Discovering Political Slang in Readers' Comments

Nabil Hossain
Dept. Computer Science
University of Rochester
Rochester, New York
nhossain@cs.rochester.edu

Thanh Thuy Trang Tran
Dept. Computer Science
Bard College
Annandale on Hudson, New York
tt7210@bard.edu

Henry Kautz
Dept. Computer Science
University of Rochester
Rochester, New York
kautz@cs.rochester.edu

Abstract

Slang is a continuously evolving phenomenon of language. The rise of social media has resulted in numerous slang terms circulating across the globe. In this paper, we aim to find novel and creative slang in the comments sections of online political news articles covering the 2016 US Presidential Election. First, we define *creative political slang* and partition it into sub-classes. Next, we extract a dataset of partisan news articles and comments ranging from the left wing to the right wing. Then, we develop *PoliSlang*, an unsupervised algorithm for detecting creative slang, evaluating its performance using expert human judgments. Finally, we use this algorithm to compare and contrast political slang usage by commenters across different news media.

Introduction

Online news sites played an out-sized and unprecedented role in the 2016 United States Presidential Election (Faris et al. 2017; Silverman 2016). Along with the rise of the readership and power of such sites, there was a rise of politically-oriented online communities that interacted through the reader comments for the sites' articles. When we examined reader comments from a range of news websites, including the New York Times, Politico, and Breitbart, we discovered an interesting linguistic phenomenon: a high level of creation and use of novel political slang. The vast majority of the slang terms we found never appeared in the actual content of articles.

We develop an algorithm for finding creative political slang in reader comments, and we perform a descriptive study of their use across a range of politically slanted sources of articles published during the 2016 US Presidential Election. Our contributions are: (i) defining a novel slang detection task, which is to find new, creative slang in political discourse; (ii) creating a large dataset of comments associated with political news articles; (iii) designing **PoliSlang**, an unsupervised algorithm for detecting creative slang; and (iv) analyzing a range of creative political slang phenomena in news site comments, including categorizing the kinds of linguistic mechanisms used to coin slang words and measuring similarities and differences in the distributions and rates of creative slang usage across different news sites in the left to right political spectrum.

Slang

The function of slang is to connect and bond members of a subcommunity (Partridge 2012). Like language, slang continuously evolves over time: some slang terms lose stigma and disappear with passage of time, others become accepted into the common culture, while new slang expressions form as groups start using new terminology. We define a **creative political slang** (CPSlang) as a recently-coined, non-standard word that conveys a positive or negative attitude towards a person, a group of people, an institution, or an issue that is the subject of discussion in political discourse. Examples of the classes of creative political slang are shown in Table 1. A *creative slang* is defined the same way, except that the target of the slang does not necessarily have to be a subject of discussion in political discourse.

There are other kinds of slang that we do not consider. These include codeword slang (Magu, Joshi, and Luo 2017) (*e.g.*, using the term "Google" to represent black people in order to evade censorship), meaning reversal (*e.g.*, "bad" instead of "good"), and implicit slang (*e.g.*, "mother" instead of "motherf****r"), and others. Moreover, our notion of CPSlang is restricted to unigrams, and we leave the analysis of slang phrases to future work.

Creative Slang Detection

Dataset

Since our purpose is to analyze creative slang usage associated with the 2016 US Presidential Election, we study the reader comments to news articles covering the election.

We choose a diverse group of partisan news sites ranging from the left-leaning to the right-leaning which allows for interesting comparisons of CPSlang usage across the whole political spectrum. To obtain the political leanings of these news sources, we make use of Allsides[1], a crowdsourcing system that ranks the political biases of news from 1 (far left) to 5 (far right). Our **News and Comments Dataset (NCD)** consists of articles and their accompanying users' comments from online publications of the news sources

[1] www.allsides.com

Class	Description	Example
Abbreviation	shortening of a word	repub (republican), huffpo (huffington post)
Acronym	word created from first letters of other words	maga (Make America Great Again), eussr (EU and USSR)
Nicknaming	assigning an offensive nickname	jebbie (Jeb Bush), presbo (President Obama), drumpf (Donald Trump)
Portmanteau	combining two words into one	killary (killer + hillary), trumpanzee (trump + chimpanzee), libtard (liberal + retard), retardican (retard + republican)
Prefix	adding a prefix to a word	uniparty, antiobama, prolife, uberwealthy
Spell	intentional misspelling	mooselimb (Muslim), obammo (Obama), gubmint (government)
Suffix	adding a suffix to a word	clintonian, islamophobe, trumpism

Table 1: Different classes of creative political slang and their examples.

Dataset	Articles	Comments
Tatar *et al.* (2014)	271,407	3,366,884
- 20minutes	231,230	2,635,489
- telegraff	40,287	731,395
Tsagkias *et al.* (2009)	290,375	1,894,925
Lichman (2013)	422,937	N/A
NCD (ours)	**87,128**	**14,965,120**
- Breitbart	54,475	13,895,687
- New York Times	16,254	873,421
- Politico	16,399	196,012

Table 2: Our News and Comments Dataset compared with related news corpora.

New York Times (NYT) (bias rating = 2), Politico (bias rating = 3) and Breitbart (bias rating = 5). To obtain articles relevant to the 2016 Election, we query the "Politics" sections of these news sites, retrieving only those articles (and their comments) that mention any of the following key words: "trump", "donald", "clinton", "hillary", "obama", "president", "immigration", "democrat", "republican", and "election". A comparison of our dataset with related news datasets is presented in Table 2.

PoliSlang: Creative Slang Detector

PoliSlang, our algorithm for detecting creative slang, takes as input a set of news reader comments and proceeds as follows:

Step 1: Pre-process. First, we start cleaning up the comments by removing punctuation, tokenizing the resulting text by splitting at white-spaces, and removing stopwords and punctuation.

Step 2: Normalize. In this step, words are reduced to their root forms, which involve converting plurals into singular forms, truncating possessive nouns to their equivalent common noun forms by detecting and removing apostrophe and "s" where necessary, *etc.*

Step 3: Remove dictionary words. Words that appear in any of a set of popular English dictionaries are removed.

Step 4: Identify misspellings. Handling misspellings is tricky because we need to differentiate between intentional spelling mistakes, which can be creative slang, and unintended spelling errors and typos. Our solution is to use a list of most common human spelling errors as found in several online sources.

Step 5: Remove common slang. This uses a list of 4,012 words to remove common curse words, insults, adult slang, Internet slang and interjections, and other historically frequent slang.

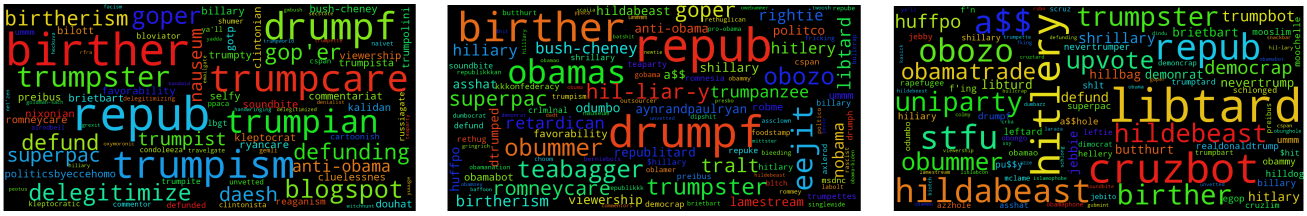
Step 7: Remove named entities. PoliSlang's final step is removing entities, such as names of people, locations, organizations, *etc.* Using a Named Entity Recognizer (NER) to find entities is a challenge in our case since NERs are trained on formally and grammatically correct language data, whereas our user comments are often very unstructured and do not follow formal rules of English. Therefore, our approach to find and remove entities is to remove words that have an entry in Wikipedia or appear in either a list of common surnames or a list of the names of members of Congress.

Any word that has passed the sequence of filters described above is considered a creative slang. Figure 1 shows word clouds of frequency-weighted potential creative slang terms discovered by PoliSlang in our datasets.

Evaluation

We collected 400 creative slang candidates by choosing from the frequency-per-comment sorted lists of PoliSlang filtered words for each dataset. In order to ensure that the words are evenly spread across the sources, we repeatedly pick the most frequent unpicked word from each source in a cyclical fashion. This selects at least 133 most frequent candidates per source. Next, we annotate these terms using three human judges who are knowledgeable in US politics and political discourse, and we accept labels by majority vote.

On the 400 candidates set, PoliSlang's precision in detecting slang, creative slang, and CPSlang, respectively, are 64.5%, 59.5% and 51%. These are encouraging results given the noisy, unstructured and conversational nature of our comments dataset. Furthermore, PoliSlang is effective in searching for newly-introduced and trending political slangs, since 82.8% of the CPSlang words it discovered did not appear until the run up to the 2016 election.



(a) NYT (b) Politico (c) Breitbart

Figure 1: Frequency word clouds of potential creative slang in our news datasets as discovered by PoliSlang.

Type	NYT	Pol	BB
Abbreviation	14.25	15.25	7.84
Acronym	1.02	0.43	0.45
Nicknaming	8.07	13.77	4.36
Portmanteau	23.94	42.2	71.96
Prefix	3.04	1.6	2.93
Suffix	47.23	21.84	9.02
Spell	2.45	4.91	3.43

Table 3: Percentage breakdown of CPSlang usage into its subclasses in each comments dataset.

Analysis and Discussion

Quantitative Breakdown of CPSlang

We analyze how the reader communities of our news sites make use of the different subclasses of CPSlang. Using our judge-annotated CPSlang terms discovered by PoliSlang, we calculate the percentage contribution of each subclass towards the overall CPSlang usage in each comments dataset. The results are shown in Table 3.

Portmanteaus are highly popular in generating CPSlang in all three reader communities, with Breitbart commenters using an unusually high proportion. A possible motivation for substantial portmanteau use in all three datasets could be their "catchy" or "sticky" nature — brought about by the humorous mockery of political entities — which makes criticism effective. The commenting communities of NYT and Politico make substantial use of suffixes in their CPSlang. As we will see below, suffixes are much milder in offending compared to portmanteaus.

Qualitative Comparison of Communities

As shown in Figure 1, the word clouds of CPSlang discovered by PoliSlang demonstrate that most of the novel slang words are derogatory terms addressed towards a politician (*e.g.*, "drumpf" for Trump, "cruzbot" for Ted Cruz, "hitlery" for Hillary, *etc.*) or a person or a group of people sharing certain beliefs or political ideologies or supporting certain politicians (*e.g.*, "libtard", "trumpster"). These word clouds also depict the motivational differences between reader communities of the news sites. NYT readers generate a majority of slang towards republicans and their presidential candidate Donald Trump (*e.g.*, "goper", "trumpism"), as is expected of followers of a left-wing news provider. Commenters in the center-leaning site Politico use CPSlang towards both the left wing (*e.g.*, "hil-liar-y", "obummer") and the right wing ("romneycare", "retardican"). Breitbart comments show quite a high proportion of CPSlang towards left-leaning groups (*e.g.*, "libtard", "democrap") and political individuals (*e.g.*, "obummer", "hildabeast").

Moreover, the Breitbart reader community is much more intense in criticizing their opposition compared to the NYT community. In NYT, CPSlang towards Trump involve mainly using suffixes to create mild offenses (*e.g.*, "trumpster", "trumpism", "trumpian", "trumpist"), whereas Breitbart readers offend Clinton with portmanteaus that directly attack her character by invoking negative frames (*e.g.*, "hitlery", "hildebeast", "shrillary").

Comparing Communities by CPSlang Usage

We measure the similarity in CPSlang usage between the three news reader communities using the generalized *Jaccard similarity coefficient* (Leskovec, Rajaraman, and Ullman 2014), which can calculate similarities between two vectors of non-negative real numbers.

First, we create normalized frequency vectors of the 204 CPSlang terms (found in the 400 candidates annotated) in each comments dataset. Given the average frequency per comment of the i-th CPSlang term in dataset X is f_i, then its normalized frequency x_i is:

$$x_i = \frac{f_i}{\sum_{j=1}^{204} f_j}$$

Next, for each pair of datasets X and Y, we compute the Jaccard similarity coefficient between their normalized frequency vectors as follows:

$$J(X,Y) = \frac{\sum_{i=1}^{204} min(x_i, y_i)}{\sum_{i=1}^{204} max(x_i, y_i)}$$

After applying these steps to our comments datasets, we get the following results:

X	Y	J(X,Y)
NYT	Politico	0.31
Politico	Breitbart	0.23
NYT	Breitbart	0.11

The results show that when using CPSlang, both NYT and Breitbart commenters share very little similarity with each other than they share with Politico commenters. This is most likely explained by the differences in political biases among the reader communities of these sites.