An application of automatic sentiment analysis methods in web-political discussions

Antoni Sobkowicz*, **Marek Kozłowski***

*Ośrodek Przetwarzania Informacji - Państwowy Instytut Badawczy
al. Niepodległości 188b, 00-608 Warszawa, Poland
{antoni.sobkowicz, marek.kozlowski}@opi.org.pl

Abstract

The article presents analysis of Polish Internet political discussion forums, characterized by significant polarization and high levels of emotion. The study compares samples of discussions gathered from the Internet comments concerning the last Polish election candidates. The authors compare three dictionary based sentiment analysis methods (built using different sentiment lexicons) with two machine learning ones. The best performing algorithm is giving results closely corresponding to human evaluations.

1. Introduction

The crucial part of information acquisition has always been to find out what other people think. As the availability and popularity of opinion-rich resources such as web reviews and comments on web fora keps growing, as more and more people start using information technologies in order to seek out and understand the opinions existing within society, new areas of study arise. Internet discussion fora are a very promising field for conducting research on human communication patterns, which encompass the content, timing, and emotional tone of the communication. Such discussions are rarely moderated, allowing for various kinds of expressions, ranging from elaborate texts to simple phrases full of emoticons.

The growing relevance of political communication in social media, particularly microblogging, implies a fundamental change in traditional political communication, which has usually been exclusively initiated and managed by political actors as well as journalists. However, as this field is relatively young, more research is needed to better understand the principles of communication on social media platforms.

We hasten to point out that consumption of goods and services is not the only motivation behind people's seeking out or expressing opinions online. A need for political information is another important factor. For example, Rainie and Horrigan (Rainie and Horrigan, 2007) studied a sample of over 2500 American adults from the 31% of Americans — over 60 million people — that were 2006 campaign internet users, defined as those who gathered information about the 2006 elections online and exchanged views via email. 28% said that a major reason for these online activities was to get perspectives from within their community, and 34% said that a major reason was to get perspectives from outside their community.

First, we verify how comments on Polish Twitter and web fora can be used for building the sentiment lexicon concerning political discussions in the election period. Second, we investigate how sentiment lexicons built in such a way can be used in order to create the relevant training data set, which is needed for utilizing the machine learning approaches. Third, we train the Naive Bayes and Maximum Entropy classifiers and perform 10-fold cross validation. Finally, we evaluate several methods using lexicon build using forum comments on two data sets.

2. Related Work

The main set of sentiment analysis problems shares the following general task: analyze an opinionated piece of text and classify the opinion as falling under one of two opposing sentiment polarities or define its position on the continuum between these two polarities. A large portion of work in sentiment-related classification falls within this category. Much research on sentiment polarity classification has been conducted in the context of reviews (e.g., "thumbs up" or "thumbs down" for movie reviews). While in this context "positive" and "negative" opinions are often evaluative (e.g. "like" vs. "dislike"), there are problems where the interpretation of "positive" and "negative" is subtly different. One example is determining whether a political speech is in support of or in opposition to the issue under debate (Pang and Lee, 2008). A related task is classifying predictive opinions on election fora into "likely to win" and "unlikely to win" (Pang and Lee, 2008). Since all these problems are concerned with two opposing subjective classes, they are often amenable to similar techniques as machine learning tasks. In our work, the focus is on the three label-classification problem (positive, negative, neutral) in the context of the Internet comments analysis.

Internet discussions involve high numbers of people, in contrast to traditional media with their relatively passive audiences. In some Internet environments, such as social media hubs (e.g., Twitter, Facebook, blogs) and discussion forums, the users immediately express their views. With the rise of weblogs and the increasing tendency of online publications to turn into message-board-style reader feedback venues, informal political discourse is becoming an important feature of the intellectual landscape of the Internet, creating a challenging area for experimentation in techniques for sentiment analysis. Mutz and Martin (Mutz, 2001) defined the hypothesis that media would surpass face-to-face communications across political divides. The Internet-based discussions provide not only access to facts and opinions "packaged" by professionals and presented in a concise form, typical for press or TV, but also exposure to raw, diverse views of "ordinary people," and

they complement the traditional media in this eye-opening role.

Wojcieszak (Wojcieszak, 2010) studied self-organized, assortative grouping of people sharing the same interests and political views into online communities that may separate from each other and become internally homogeneous. The descriptions of the emotional attitude of such communities become impossible to synthesize into one coherent personal worldview, especially when accompanied by the polarized traditional media and biased selection of information sources by the users. The communication between supporters of the conflicted camps would be difficult especially in the face-to-face mode due to the strong emotions that divide the society.

MacKuen, Wolak, Keele, and Marcus (MacKuen et al., 2010) introduce an interesting concept of two idealized types of participants in political debates: a deliberative citizen,who considers all arguments, including these opposite to his views, and a partisan combatant, passionate supporter of a single viewpoint. In real-life situations, people's behaviour falls somewhere between these two extremes. The authors argue that it is emotions that distinguish the deliberative from the combative stance.

Tony Mullen and Robert Malouf (Mullen and Malouf, 2006) describe preliminary statistical tests on a new dataset of political discussion group postings, which indicate that posts made in direct response to other posts in a thread have a strong tendency to represent a political viewpoint in opposition to the original post.

Mining the Sentiment from political Web posts is presented in the paper (Durant and Smith, 2006). Sentiment classification of weblog posts, political weblog posts in particular, appears to be a more difficult problem than classification of traditional text because of the interplay of the images, hyperlinks, the style of writing and language used within weblogs. Using a dedicated dataset gathered from a two-years' worth of political weblogging, the authors investigated how correctly Naive Bayes and SVM classifiers predict the political category of a given post.

In the paper by (Bermingham and Smeaton, 2011), the recent Irish General Election was used as a case study for investigating the potential to model political sentiment through mining social media. The proposed approach combines sentiment analysis using supervised learning and volume-based measures. Evaluation was done against the conventional election polls and the final election result.

Stieglitz and Dang-Xuan (Stieglitz and Dang-Xuan, 2012) conducted research on the sentiment analysis of Twitter messages and their "retweetability". The paper examines whether the sentiment occurring in the politically relevant tweets has an effect on how often these tweets will be retweeted. In the data set of 64,431 political tweets, a positive relationship was found between the quantity of words indicating affective dimensions (including positive and negative emotions associated with certain political parties or politicians) in a tweet, and its retweet rate.

Paltoglou et al. (Paltoglou et al., 2010) employed Maximum Entropy, Naive Bayes and Lexicon based methods in order to analyze the sentiment of data originating from BBC and Digg. The results show that the Lexicon based methods outperform machine learning methods. However, Naive Bayes scores higher results than Maximum Entropy Classifier.

3. Experiments

3.1. Data Sources

Political comments were gathered from two major Polish news websites, onet.pl and wp.pl. The comments selected cover 5 topics that were important for the public opinion during the three months after the first round of presidential elections in Poland. Three of the topics covered main candidates (Andrzej Duda, Bronisław Komorowski, Paweł Kukiz), last two covered the current prime minister (Ewa Kopacz), and the shadow prime minister and one of campaign leads (Beata Szydło). We gathered 1,533,035 comments from 2057 articles published between 21 May 2015 and 28 August 2015. We used semi-automatic crawling software written in Python and C# that took list of articles (manually gathered) and crawled subsequent pages for news and comments. The dataset containing these comments will hereinafter be called POL2015.

We also gathered two datasets that were used in the creation of sentiment lexicon and in Machine Learning algorithm training. The first dataset contains around 31,095 tweets with automatically tagged sentiment, ranging from -1 (negative) to 1 (positive), with no neutral sentiment. Automatic sentiment tagging was based on an algorithm, provided by Twitter, that interprets tweets with ":)" emoticons as positive and ":(" as negative. This dataset will be hereinafter called TW2015. The second dataset (POL2012) contains 6,500 political texts from 2011, gathered by Sobkowicz and Sobkowicz (Sobkowicz and Sobkowicz, 2012), with manually tagged sentiment, ranging from -3 (very negative) to 1 (positive). The datasets were annotated by a single researcher (due to time and financial constrains), which may have skewed the objectivity of sentiment.

Political texts and Twitter texts are vastly different, as shown in the example below. Typical Twitter text is shorter than 140 letters (a Twitter's constraint by design) and often contains tokens that are not words (like links or emoticons), as in examples:

- "RT @przepisy_dzieci: Kokosowa #kasza manna - pyszne #przepisy dla dzieci :) http://t.co/uC1M5GS5yY" ("RT @przepisy_dzieci: Coconut #farina - tasty #recipe for kids :) http://t.co/uC1M5GS5yY")

- "@jerry72p tu chodzi o grubsza sprawe.Niebawem powinna wyciec :)" ("@jerry72p it's about something bigger.It should leak soon :)")

Political comments in the other hand are often longer, have better grammar and rarely have user handles, links or included:

- "Pan Duda to chyba czytać nie umie, bo wszystko mówi z pamięci nie to co pan Komorowski duka,stęka, ale jakoś przeczyta co mu napiszą". ("Mr. Duda probably does not know how to read, because

he speaks from memory only unlike Mr. Komorowski stamms,groans, but reads what they write for him.")

- "Właśnie wymieniłeś same zalety. Niestety poprzedni miał tylko dziadka - Osip Szczynukowicz. Niektórym to wystarczyło." ("You mentioned only advantages. Unfortunately previous had only grandfather - Osip Szczynukowicz. For some it was enough.")

3.1.1. Manual Sentiment Tagging

Manual sentiment tagging was done by the authors, which means that all assessments are highly subjective and could vary for a different group of people, or even for the same person re-reading the analyzed text later.

The neutral emotion tag can additionally increase error rate, since texts with clearly positive and negative emotions can be deemed neutral as more and more diverse extreme emotions appear in texts. One reason for this is that, for humans, negative emotions have higher impact (Peeters and Czapinski, 1990), which transfers to higher confidence in the accuracy of one's formed impression when it was formed more on the basis of negative traits than positive traits (Baumeister et al., 2001). This personal impression may serve as an incentive to change the sentiment in order to overcome the negative bias.

3.1.2. Sentiment Lexicon Creation

Both sentiment lexicons were created automatically from the sentiment-tagged data sets using the token sentiment value generation procedure described by Kiritchenko in (Kiritchenko et al., 2014). The method generates sentiment value s for each word (token) t based on the pointwise mutual information (PMI):

$$s(t) = PMI(t, positive) - PMI(t, negative) \quad (1)$$

$$PMI(t, positive) = log_2(\frac{freq(t, E_1) * count(W)}{freq(t, W) * count(E_1)}) \quad (2)$$

where $freq(t, E_1)$ is the number of times the token t occurs in the collection E_1 (positive tokens) and $count(W)$ is the number of different tokens t in the collection W (all tokens). $freq(t, W), freq(t, E_{-1})$ are described in similar way, as is $PMI(t, negative)$.

It allows for building lexicons without human effort and ensures that the dictionaries were created in the same conditions.

3.2. Methods

Comments were analyzed using five methods: three dictionary based, using two different sentiment lexicons, and two machine learning based ones (Naive Bayes and Maximum Entropy Classifiers) using two different training sets.

The dictionary based methods process an input text as a bag-of-words. Simple Dictionary Based (SDB) method takes the sentiment value of each word and returns the sum of sentiment values (s). $\log^2$ Dictionary Based (LDB) algorithm (inspired by (Sobkowicz and Sobkowicz, 2012)) works in a similar way, but the final sentiment S value is derived from the sentiment sum s using the following equation: $S = 0.8 * s * \log^2(\frac{h}{w})$, where h is the number of words that have a sentiment value, and w is the number of all words in an input text. NL Dictionary Based (NLDB) method employs a more complicated procedure described in (Sobkowicz, 2015). In short, NLDB takes the text, lemmatizes each word and checks its part of speech. Using this information, it applies different weights to the sentiment values of each word, depending on its location and predecessors. Finally, it saves the sum of all sentiment values obtained that way.

The machine learning methods evaluated are Maximum Entropy Classifier (MEBoW) and Naive Bayes Classifier(NB). Both methods were trained using the same training data set. Maximum Entropy is logistic regression based classifier, used when more than two outcome classes are needed, as described in (Greene, 2008). Naive Bayes classifier is a simple probabilistic classifier based on the Bayes' theorem, that was first introduced in 1950s (after (Russell and Norvig, 1995)).

Output from all methods was normalized to integer values between -1 (negative) and 1 (positive), where 0 means neutral.

3.3. Evaluation Procedure

Machine learning algorithms were evaluated using 10-fold cross-validation. All folds have comments containing all three emotion values (-1, 0, 1). Each method was tested on two data sets[1]: 950 manually sentiment annotated comments from POL2015 (hereinafter POL2015T) and 650 comments from POL2012 (hereinafter POL2012T). The data sets used to train machine learning algorithms and create sentiment lexicons consist of 5850 comments from POL2012 (TRAIN-POL) and all text from the Twitter set TW2015 (TRAIN-TWIT).

Text emotion

	-1	0	+1	Total
POL2012T	327	310	13	650
POL2015T	644	146	160	950
TRAIN-POL	2387	3304	159	5850
TRAIN-TWIT	14053	0	17042	31095

Table 1: Comment distribution in POL2012T and POL2015T

The evaluation procedure work as follows:

1. Preparation
   (a) (In case of machine learning based algorithms) Training the classifier
   (b) (In case of dictionary based algorithms) Loading the sentiment lexicon
2. Loading the test data set
3. Sentiment Classification

[1]Datasets are available on http://opi-lil.github.io/datasets/ website.

4. Comparing algorithm's sentiment values with the gold standard ones (manually tagged)

The reported sentiment values were measured using the raw accuracy (number of texts with correctly detected sentiment). The evaluation was performed for two cases: binary sentiment detection (texts with manually tagged neutral sentiment were ignored) and full sentiment detection (including neutral sentiment detection).

4. Results

We performed two kinds of experiments: one narrowed to the machine learning algorithms, the second one for all the methods.

Table 2 contains results for 10-fold cross-validation of Maximal Entropy and Naive Bayes classifiers, using TRAIN-POL as a training data set. Cross-validation was done using full, 3 level sentiment categorization (-1: negative, 0: neutral, 1: positive).

10-fold cross-validation results

	TRAIN-POL
MEBoW	0.79
NB	0.45

Table 2: Results for Accuracy on TRAIN-POL data set.

MEBoW provides better 10-fold cross-validation results than Naive Bayes. This shows that Maximum Entropy performs better than simple Naive Bayes for analysing 3 category sentiment.

Table 3 and 4 contain the results for dictionary based and machine learning based algorithms using training data sets TRAIN-POL and TRAIN-TWIT respectively. All methods were tested on POL2012T and POL2015T data sets. Columns labeled FULL contain accuracy results for 3-category sentiment detection and columns labeled as BIN contain accuracy results for binary sentiment classification (positive vs negative). Results show that for all algorithms except MEBoW the full sentiment detection accuracy is lower than the binary sentiment detection. MEBoW reports a slightly better 3-category sentiment accuracy when the test data set shares the origin with training data set, but performs noticeably worse on test data sets of different origin. MEBoW scores for TRAIN-TWIT data set were so low in comparison to NB results that this method was omitted in further discussion.

Table 5 and figure 1 contain accuracy comparison for algorithms trained using TRAIN-POL and TRAIN-TWIT data sets and tested on POL2015 data set. Column labels are the same as in previous tables. Results show that training on TRAIN-POL data set, based on tagged political texts from web portals, provides better results than using a TRAIN-TWIT, which was based on general Twitter messages.

5. Conclusions

The paper compares three dictionary based sentiment analysis methods, built using different sentiment lexicons, with two machine learning based sentiment classifiers, using Internet comments concerning the last Polish election

Accuracy results for test data

	POL2012T		POL2015T	
	FULL	BIN	FULL	BIN
MEBoW	0.73	0.78	0.39	0.35
NB	0.52	0.95	0.65	0.76
SDB	0.52	0.97	0.68	0.78
NLDB	0.53	0.94	0.63	0.74
LDB	0.53	0.94	0.69	0.73

Table 3: Results for Accuracy on POL2012T and POL2015T test data sets, algorithms were trained with TRAIN-POL data set.

Accuracy results for test data

	POL2012T		POL2015T	
	FULL	BIN	FULL	BIN
NB	0.27	0.44	0.25	0.3
SDB	0.26	0.46	0.31	0.37
NLDB	0.23	0.44	0.28	0.34
LDB	0.25	0.45	0.3	0.35

Table 4: Results for Accuracy on POL2012T and POL2015T data sets, algorithms were trained with TRAIN-TWIT data set.

candidates. We use Twitter and Internet forum's comments both as training/test data sets and for building the sentiment lexicon concerning political discussions in the election period.

Results show that Naive Bayes algorithm does not work well with 3 category input. Although it has a better binary accuracy than Maximum Entropy, its performance on texts with neutral sentiment falls short. Detecting neutral sentiment is a difficult task (although there are works that try to combat this problems, for example by detecting if text is emotive at all, see (Ptaszynski et al., 2014)); all evaluated methods except Maximum Entropy report worse accuracy on both test datasets.

Comparison of dictionary based methods with Naive Bayes shows that their accuracy is nearly the same on all test data sets. Results above 0.75 for binary sentiment classification are high enough to be usable in real world scenarios, even taking into consideration the bias resulting from the manual sentiment tagging. Notably, SBD, the simplest of dictionary based methods, provides the best overall results. This may be due to the fact that dictionary based methods are less sensitive to missing data. The Maximum

Accuracy comparison for POL2015

	TRAIN-POL		TRAIN-TWIT	
	FULL	BIN	FULL	BIN
NB	0.65	0.76	0.25	0.3
SDB	0.68	0.78	0.31	0.37
LDB	0.69	0.73	0.3	0.35
NLDB	0.63	0.74	0.28	0.34

Table 5: Results for Accuracy on POL2015T using TRAIN-POL and TRAIN-TWIT training data sets for algorithm training and sentiment lexicon generation.

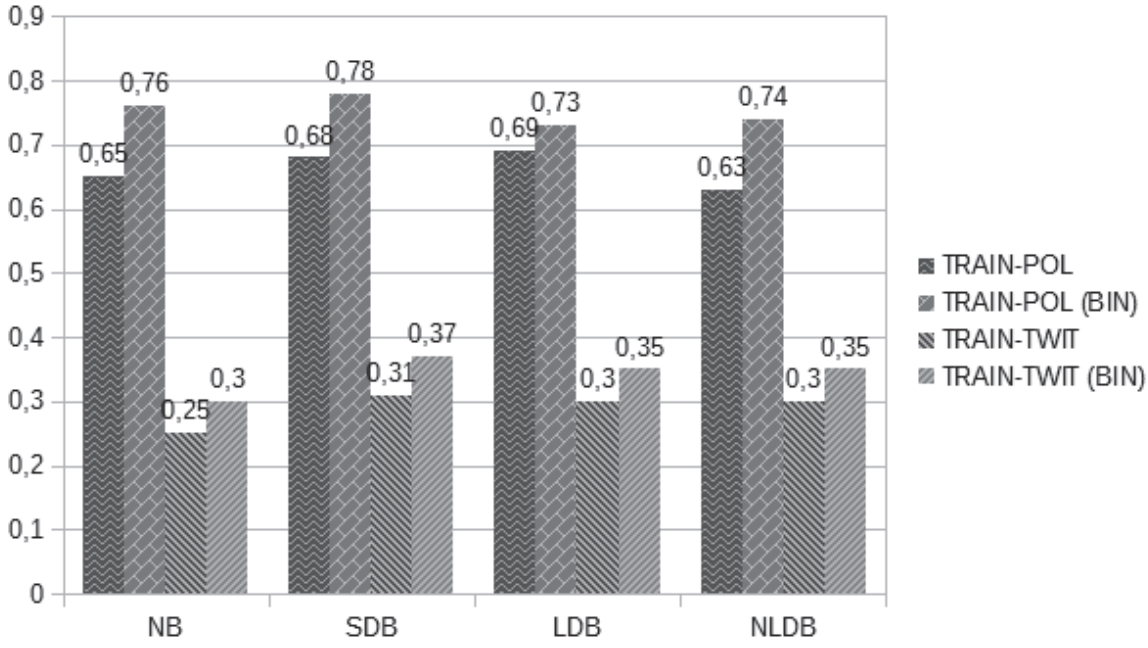

Figure 1: Results for Accuracy on POL2015 using TRAIN-POL and TRAIN-TWIT training data sets for algorithm training and sentiment lexicon generation.

Entropy algorithm, while performing well on a data set similar to the training set, works much worse for data coming from other sources.

Experiments using Twitter data set for machine learning training and sentiment lexicons creation show that general, non-filtered Twitter data cannot be easily used to build relevant models/lexicons for analysing political texts from the web portals. This is because political texts very often contain domain-specific slang not represented in other social media. Interestingly, despite political changes during last 3 years, the training data from 2012 seems to work well on texts from 2015. This indicates that Polish political slang is fairly stable.

Overall, the results show that while high binary accuracy can be easily achieved, detecting all three sentiment categories with high accuracy (>0.75) is hardly possible using only text processing. One way to improve the algorithms is to use some additional features, e.g. ones from social network analysis.

6. References

Baumeister, Roy F, Ellen Bratslavsky, Catrin Finkenauer, and Kathleen D Vohs, 2001. Bad is stronger than good. *Review of general psychology*, 5(4):323.

Bermingham, Adam and Alan F Smeaton, 2011. On using twitter to monitor political sentiment and predict election results.

Durant, Kathleen T and Michael D Smith, 2006. Mining sentiment classification from political web logs. In *Proceedings of Workshop on Web Mining and Web Usage Analysis of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (WebKDD-2006), Philadelphia, PA*.

Greene, William H, 2008. The econometric approach to efficiency analysis. *The measurement of productive efficiency and productivity growth*:92–250.

Kiritchenko, Svetlana, Xiaodan Zhu, and Saif M Mohammad, 2014. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*:723–762.

MacKuen, Michael, Jennifer Wolak, Luke Keele, and George E Marcus, 2010. Civic engagements: Resolute partisanship or reflective deliberation. *American Journal of Political Science*, 54(2):440–458.

Mullen, Tony and Robert Malouf, 2006. A preliminary investigation into sentiment analysis of informal political discourse. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*.

Mutz, Diana C, 2001. Facilitating communication across lines of political difference: The role of mass media. In *American Political Science Association*, volume 95. Cambridge Univ Press.

Paltoglou, Georgios, Stéphane Gobron, Marcin Skowron, Mike Thelwall, and Daniel Thalmann, 2010. Sentiment analysis of informal textual communication in cyberspace. *In Proc. Engage 2010, Springer LNCS State-of-the-Art Survey*:13–25.

Pang, Bo and Lillian Lee, 2008. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.

Peeters, Guido and Janusz Czapinski, 1990. Positive-negative asymmetry in evaluations: The distinction between affective and informational negativity effects. *European review of social psychology*, 1(1):33–60.

Ptaszynski, Michal, Fumito Masui, Rafal Rzepka, and Kenji Araki, 2014. Emotive or non-emotive: That is the question. *ACL 2014*:59.

Rainie, L. and J. Horrigan, 2007. Election 2006 online.

Russell, Stuart and Peter Norvig, 1995. Artificial intelligence: a modern approach.

Sobkowicz, Antoni, 2015. Automatic sentiment analysis in polish language. In *Machine Intelligence and Big Data in the Industry*.

Sobkowicz, Pawel and Antoni Sobkowicz, 2012. Two-year study of emotion and communication patterns in a highly polarized political discussion forum. *Social Science Computer Review*:448–469.

Stieglitz, Stefan and Linh Dang-Xuan, 2012. Political communication and influence through microblogging–an empirical analysis of sentiment in twitter messages and retweet behavior. In *System Science (HICSS), 2012 45th Hawaii International Conference on*. IEEE.

Wojcieszak, Magdalena, 2010. "don't talk to me": effects of ideologically homogeneous online groups and politically dissimilar offline ties on extremism. *New Media & Society*.