in greater detail – in the context of linguistics, sentiment analysis, and analysis of human interactions in CMC. We believe that corpus this large, coming from very bi-polar community can be very interesting for researchers.

8. References

Eisenstein, J., & Xing, E. (2010). *The CMU 2008 political blog corpus*. Carnegie Mellon University, School of Computer Science, Machine Learning Department.

Guerini, M., Giampiccolo, D., Moretti, G., Sprugnoli, R., & Strapparava, C. (2013). The new release of corps: A corpus of political speeches annotated with audience reactions. In *Multimodal Communication in Political Speech. Shaping Minds and Social Action* (pp. 86-98). Springer Berlin Heidelberg.

Longhi, J., Wigham, C. R.. (2015) Structuring a CMC corpus of political tweets in TEI: corpus features, ethics, and workflow. *Corpus Linguistics 2015*

Merz N., Regel, S., Lewandowski, J. (2016). The Manifesto Corpus: A new resource for research on political parties and quantitative text analysis. *Research & Politics*

Osenova, P., & Simov, K. (2012). The Political Speech Corpus of Bulgarian. In *LREC* (pp. 1744-1747).

Sobkowicz, P., Thelwall, M., Buckley, K., Paltoglou, G., & Sobkowicz, A. (2013). Lognormal distributions of user post lengths in Internet discussions-a consequence of the Weber-Fechner law?. *EPJ Data Science*, *2*(1), 1-20.

Sobkowicz, P., & Sobkowicz, A. (2012). Two-year study of emotion and communication patterns in a highly polarized political discussion forum. *Social Science Computer Review*, 0894439312436512.

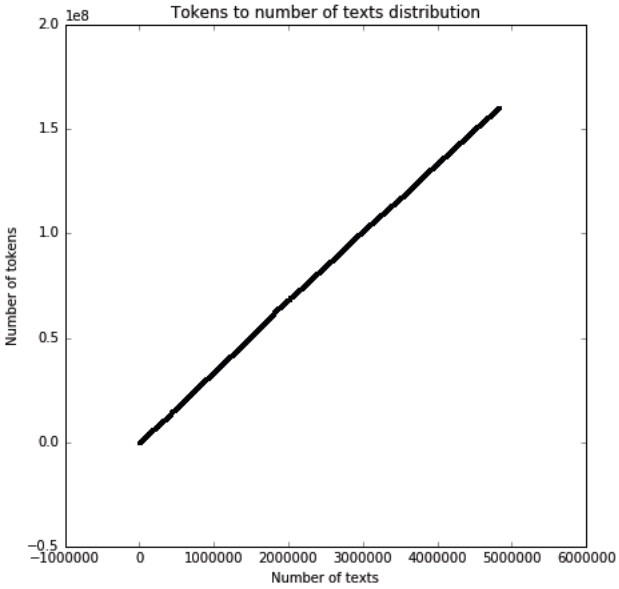

Figure 3: Distribution of number of non-unique tokens to number of texts in corpus.

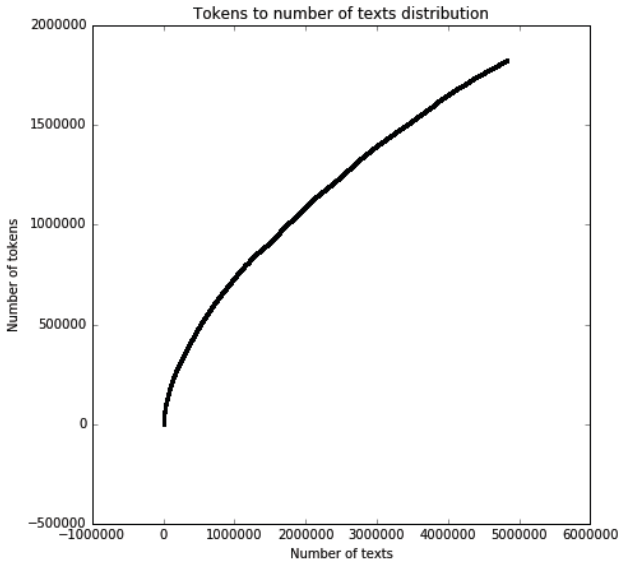

Figure 4: Distribution of number of unique tokens to number of texts in corpus.

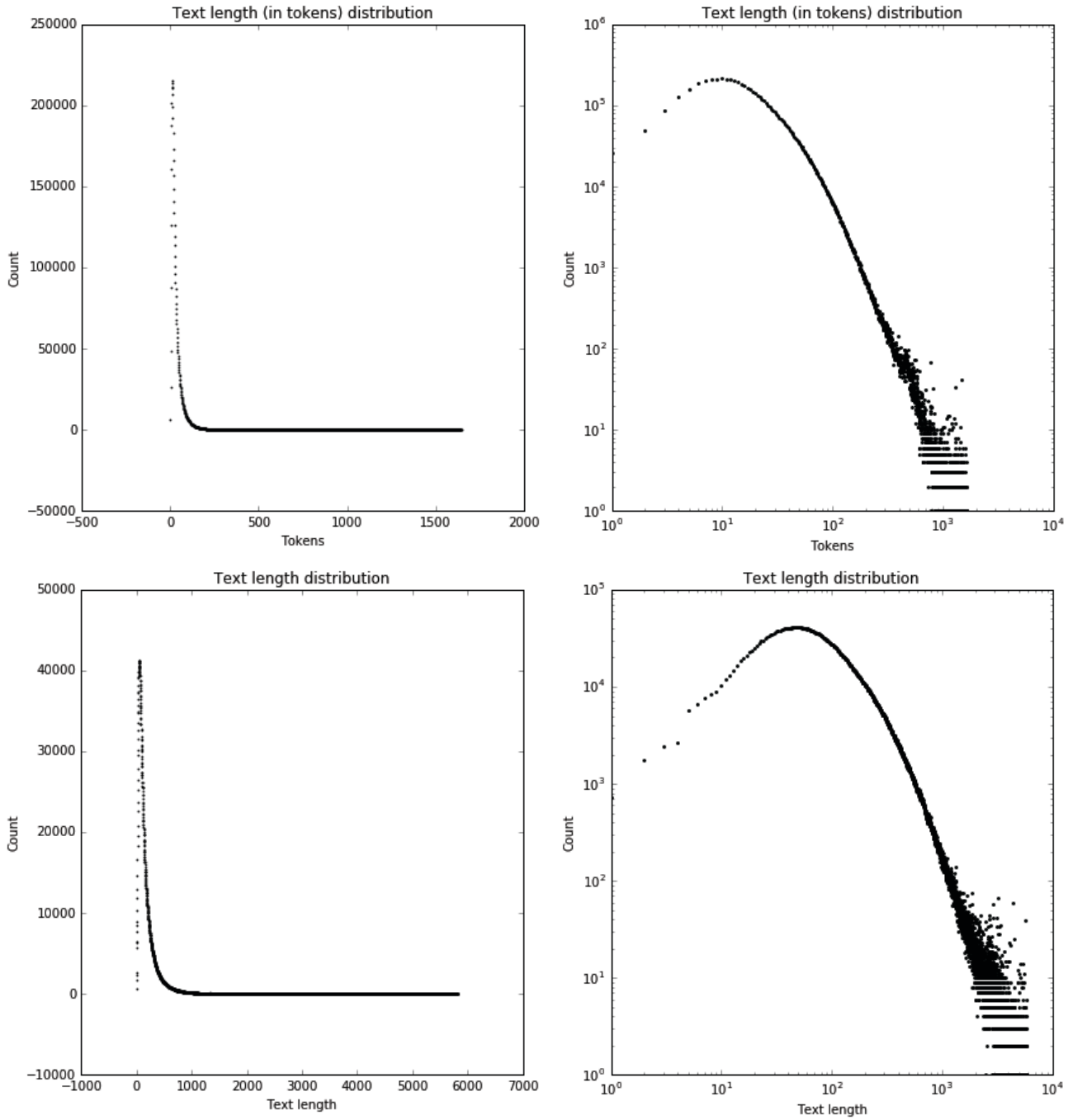

Figure 2: Distributions of text length in the corpus, both in raw character length and in token length. Right figures show the distribution in log-log scale. Both distributions seem to follow log-normal distribution, which seems to be the case for most of human-created according to other research.