Title

How Do General-Purpose Sentiment Analyzers Perform when Applied to Health-Related Online Social Media Data?

Permalink

https://escholarship.org/uc/item/6fq4x80m

Authors

He, Lu
Zheng, Kai

Publication Date

2019-08-21

DOI

10.3233/shti190418

Published in final edited form as:
Stud Health Technol Inform. 2019 August 21; 264: 1208–1212. doi:10.3233/SHTI190418.

How do General-Purpose Sentiment Analyzers Perform when Applied to Health-Related Online Social Media Data?

Lu He[a], **Kai Zheng, PhD**[a]

[a]Department of Informatics, University of California, Irvine, City Irvine, CA, USA

Abstract

Sentiment analysis has been increasingly used to analyze online social media data such as tweets and health forum posts. However, previous studies often adopted existing, general-purpose sentiment analyzers developed in non-healthcare domains, without assessing their validity and without customizing them for the specific study context. In this work, we empirically evaluated three general-purpose sentiment analyzers popularly used in previous studies (Stanford Core NLP Sentiment Analysis, TextBlob, and VADER), based on two online health datasets and a general-purpose dataset as the baseline. We illustrate that none of these general-purpose sentiment analyzers were able to produce satisfactory classifications of sentiment polarity. Further, these sentiment analyzers generated inconsistent results when applied to the same dataset, and their performance varies to a great extent across the two health datasets. Significant future work is therefore needed to develop context-specific sentiment analysis tools for analyzing online health data.

Keywords

Social Media; Computing Methodologies

Introduction

Increasingly, patients use the internet to ask questions related to their health conditions and write about their experience coping with diseases. According to a survey conducted by the Pew Internet Project, in the U.S., one in five patients' living with chronic diseases participated in creating online content about their health or medical issues through social media websites [1]. As such content is readily available and contains rich information and insights, researchers have started to utilize it as a new source of data to conduct novel research studies. In this paper, we refer to such content hereafter as "online health data."

Sentiment analysis, also known as opinion mining, is a branch of natural language processing (NLP) for describing emotions from text. It provides a computational means to automatically classify positive, negative, or neutral attitude toward a subject (e.g. a movie) based on the opinions expressed in a piece of text (e.g. a movie review). In health-related disciplines, sentiment analysis has been widely applied to analyze online health data to

Address for correspondence Lu He, lhe11@uci.edu.

investigate in topics such as public opinions of health policies (e.g. the Affordable Care Act) [2], patient attitudes toward a medical treatment or intervention (e.g. vaccination) [3,4], consumer rating of healthcare services or products (e.g. hospital services; drugs and cosmetic products) [5,6] and patient journeys (e.g. stigmatization related to Alzheimer's Disease) [7].

The proliferation of research work that has applied sentiment analysis to online health data is attributable in part to the availability of several general-purpose sentiment analyzers (e.g. Stanford Core NLP Sentiment Analysis). However, all of these analyzers were initially developed in non-healthcare contexts; and were trained on non-healthcare datasets. It is therefore imperative to evaluate their validity before applying them to analyze online health data. Unfortunately, in reviewing the relevant literature, we found that previous studies used these general-purpose sentiment analyzers in a rather arbitrary manner. Most of the studies did not provide adequate justifications as to why a particular sentiment analyzer was chosen; whether it was appropriate for a study context; and whether other analyzers might produce better results.

In this paper, we aimed to address this gap by empirically evaluating the performance of three general-purpose sentiment analyzers that have been most popularly used in previous studies: Stanford Core NLP Sentiment Analysis, TextBlob, and VADER. We applied these sentiment analyzers on two datasets, representing two typical analytical scenarios with online health data: public opinions regarding health interventions and healthcare policy. We also used a non-domain specific twitter dataset to serve as the baseline and compared the performances of the analyzers on the baseline dataset with the other two health-related datasets. Through the empirical evaluation, we aimed to answer the following research questions:

1. Do different sentiment analyzers produce consistent results when applied to the same online health dataset?

2. Does the same general-purpose sentiment analyzer perform differently when applied to different online health datasets concerning different health topics?

3. Are these general-purpose sentiment analyzers adequate enough to generate useful results without retuning for online health data?

Answering these questions may help the research community establish an evidence base as regards the validity of these general-purpose analyzers when applied in studies that involve online health data. The results may also provide insights into how to properly select the right sentiment analyzer for a particular study context, and how to improve their performance in future research.

Related Work

Most of the existing sentiment analyzers and lexicons were developed based on movie reviews or product reviews, possibly because of the availability of large amounts of labeled data for training. The Hu&Liu sentiment lexicon, one of the earliest tools for sentiment analysis, was curated by manually grouping words contained in e-Commerce product reviews into different sentiment categories [8]. Similarly, the most popularly used sentiment analyzer, the Stanford CoreNLP Sentiment Analysis tool, based its sentiment treebank and model training on a movie review dataset [9,10].

In health-related studies, Korkontzelos et al. used sentiment analysis to detect adverse drug reaction (ADR) based on twitter data, and demonstrated a higher level of accuracy as compared to conventional approaches [11]. Davis et al. used a sentiment lexicon named labMT15 to assess public opinions expressed in tweets regarding the Affordable Care Act (ACA), and showed that the results were highly consistent with what could only be obtainable previously through expensive polls [2]. In another study, Du et al. used the supervised machine-learning method to automatically classify tweets based on the sentiments toward HPV, and to study the evolution of the sentiments over time using a time series analysis [3]. More recently, Burnap et al. incorporated sentiment scores into their feature sets to develop supervised machine learning models that automatically detect suicide-related tweets [12]; and Ji et al. extracted negative sentiments from online social media data to understand public health concerns regarding disease epidemics [13].

While researchers may opt to train machine-learning classifiers on their own datasets, doing so requires a significant amount of manually annotated data, in addition to sophisticated skills in developing, training, and testing machine learning models. As a result, most of the previous studies leveraged existing, general-purpose sentiment analyzers that are readily available and are relatively easy to adopt. However, as it has been previously demonstrated, sentiment analysis models trained in one domain may perform poorly when applied in another domain without adaptation. In a study on Ebola-related social media discussions, Lu et al. noticed a significant level of disagreement between the results generated by different sentiment analyzers that they experimented with [14]. In reviewing the previous studies, we also found that there was generally a lack of discussions on the rationale of choosing a particular sentiment analyzer; and very few studies validated the analyzer chosen before applying it to their datasets. It thus remains unknown whether the results reported in these studies, based on general-purpose sentiment analyzers developed in non-healthcare domains, are reliable and repeatable.

Methods

Sentiment Analyzers Studied

In this paper, we evaluated three sentiment analyzers that had been most popularly used in previous studies: Stanford Core NLP Sentiment Analysis, TextBlob, and VADER.

The Stanford Core NLP Sentiment Analysis tool was developed by the Stanford NLP group as a module in the Stanford Core NLP toolkit [9]. Its sentiment analysis model was trained using a recursive neural tensor network on a movie review dataset made available by Pang and Lee [10]. Unlike earlier sentiment analyzers, in which bag of words was used and word orders were ignored, the Stanford Sentiment Analysis tool parses input text into sentiment trees, where each leaf node refers to a word; every word in the input text is thus represented as a vector. Instead of producing numerical sentiment score, the results generated by the

Stanford Sentiment Analysis tool are in the form of discrete categories, namely "very negative," "neutral," "positive," and "very positive."

TextBlob is a widely distributed Python Library with an easy-to-use API that can be called upon by other programs to analyze sentiments of text, in addition to performing other common NLP tasks such as part-of-speech tagging and tokenization [15]. Its sentiment analyzer has two implementations: one based on a collection of semantic patterns; and the other based on a Naïve Bayes learning module. TextBlob returns sentiment analysis results in the form of numerical polarity ranging from −1 (most negative) to 1 (most positive). It also produces a companion subjectivity score in the range of 0 (very objective) to 1 (very subjective).

VADER, which stands for Valence Aware Dictionary and sEntiment Reasoner, is an open-source sentiment analyzer developed and maintained by Hutto and Gilbert. VADER is a lexicon and rule-based tool optimized for classifying sentiments expressed in user-generated text in social media [16]. The rules it utilizes are heuristics derived from a manual review of a set of tweets by multiple independent human judges. The authors have also incorporated many features that are not found in other sentiment analyzers, such as punctuation, capitalization, slangs, emoticons, and degree modifiers. The results produced by VADER are in the form of sentiment polarity (positive vs. negative), and a numeric sentiment intensity score on a scale from −4 (extremely negative) to +4 (extremely positive).

Annotated Datasets for Evaluation

To evaluate the performance of these general-purpose sentiment analyzers, we leveraged two publicly available online health datasets that were annotated in previous studies for the purposes of understanding public opinions of the Health Care Reform (hereafter referred to as the "HCR" dataset) and public opinions of HPV (the "HPV" dataset), respectively. We also used a general-purpose twitter dataset to serve as the baseline.

The first dataset, or the HCR dataset, contains tweets that include the hashtag "#hcr" in March 2010. The original annotated dataset also includes 8 different targets, for instance, Obama, HCR, Liberals, etc. The annotated sentiments include 5 categories, positive, negative, neutral, irrelevant and unsure. For our study purpose, we only included the tweets that targeted on HCR and those that expressed positive, negative or neutral sentiments. This resulted in a total of 961 tweets, with 290 of them (30%) being labeled as positive, 422 (44%) being labeled as negative and 249 (26%) being labeled as neutral.

The second dataset used in this study was curated and made available by Du et al. This dataset was created by three human annotators in an attempt to discover hierarchical sentiment categories of public opinions regarding the HPV vaccination [3]. The manually annotated results consist of three major categories: "negative," "positive/neutral," and "unrelated." The original dataset contained a total of 6,000 tweet IDs. After removing invalid tweet IDs (e.g. those that were deleted, or are no longer publicly accessible), 4,616 tweets are available to use in this study. As our study focused on sentiment analysis, we further excluded those tweets that were labeled as "unrelated," which left us 3,211 tweets to work with. Among these tweets, 1,084 (34%) were annotated as negative, and 998 (31%) were positive and 1129 (35%) were neutral.

The baseline dataset contains non-domain specific tweets that have been annotated. It includes tweets regarding products (e.g., “am loving new malcolm gladwell book - outliers”), personal experiences or feelings (e.g., “so tired. I didn’t sleep well at all last night.”), opinions (e.g., “If Google’s self-driving car is the future, I don’t want to be a part of it. #savethemanuals”). It contains 381 positive tweets (38.9%), 387 negative tweets (39.5%) and 212 neutral tweets (21.6%).

Study Design

We separately applied the three sentiment analyzers to each of the annotated datasets. To answer the first research question (“do different sentiment analyzers produce consistent results when applied to the same online health dataset”), we calculated the inter-rater agreement rate between the results generated by different sentiment analyzers when applied to the same dataset. To answer the second research question (“does the same general-purpose sentiment analyzer perform differently when applied to different online health datasets concerning different health topics”), we computed the precision-recall-F1 metrics of the sentiment analyzers across the two annotated datasets. We then compared the classification results (i.e. sentiment polarity) to the ground truth—human annotations, to answer the third research question (“are these general-purpose sentiment analyzers adequate enough to generate useful results without retuning for online health data”). Last, we performed term frequency-inverse document frequency (TF-IDF) analysis to extract and compare the important words in each sentiment category across the three datasets.

Results

The overall results - the precision, recall, F1-measure metrics of each sentiment analyzer on each dataset are reported in Table 1, Table2 and Table 3. The recall measure should be interpreted with caution, as with skewed datasets it is easy to achieve a high recall by incorrectly labeling the data. Stanford NLP sentiment analyzer performed poorly on both the HCR and the HPV dataset, compared to the baseline dataset, as it has extremely low precision (e.g., 11%, 7.3%) in detecting positive sentiments on the HCR dataset and the HPV dataset. VADER performed fine on the HCR and the HPV dataset with around 42 % to 51% F1-measures in each sentiment category, however, the performance on the baseline dataset (65%~74%) still beats these numbers. TextBlob performed the worst in distinguishing sentiment classes, especially that it assigned most labels as neutral. It performed equally poorly on the HCR and the HPV dataset with extremely low precision for the positive and negative sentiment class, but had decent performance on the baseline dataset. Therefore, the three sentiment analyzers all performed poorly on the two health-related datasets, but had satisfying performance (i.e., most having a precision or a recall higher than 60% in detecting the three sentiment categories) on the baseline dataset that is non-domain specific.

Table 4 reports the inter-rater agreement rates when applying these sentiment analyzers to the first annotated dataset (HCR). Among the three sentiment analyzers, Stanford NLP and

TextBlob exhibit a high degree of agreement with each other, 46.68%; while the results obtained by the VADER Sentiment Analysis tool correlate poorly with the other two analyzers, for instance, Stanford NLP in particular (29.88%). Shown in Table 5 are inter-rater agreement rates when applying each of the three sentiment analyzers to the second annotated dataset (HPV). Similarly, Stanford NLP and TextBlob have the highest agreement (39.3%). However, the results produced by VADER and by the Stanford NLP Sentiment Analysis tool are poorly correlated; the inter-rater agreement ratio is only 38.0%. In Table 6 we present the inter-rater agreement rates of the three sentiment analyzers on the baseline dataset. While the agreement rates are higher than those of the HPV and the HCR datasets, they are still unsatisfying with only around 50% agreement ratios of each pair of the three sentiment analyzersTo further explore possible reasons why general-purpose sentiment analyzers perform significantly worse on health-related datasets than on baseline, non-domain specific dataset, and how different health-related datasets are from the non-domain specific dataset, we used term frequency-inverse document frequency (TF-IDF) analysis. Table 6 to Table 8 present the top-5 weighted TF-IDF unigrams of the three datasets for each sentiment category.

The TF-IDF analysis may in part explain why the sentiment analyzers tend to assign too many neutral labels and fail to recognize negative and positive sentiment classes. For instance, words and phrases such as “danish” itself is neutral, but tweets that express negative sentiments toward HPV often include such words to form their arguments. However, the three sentiment analyzers failed to recognize them as negative, because without a proper context, they are neutral. A sample tweet “Vaccines trigger genetically modified diseases” is labeled as neutral by VADER and TextBlob. While we do not have space to include the top-10 unigrams and bigrams of the baseline dataset, we observed that the top-10 weighted unigrams are sentiment words such as “love”, “happy”, “amazing”, “better”, “thank”, etc, which do not appear in the unigrams in both the HCR and the HPV datasets. Therefore, the significant differences in performance of the three sentiment analyzers on health-related datasets and general datasets that are not in health-domain may be explained in part by the different ways of sentiment expression – in health-related social media data, people use words that are neutral to form arguments and express negative or positive sentiments, and such expressions cannot be accurately captured by general-purpose sentiment analyzers. We also noticed that for the HCR dataset, many of the combined word phrases such as “passthedamnbill” and “killthebill” are used to express sentiments, however current NLP tools may have difficulties parsing and splitting those combined words, and therefore makes it hard to sentiment analyzers to detect the underlying sentiments.

Discussion

Our results indicate that the three general-purpose sentiment analyzers popularly used in previous studies produced inconsistent classification results when applied to the same online health dataset, and their performance varies to a great extent across different datasets. Among these three analyzers, VADER and the Stanford Sentiment Analysis tool have the lowest degree of inter-rater agreement. This may be due to the fact that these two sentiment analyzers were developed from distinct domains: VADER drew its lexicon and rules primarily from social media data (tweets); while the Stanford Sentiment Analysis tool was developed based on movie reviews. This finding indicates that in sentiment analysis, the utility of a high-performing sentiment analyzer trained in one domain may not be transferable to other domains. Thus, when deciding what sentiment lexicons or sentiment analyzers to use, researchers should be aware of the contexts in which they were originally developed, in addition to its underlying classification mechanisms.

Overall, the performance of these three general-purpose sentiment analyzers is unsatisfactory (e.g., having extremely low precision in detecting positive sentiments or falsely labeling tweets as neutral) when applied to online health data, while they have decent performance on non-health related social media data. The Stanford Core NLP Sentiment Analysis tool may fall short because of the nature of user-generated content online, especially tweets, that contain an excessive number of anonyms, abbreviations, hashtags, and URLs; as compared to movie reviews that are generally well structured. However, this does not explain why tools such as VADER, which was specifically designed to process social media data, missed a vast majority of the text containing negative sentiments in the HPV dataset and HCR dataset. It is possible that “negativity,” “positivity,” and “neutrality” of sentiments expressed in the context of health-related discussions may be interpreted differently by human annotators, in contrast to sentiments conveyed in other types of social media exchanges, demonstrated by the TF-IDF analysis above. Thus, it is critical for health sciences researchers to be mindful of the limitation of the sentiment analyzer(s) that they choose to use.

The findings from this study provide insights into future work on how to improve the utility of sentiment analysis in studies that involve online health data. First, based on our review of previous work, it appears that a lack of comprehensive understanding of the state-of-the-art of sentiment analysis tools impedes researchers from picking the right tool for their studies, and from providing adequate rationale to justify their choices. Therefore, it will be very valuable to have a “road map” of the history and recent development of sentiment analyzers and lexicons, especially their context of development, working mechanisms, and intended use. Second, the existing general-purpose sentiment analyzers need to be significantly adapted when used to analyze online health data. This requires a thorough understanding of the nature of health-related social media discussions, and of the specific health-related topic being studied; for example, the algorithm and lexicon appropriate for determining sentiment polarity in detection of drug side effects can be very different from what is appropriate for use in understanding the public’s opinions toward a health policy. Future work is thus called for to explore how to develop context-specific lexicons and sentiment analyzers that are optimized for analyzing different types of online health data. Third, pre-study validation and post- study error analysis are essential to understand the utility and limitation of an existing sentiment analyzer, which should be performed and reported in every study that involves sentiment analysis. Unfortunately, most of the previous studies that we reviewed simply reported the results produced by a general-purpose sentiment analyzer, without conducting any assessment of the validity of the results. Lastly, knowing that sentiment analyzers usually do not perform well when applied across domains, the research community may consider developing domain adaptation techniques that can be readily applied to extend the capability of existing general-purpose sentiment analyzers, e.g., by means of re-training sentiment analysis models, adjusting heuristics and rules, or swapping lexicons.

Conclusion

In this study, we empirically evaluated the performance of three general-purpose sentiment analyzers on two different online health datasets. The results show that these general-purpose sentiment analyzers were unable to produce consistent results when applied to the same dataset, and their performance varies when applied to different datasets. These findings suggest that general-purpose sentiment analyzers developed in non-healthcare domains may perform poorly on online health data. Future work is thus needed to identify ways to tailor them, or develop new sentiment analyzers optimized for the health context.

References

[1]. Social Media and Health | Pew Research Center, (2010). https://www.pewinternet.org/2010/03/24/social-media-and-health/ (accessed March 29, 2019).

[2]. Davis MA, Zheng K, Liu Y, and Levy H, Public Response to Obamacare on Twitter, J. Med. Internet Res 19 (2017) e167. 10.2196/jmir.6946. [PubMed: 28550002]

[3]. Du J, Xu J, Song H, Liu X, and Tao C, Optimization on machine learning based approaches for sentiment analysis on HPV vaccines related tweets, J. Biomed. Semant 8 (2017). 10.1186/s13326-017-0120-6.

[4]. Du J, Xu J, Song H-Y, and Tao C, Leveraging machine learning-based approaches to assess human papillomavirus vaccination sentiment trends with Twitter data, BMC Med. Inform. Decis. Mak 17 (2017). 10.1186/s12911-017-0469-6.

[5]. Huppertz JW, and Otto P, Predicting HCAHPS scores from hospitals' social media pages: A sentiment analysis, Health Care Manage. Rev (2017) 1. 10.1097/HMR.0000000000000154.

[6]. Isah H, Trundle P, and Neagu D, Social media analysis for product safety using text mining and sentiment analysis, in: 2014 14th UK Workshop Comput. Intell. UKCI, 2014: pp. 1–7. 10.1109/UKCI.2014.6930158.

[7]. Oscar N, Fox PA, Croucher R, Wernick R, Keune J, and Hooker K, Machine Learning, Sentiment Analysis, and Tweets: An Examination of Alzheimer's Disease Stigma on Twitter, J. Gerontol. Ser. B 72 (2017) 742–751. 10.1093/geronb/gbx014.

[8]. Hu M, and Liu B, Mining Opinion Features in Customer Reviews, (n.d.) 6.

[9]. Manning C, Surdeanu M, Bauer J, Finkel J, Bethard S, and McClosky D, The Stanford CoreNLP Natural Language Processing Toolkit, in: Proc. 52nd Annu. Meet. Assoc. Comput. Linguist. Syst. Demonstr., Association for Computational Linguistics, Baltimore, Maryland, 2014: pp. 55–60. 10.3115/v1/P14-5010.

[10]. Pang B, and Lee L, Opinion mining and sentiment analysis, (n.d.) 94.

[11]. Korkontzelos I, Nikfarjam A, Shardlow M, Sarker A, Ananiadou S, and Gonzalez GH, Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts, J. Biomed. Inform 62 (2016) 148–158. 10.1016/j.jbi.2016.06.007. [PubMed: 27363901]

[12]. Burnap P, Colombo W, and Scourfield J, Machine Classification and Analysis of Suicide-Related Communication on Twitter, in: Proc. 26th ACM Conf. Hypertext Soc. Media - HT 15, ACM Press, Guzelyurt, Northern Cyprus, 2015: pp. 75–84. 10.1145/2700171.2791023.

[13]. Ji X, Chun SA, and Geller J, Monitoring Public Health Concerns Using Twitter Sentiment Classifications, in: 2013 IEEE Int. Conf. Healthc. Inform., 2013: pp. 335–344. 10.1109/ICHI.2013.47.

[14]. Lu Y, Hu X, Wang F, Kumar S, Liu H, and Maciejewski R, Visualizing Social Media Sentiment in Disaster Scenarios, in: Proc. 24th Int. Conf. World Wide Web - WWW 15 Companion, ACM Press, Florence, Italy, 2015: pp. 1211–1215. 10.1145/2700171.2791023.

[15]. TextBlob: Simplified Text Processing — TextBlob 0.15.2 documentation, (n.d.). https://textblob.readthedocs.io/en/dev/index.html (accessed March 30, 2019).

[16]. Hutto CJ, and Gilbert E, VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text, (n.d.) 10.

Table 1-

Precision, recall, F1-measure, Stanford NLP

		Dataset		
Class	**Metric**	**Baseline**	**HCR**	**HPV**
Positive	Precision	54%	11%	7.3%
	Recall	73%	62.7%	47%
	F1-measure	62%	18.8%	12.7%
Negative	Precision	39%	47.2%	58.2%
	Recall	71%	46.5%	35.8%
	F1-measure	51%	46.8%	44.3%
Neutral	Precision	74%	38.6%	39%
	Recall	34%	20%	34%
	F1-measure	47%	26.3%	36.6%

Table 2-

Precision, recall, F1-measure, VADER

		Dataset		
Class	**Metric**	**Baseline**	**HCR**	**HPV**
Positive	Precision	82.6%	54.8%	39%
	Recall	67.6%	38.2%	45%
	F1-measure	74%	45%	42%
Negative	Precision	64.7%	35.4%	57%
	Recall	90.7%	56.4%	49%
	F1-measure	75.5%	43.5%	53%
Neutral	Precision	68.8%	46.6%	50%
	Recall	62%	41.4%	51%
	F1-measure	65.3%	43.9%	51%

Table 3-

Precision, Recall, F1-measure, TextBlob

		Dataset		
Class	**Metric**	**Baseline**	**HCR**	**HPV**
Positive	Precision	59%	7.6%	3%
	Recall	72%	40.7%	49%
	F1-measure	65%	12.8%	5.9%
Negative	Precision	40%	2.6%	3.8%
	Recall	90%	68.8%	65%
	F1-measure	55%	5%	7.1%
Neutral	Precision	82%	96.8%	97.6%
	Recall	36.8%	27%	36%
	F1-measure	51%	42.2%	52%

Table 4-

Inter-rate agreement rates, the HCR dataset

Analyzer A	Analyzer B	Agreement Ratio
Stanford	TextBlob	46.68%
Stanford	VADER	29.88%
VADER	TextBlob	34.95%

Table 5-

Inter-rate agreement rates, the HPV dataset

Analyzer A	Analyzer B	Agreement Ratio
Stanford	TextBlob	39.3%
Stanford	VADER	38.0%
VADER	TextBlob	38.5%

Table 6-

Inter-rate agreement rates, the Baseline dataset

Analyzer A	Analyzer B	Agreement Ratio
Stanford	TextBlob	51.0 %
Stanford	VADER	42.6%
VADER	TextBlob	42.7%

Table 7:

Top-5 positive unigrams

Baseline	HCR	HPV
love	Passthedamnbill	Act2015
happy	Affordable	availability
Lebron	Human	post2015
earlier	Petition	saving
seats	stupakpitts	literally

Table 8:

Top-5 negative unigrams

Baseline	HCR	HPV
fucking	handsoff	neutral
warner	takeover	trigger
fuck	killthebill	victim
driverless	tax	injury
cable	codered	danish

Table 9:

Top-5 neutral unigrams

Baseline	HCR	HPV
Driving	defazio	slightly
deflategate	holding	stance
Google	schedule	callaghan
Check	breaking	heather
flight	association	project