Classifying opinions in online political discussion – comparison of two methods

Abstract

Bla bla

keywords:

1 Introduction

Over the past years there has been an alarming growth in hate against minorities like Muslims, Jews, Gypsies and gays in Europe, driven by right wing populism parties and extremist organizations (Fekete, 2013; Wilson and Hainsworth, 2013). A similar increase in hate speech has been observed on the Internet (Goodwin et al., 2013; Bartlett et al., 2013), and experts are concerned that individuals influenced by this web content may resort to violence as a result (Strømmen, 2012; Sunde, 2013). Hateful speech is not only observed on extremst sites, but also as comments on e.g. Twitter, YouTube and online newspaper articles.

Social media and online discussions contain a wealth of information which can make us able to understand the extent of hate speech on the Internet. However, it turns out that academia is lacking research on social media and online radicalization (Taylor, 2013). Opinion mining is the discipline of automatically extracting opinions from a text material and may be one important tool in the understanding online radicalization. Opinion mining has mostly been used to analyze opinions in comments and reviews about commercial products, but there are also examples of opinion mining towards political tweets and discussions, see e.g. Tumasjan et al. (2010); Chen et al. (2010). Opinion mining towards political discussions is known to be hard since citations, irony and sarcasm is very common (Liu, 2012).

Opinion classification is perhaps the most studied topic within opinion mining. It aims to classify a set of text as either positive or negative and sometimes also neutral. There are mainly two approaches, one based on machine learning and one based on based on using a list of words with given sentiment scores (lexical approach). One simple lexical approach is to count the number of words with positive and negative sentiment in the document as suggested by Hu and Liu (2004). One may classify the opinion of larger documents like movie or product reviews or smaller documents like tweets, comments or sentences. See Liu (2012), chapters three to five and references therein for the description of several opinion classification methods.

In this paper we focus on classifying the opinion toward religious/political topics, say the Quran, in political discussion by using the lexical-based approach. One intuitive approach is to find both the keyword (e.g. Quran) and the words with sentiment in the sentence and classify the sentiment of the sentence based on the polarity of these sentiment words. We expect that statistically the importance of a sentiment word towards the keyword is related on the number of words between the sentiment and key word as suggested by Ding et al. (2008). Two other approaches is to automatically parse the material and either use the distance between key and sentiment word in the parse tree or develop grammatical dependence paths, see e.g. Jiang et al. (2011). The aim of this paper is to compare the performance of a word distance method (Ding et al., 2008) with a developed method based on distance in parse tree and grammatical dependence paths to classify opinions in political discussions.

The paper is organized as follows.

2 Opinion mining methods

In this Section we present two methods to classify sentences to either positive, neutral or negative towards a keyword. Both methods follow the same general algorithm presented below which is inspired by Ding et al. (2008) and is based on a sentiment lexicon. Both keywords, sentiment words and sentiment shifters can in general appear several times in a sentence. Sentiment shifters is words that potentially shift the sentiment of a sentence from positive to negative or negative to positive. E.g. “not happy” have the opposite polarity than just “happy”. We assume that we have a list of sentiment words each associated with a sentiment score representing the polarity and strength of the sentiment word (sentiment lexicon). Let $kw_i, i \in \{1, 2, \ldots, I\}$ represent appearance number i of the keyword in the sentence. Further let $sw_j, j \in \{1, 2, \ldots, J\}$ be appearance number j of a sentiment word in the sentence. Finally let $\mathbf{ss} = (ss_1, ss_2, \ldots, ss_K)$ repressent the sentiment shifters in the sentence. We compute a sentiment score, S, for the sentence as follows

$$S = \frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{J}\mathbf{imp}(kw_i, sw_j)\mathbf{shift}(sw_j, \mathrm{ss}) \tag{1}$$

where the function Imp computes the importance of the sentiment word sw_j on the keyword appearance kw_i. This will be computed in different ways as described below. Further, the function $\mathbf{shift}(sw_j, \mathrm{ss})$ computes whether the sentiment of sw_j should be shifted based on all the sentiment shifters in the sentence. It returns -1 (sentiment shift) if some of the sentiment shifters is within d_p words in front or d_n words behind sw_j, respectively. Else the function, returns 1 (no sentiment shift). We classify the opinion towards the keyword to be positive, neutral or negative if $S >= t_p$, $t_p > S > t_n$ and $S <= t_n$, respectively. The parameters d_p, d_n, t_p and t_n is tuned using a training set.

2.1 Word distance method

For the word distance method we use the following Imp function

$$\mathbf{imp}(kw_i, sw_j) = \frac{\mathbf{sentsc}(sw_j)}{\mathbf{worddist}(kw_i, sw_j)} \tag{2}$$

where $\mathbf{sentsc}(sw_j)$ is the sentiment score of sw_j from the sentiment lexicon and $\mathbf{worddist}(kw_i, sw_j)$ is the number of words between kw_i and sw_j in the sentence plus one.

2.2 Parse tree method

LEGGE INN FRA LILJA OG PER ERIK OM PARSING OG DEPENDENCE PATHS.

Let $\mathcal{D}$ denote the set of all important grammatical relations. The function $\mathbf{gram}(kw_i, sw_j)$ returns the grammatical relation, and if $\mathbf{gram}(kw_i, sw_j) \in \mathcal{D}$, then the function $W_{\mathbf{dep}}(kw_i, sw_j) \in [0, 1]$, return the importance of the grammatical relation. Further let $\mathbf{treedist}(kw_i, sw_j)$ return the number of words between the two words in the parse tree plus one. The Imp function is computed as follows. If $\mathbf{gram}(kw_i, sw_j) \in \mathcal{D}$ we use

$$\begin{aligned}\mathbf{imp}(kw_i, sw_j) &= \\ &\alpha \cdot \mathbf{sentsc}(sw_j) W_{\mathbf{dep}}(kw_i, sw_j) \\ &+(1-\alpha)\cdot \frac{\mathbf{sentsc}(sw_j)}{\mathbf{treedist}(kw_i, sw_j)}\end{aligned} \tag{3}$$

where $\alpha \in [0, 1]$ is parameter that weights the score from the important dependence path and the tree distance may be tuned from a training set. If $\mathbf{gram}(kw_i, sw_j) \notin \mathcal{D}$ we simply use

$$\mathbf{imp}(kw_i, sw_j) = \frac{\mathbf{sentsc}(sw_j)}{\mathbf{treedist}(kw_i, sw_j)} \tag{4}$$

Note that when $\alpha = 0$, (3) reduces to (4).

2.3 Statistical analysis of classification performance

We compare the classification performance of a set of M different methods, denoted as $\Pi_1, \Pi_2, \ldots, \Pi_M$, using a random effect logistic regression. Let the stochastic variable $Y_{tm} \in \{0, 1\}$ represents whether method Π_m, $m \in \{1, 2, \ldots, M\}$ classified the correct opinion to sentence number $t \in \{1, 2, \ldots, T\}$, where T is the number of sentences in the test set. We let Y_{tm} being the dependent variable of the regression model. The different methods $\Pi_1, \Pi_2, \ldots, \Pi_M$ is included as a categorical independent variable. We also assume that classification performance of the different methods depends on the sentence to be classified, thus the sentence number therefore is included as a random effect. The model is then formulated as

$$\begin{aligned}&P(Y_{tm} = 1) = \mathrm{logit}(\mu + \Pi_m + \epsilon_t + \epsilon_{tm}) \\ &P(Y_{tm} = 0) = 1 - P(Y_{tm} = 1)\end{aligned} \tag{5}$$

where ϵ_t is the sentence number random effect and ϵ_{tm} is additional random noise. Fitting the model to the observed classification performance of the different methods we are able to see if the probability of classifying correctly significantly vary between the methods.

The statistical analysis is performed using the statistical program R (R Core Team, 2013) and the R package `lme4` (Bates et al., 2013).

3 Real data example

3.1 Text material

We did not find any suitable annotated text material related to political discussions and therefore created our own. We manually selected 46 debate articles from the Norwegian online newspapers *NRK Ytring, Dagbladet, Aftenposten, VG and Bergens Tidene*. To each debate article there were attached an discussion thread where readers could express their opinions and feelings towards

Table 1: Manual annotation of training and test set.

	Negative	Neutral	Positive
Training	174 (46%)	162 (42%)	46 (12%)
Test	102 (33%)	182 (59%)	24 (8%)

the content of the debate article. All the text from the debate articles and the subsequent discussions were collected using text scraping (Hammer et al., 2013). The debate articles were related to religion and immigration and we wanted to classify the opinion towards all words with stem: *islam, muslim, quran, allah, muhammed, imam and mosque*. These are topics that often creates a lot of active discussions and disagreements.

We automatically divided the material into sentences and all sentences containing at least one keyword and one sentiment word were kept for further analysis. If a sentence contained more than one keyword, e.g. both islam and quran, the sentence were repeated one time for each keyword. We could then classify the opinion towards each of the keywords in the sentence consecutively. To assure that we do not underestimate the uncertainty in the statistical analysis, we see each repetition of the sentence as the same sentence with respect to the sentence random effect in the regression model in Section 2.3

Each sentence were manually annotated whether the commenter were positive, negative or neutral towards the keyword in the sentence. Each sentence were evaluated individually. The sentences were annotated based on all our knowledge, e.g. a sentence like "Muhammed is like Hitler" would be annotated as a negative opinion towards Muhammed. Further, if a commenter presented a negative fact about the keyword, the sentence would be denoted as negative.
INTERANNOTARY!!
Finally the material were divided in to two parts where the first half of the debate articles with subsequent discussions where in the training set and the rest were in the test set. The researcher working on finding the important dependence paths, did only use the training set and did never see the test set before the decisions about dependence paths were decided. After the division, the training and test set consisted of a total of 382 and 308 sentences, respectively. Table 1 summarizes the opinions in the sentences

3.2 Sentiment lexicon and sentiment shifters

Unfortunately, no sentiment list exists for the Norwegian language and therefore we developed our own by manually translating the AFINN list (Nielsen, 2011). We also manually added 1590 words relevant to political discussions like 'deport', 'expel', 'extremist', 'terrorist' and so on, ending up with a list of 4067 Norwegian sentiment words. Each word were given a score from -5 to 5 ranging from words with extremely negative sentiment (e.g. 'behead') to highly positive sentiment words (e.g. 'breathtaking').

Several Norwegian sentiment shifters were considered but only the basic shifter 'not' improved the opinion classification and therefore only this word were used in the method.

3.3 Classification performance

In this study we compared four different methods based on the general algorithm in (1).

- We use the **imp**-function presented in (2). We denote this method WD (word distance).
- For this method and the two below we use the **imp**-function in (3). Further we set $\alpha = 0$ meaning that we do not use the important dependence paths. We denote this method A0 ($\alpha = 0$).
- We set $\alpha = 1$ and for all dependency paths we set $W_{\mathbf{dep}} = 2/3$. We denote this method CW (constant weights).
- We set $\alpha = 1$ and for $W_{\mathbf{dep}}$ we use the weights presented in Table ??????. We denote this method OD (optimal use of dependence paths)

For each method we used the training set to manually tune the parameters d_p, d_n, t_p and t_n of the method. The parameters were tuned to optimize the number of correct classifications.

Table 2 shows the optimal parameter values of d_p, d_n, t_p and t_n and the classification performance for different methods. We see that the sentiment shifter 'not' only have a positive effect on classification performance when it is in front of the sentiment word. We see that using dependence tree distances (method A0) the classification results is significantly improved compared word distance in the sentence (method WD) (p-value = 0.047). Also including the important dependence

Table 2: The second to the fifth column show the optimal values of the parameters of the model. The sixth column show the number of correct classifications and the last column shows the p-value that the method performs better than WD.

	d_p	d_n	t_p	t_n	Correct	p-val
WD	2	0	0.7	0.0	145 (47%)	
A0	2	0	2.0	0.3	161 (52%)	0.047 *
CW	2	0	2.0	0.3	161 (52%)	0.048 *
OD	2	0	2.0	0.3	162 (53%)	0.031 *

paths (method OD) further improves the classification performance (p-value = 0.031).

4 Closing remarks

In this paper we have analyzed if parsing can improve

References

Jamie Bartlett, Jonathan Birdwell, and Mark Littler. 2013. The rise of populism in Europe can be traced through online behaviour... Demos, http://www.demos.co.uk/files/Demos_OSIPOP_Book-web_03.pdf?1320601634. [Online; accessed 21-January-2014].

Douglas Bates, Martin Maechler, and Ben Bolker, 2013. *lme4: Linear mixed-effects models using S4 classes*. R package version 0.999999-2.

Bi Chen, Leilei Zhu, Daniel Kifer, and Dongwon Lee. 2010. What Is an Opinion About? Exploring Political Standpoints Using Opinion Scoring Model. In *AAAI*.

Xiaowen Ding, Bing Liu, and Philip S. Yu. 2008. A Holistic Lexicon-based Approach to Opinion Mining. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 231–240, New York, NY, USA. ACM.

Liz Fekete. 2013. Pedlars of hate: The violent impact of the European far Right. Institute of Race Relations, http://www.irr.org.uk/wp-content/uploads/2012/06/PedlarsofHate.pdf. [Online; accessed 21-January-2014].

Matthew Goodwin, Vidhya Ramalingam, and Rachel Briggs. 2013. The New Radical Right: Violent and Non-Violent Movements in Europe. Institute for Strategic Dialogue, http://www.strategicdialogue.org/ISD%20Far%20Right%20Feb2012.pdf. [Online; accessed 21-January-2014].

Hugo Hammer, Alfred Bratterud, and Siri Fagernes. 2013. Crawling Javascript websites using WebKit with application to analysis of hate speech in online discussions. In *Norwegian informatics conference*.

Minqing Hu and Bing Liu. 2004. Mining and Summarizing Customer Reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, New York, NY, USA. ACM.

Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. 2011. Target-dependent Twitter Sentiment Classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 151–160, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Finn Årup Nielsen. 2011. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *CoRR*, abs/1103.2903.

R Core Team, 2013. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Øyvind Strømmen. 2012. *The Dark Net. On Right-Wing Extremism, Counter-Jihadism and Terror in Europe*. Cappelen Damm.

Inger Marie Sunde. 2013. Preventing radicalization and violent extremism on the Internet (Norwegian). The Norwegian Police University College 2013:1.

Hannah Taylor. 2013. Social Media for Social Change. Using the Internet to Tackle Intolerance. Institute for Strategic Dialogue, http://tsforum.event123.no/UD/rehc2013/pop.cfm?FuseAction=Doc&pAction=View&pDocumentId=46414. [Online; accessed 21-January-2014].

Andranik Tumasjan, Timm O Sprenger, Philipp G Sandner, and Isabell M Welpe. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the fourth international aaai conference on weblogs and social media*, pages 178–185.

Robin Wilson and Paul Hainsworth. 2013. Far-right Parties and discourse in Europe: A challenge for our times. European network against racism, http://cms.horus.be/files/99935/MediaArchive/publications/20060_Publication_Far_right_EN_LR.pdf. [Online; accessed 21-January-2014].