The Vaccination Debate

A way to structure online debates using Text Mining

Vrije Universiteit Amsterdam  
Applied Text Mining – L\_PAMATLW002  
Group 1  
Jip van Stijn, Mehul Verma, Stan Frinking & Sanne Hamersma

**Contents**

Introduction and motivation…………………………………………………………………...2

Pipeline of the first perspective graph ………………………………………………………...3

Improvement of extracting relevant opinions ………………………………………………...5

Evaluation of sentence polarity ……………………………………………………………….6

Evaluation of stance of article………………………………………………………………....6

The final perspective graph: onclusion and discussion………………………………………..7

Appendix………………………………………………………………………………………7

**1. Introduction and motivation**

While people have partaken in debates for centuries, the nature of these discussions have changed drastically since the emergence of the internet. As more people share their opinion and less people seem to listen, facts, science, and the very foundations of truth are being questioned in online mass-debates. A striking example of this is the vaccination debate, in which many actors have provided their opinions on the safety and desirability of national vaccination programmes for children. As the online environment is unstructured and chaotic, it is difficult to have an overview of which actors use certain arguments to support which side of the debate.

This report attempts to use Text Mining to provide insight in this chaos, structuring opinions, arguments and actors in a perspective graph. The Vrije Universiteit Amsterdam supplied us with 293 files of annotated texts about vaccination scraped from the internet.

As an end product, the perspective graph is expected to include the following features:

* Source
* Cue
* Content
* Clustered group that the content belongs to, and the most representative content of this group
* Polarity (positive, negative or neutral)
* Stance (pro-vac strong, pro-vac weak, anti-vac strong, anti-vac weak)
* Certainty

This data in a graph provides the information required to have a clear overview of the debate. It is a fast way to find all opinions grouped together and compare the stances of these groups. Certainty gives extra insight in how much value should be assigned to the opinion.

The following section describes the pipeline used to create the first perspective graph. An analysis of the patterns of errors in this graph are provided in section 3, 4 and 5 as well as the methods that were used to improve these. Afterwards, two sections will cover the evaluation of the polarity feature as well as the stance of each article, using annotated gold data provided by the VU. The final section presents the final perspective graph and covers the conclusions and discussion of the project.

The work was divided equally where each group member played a mentoring role in their specialty. Stan and Sanne, Text Mining master students, focussed more on extracting the relevant sentences and the corresponding source cue content. Mehul and Jip are following the Artificial Intelligence master and tutored on coding. They concentrated on the grouping methods and extracting the stance. Stan and Sanne helped them with the natural language processing jargon and known packages to use.

**2. Pipeline for the first perspective graph**

First a perspective graph is made using the given annotations of the texts in NAF and CoNNL format. After evaluating this perspective graph by using error analysis, building blocks were added in the NLP Pipeline. A second evaluation is performed, providing the final perspective graph. Below the steps of the NLP Pipeline are elaborated.

Extracting relevant opinions from the NAF opinion files  
The first challenge when trying to create an overview of an online debate in a structured way is to find the sentences that give any insight in the matter. Not all sentences that appear in a text are important for evaluating what the text introduces to the discussion. So the question is, how do you find the sentences that are important. In answering this question, a significant consideration is whether to focus on precision or recall. Precision is the accuracy of the selection, i.e. from the selected items, how many are relevant. Recall is, from the set of all relevant sentences, the amount of sentences that are actually selected. Since the goal of this assignment is to create an overview of the online debate, we expected that if we only include the sentences that we are quite sure of add something to this discussion, we would get a good overview. Including too many sentences, even those that may not be relevant, could result in a skewed representation of the internet.

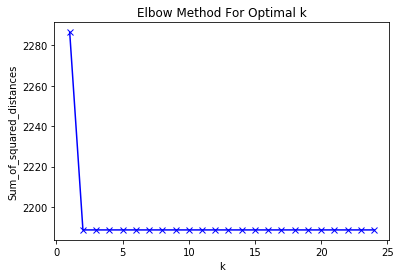
Finding source, cue content and polarity in CoNNL files  
In the CoNNL files every token is matched to an identifier and a sentence id. The triples source, cue and content have an identifier that link them together. The triples of the relevant sentences found in the previous step are found by using this linking identifiers and the sentence ids. The CoNNL files also provided the polarity, positive or negative, of a sentence. For one sentence, multiple polarities according to the cue are given. The system calculates the most occuring polarity of one sentence. In case of a tie, the polarity is set to neutral.

Using a lexicon based approach to get certainty and negation  
To classify sentences based on the degree of certainty they display, a lexicon based sentiment analysis approach was used. In this approach each word in a sentence around the source is compared with two manually created lexicons one for each class. Classification of a sentence was decided based with their word match on the type of lexicon. The two types of lexicons created for this purpose contained words which define a sentence as "exhibiting certainty" or "exhibiting uncertainty". For example, words like “concluded”, “announced”, and “declared” around the source indicate that the average tone of a sentence is certain. On the contrary, words like “might”, “doubt”, and “questionably” indicate an uncertain tone of the sentence. One of the challenges with this approach was that for a lot of sentences our sentiment analysis approach for certainty didn’t work because there was no match of any source-neighboring words in the sentence with either of the lexicon dictionaries. This caused an increase in the number of "unknown" labels. To counter this we searched for all the words of a sentence in the uncertain lexicon dictionary instead of just the source-neighboring words and classify the matches as "Uncertain". This led to a rise in the number of classifications for "uncertain" class while “unknown” was classified less.

Negation detection sentiment analysis was used to classify sentences into confirm/deny labels. To achieve this a lexicon based approach was used where we manually built a lexicon containing words which defines a sentence as "deny". To implement this, every word in a sentence was searched for a match in the lexicon dictionaries and in case of a match the label "deny" was assigned to the sentences. The problem with this approach was that creating such a lexicon dictionary is a very challenging and time consuming task. To counter this, the sentiment analysis package from NLTK nltk.sentiment.vader.negated was used. nltk.sentiment.vader.negated uses similar lexicon based approach in the backend to find negation in the words. Using NLTK improved the number of tagging for negation by 30%, which is probably because of a high quality and large lexicon dictionary used by NLTK.

Determining stance by combining clustering with conformity and certainty

In order to find the stance (pro-vax, anti-vax, mix, neutral, or unknown) of each sentence, and, ultimately, of each article, a clustering technique was performed on the content of each (source, cue, content)-triple. This was done using Google’s word embedding tool Word2Vec[[1]](#footnote-0). After removing stop words, the word embedding tool was used to create a 300 feature vector for each word in each sentence, with words usually found in a similar context having a similar vector. A vector for each content sentence was computed by adding the vectors of all words in a given sentence and then dividing it by its word count.

Scikit-learn’s KMeans clustering method was considered for grouping the content sentences based on their Word2Vec vectors, but it was deemed useful to find a most representable sentence for each cluster. Hence, Scikit-learn-extra’s KMedoids clustering tool was used. For a range of K from 1 to 25, the sum of the squared distances of each point to its cluster’s medoid was computed, creating the elbow graph as shown in Figure 1. In this graph, the optimal number of clusters appears to be 2. However, since the sum of squared distances is less reliable for data with large dimensions, this may not be an accurate conclusion. Hence, the clustering was performed three times, using K=3, K=8, and K=15.

A visual inspection of the clusters suggested that some of the clusters were better than others. For example, using each of the three values for K, there was one cluster with all the sentences that included the words “vaccines” and “autism”. The clustering showed a better performance when only a small subset of the documents was used, while the clusters were more messy when more data was added. After the knowledge graph was created, it was clear that many sentences with a high word count (>10) were often mislabeled. The meaning of these long content sentences was harder to compare to other sentences based on word embedding, as their structure was more complex, giving rise to nestedness and multiple negations. Additionally, some contents were found to solely contain one word (e.g. “that”, “what”) Hence, all sentences with a word count higher than 10 or lower than 2 was removed.

The stance of the medoids of each cluster was annotated by hand (pro-vax, anti-vax, or unknown). In a later version of this algorithm, this could be automated by using word embeddings and training a machine learning algorithm on a set of annotated pro and anti-vax statements, in order to classify the medoid sentences. In the current project, all content-sentences in a cluster were classified according to the classification of the medoid of that cluster. Then, if the negation of a content was found to be “Deny”, the classification was flipped to the opposite stance. Finally, the overall stance of each document was computed by counting the pro and anti statements in its text and classifying the document with the majority stance.

**3. Improvement of extracting relevant opinions**

The first approach was to use all sentences that have an opinion in them, according to the NAF files. However, when we took a look at the results, we found that some of the opinions weren't really opinions. Words like "not" were categorised as being positive and even a single apostrophe got a categorisation. This is why we chose to find another way of extracting relevant opinions to include in the perspective graph: backtracking from the raw text which sentences had to be included in the final result. For 3 texts, we marked the relevant opinions and then we looked at the data we had on those sentences. Soon we realised that most of them had corresponding predicates that link to the same frame from FrameNet. For example, sentence a and b both have the frame "Causation".

1. *Kennedy claims the mercury is still used as a preservative and causes brain damage (chicagotribune-com\_20170918T235148)*
2. *When these kids take their well-financed school-year abroad , lack of full immunization will put them at risk for infections. (chicagotribune-com\_20170918T235148)*

Still, we didn't get all of the results we wanted, so we looked for other cues in the data, and found the frame "Preventing". We repeated this search until we found 3 frames that were quite inclusive of the sentences we wanted to use, but also excluded most sentences we deemed unimportant for the analysis: causation, preventing, statements (only if they were also annotated with an opinion). After we finished the final perspective graph, we also found a 4th frame, "Evidence", but due to temporal issues we decided not to include this.

Since we made a frame restriction in selecting the relevant opinions, a predicate (presented as cue) was the only necessary requirement in the second perspective graph. Thus, a triple without a source or content is returned as ‘Unknown’. Even though the source is unknown, it still might be relevant information. Evaluating the triples, they all contain a content but there are 89 opinions that have an unknown source. We picked out 10 random opinions and looked in the original text if this opinion could be ascribed to the author of the text. It turned out to be incorrect to ascribe this to the author. For example:

1. *It has been estimated vaccines prevent 25 % of the deaths of these children , so 75 % still die. (Child-Health-Safety\_20170626T115833)*

The text does not elaborate on whether the author or an organization estimated this. For future work, the pipeline could be improved by adding source detection.

**4. Evaluation of sentence polarity**

To quantitatively evaluate whether the system filters out the correct polarity, gold data is required. In class students annotated polarity manually, but all in a different format. Preprocessing this data would have been too time consuming. We decided to perform a qualitative evaluation on the system's output (note: at this point the system's output isn't completely filtered yet, so these sentences may not appear in the final perspective graph).

1. *If my wife had given that vaccine 5 minutes earlier , the mother would have been convinced that the vaccine caused the seizure. (@berkeleywellness\_20170709T195101)*

Example d is assigned the polarity positive, but reading this sentence it does not carry a truly positive meaning. This is one of the examples that show that the systems output was faulty. Appendix A shows a table with a more detailed qualitative analysis. In this table it can be seen that the NAF file classifies the text quite positive, although VADER analyses it as being fairly negative. Investigating why exactly some of these sentences seem to be (wrongly) categorised as positive, we found that the opinion data only annotated one word. In sentence 13, the word ‘could’. This word is positive, however VADER looks at all the words in a sentence and gives them a score, weighing them together to get a compound score for the sentence.

Finding another way of getting the polarity of the sentences was desired. We settled on doing a sentiment analysis with VADER[[2]](#footnote-1). For each sentence we computed the compound sentiment, and when it was smaller than -0.2, it got the tag negative, and for anything bigger than 0.2, it was positive. Results in between these values were assigned the tag neutral.

**5. Evaluation of stance of article**

The computed overall stance of each article when using the lexicon-based approach for negation was evaluated against the provided GOLD data, and resulted in an overall precision score of 33%. When using the negation computed by the NLTK package, a precision score of 33% was found as well. This finding means one of two things: 1) both methods for negation computation are equally accurate, or 2) there is a different bottleneck in the algorithm that has a much stronger influence on the (rather weak) performance of the algorithm. As a visual inspection of the data showed that the NLTK-based approach resulted in significantly better negation detection, we strongly suspect the latter of the two options. While the KMedoid clustering mechanism based on word embedding seems elegant, it can still be improved substantially. For example, many sentences were not grouped together even though they had a similar meaning, and most of the cluster medoids did not have a clear stance, meaning that most content-sentences got the label ‘unknown’. Additionally, the source of each content was not used, meaning that every sentence got attributed to the article itself. This might have led to many misclassifications, as the writer of an article will often mention multiple statements of the opponents of the debate. In future versions of the algorithm, the attribution of statements to groups of sources could be an important improvement.

**6. The final perspective graph: conclusion and discussion**

The first and the final perspective graphs can be found in the zipfile. According to a qualitative evaluation, the changes in the pipeline seem to improve the system. Because there is no gold data of relevant triples nor polarity this cannot be proven by a rise in accuracy, recall or f-measure. While improvement in negation was not found by evaluation of the stance on the GOLD data, a qualitative analysis showed that the negation detection did improve, but that the algorithm still lacked in other facets, including the clustering mechanism.

During this NLP pipeline, there were some encoding issues that have an effect on the results. These encoding issues arised from the input, the opinion NAF files and influence the factuality detection and grouping since they make sentences unreadable. Also, the input files contained some annotation mistakes such as words with type ‘NaN' and missing sources. Thus, whether true value can be ascribed to these results is pendable. For future work, the pipeline could be improved by adding a source detection. The sources that are now unknown, 16 percent of all found opinions, might be found by entity recognition of persons. Another improvement would be adding a fourth frame, the “Evidence” frame. During the selection process of relevant sentences this frame showed some promising results.

**7. Appendix**

|  |  |  |  |
| --- | --- | --- | --- |
| Sent\_id | Opinion NAF | VADER | Text |
| 3 | Neg | Neg | He spoke with Berkeley Wellness about why the anti-vaccine movement is so dangerous to public health . |
| 7 | Neg | Neg | What those numbers tell us is that many children are being put at risk of serious , preventable diseases . |
| 11 | Pos | Neg | Before that outbreak , there was a report by an investigative reporter named Gary Baum . |
| 13 | Pos | Neg | He wrote an article predicting that this could become the epicenter of measles outbreak . |
| 24 | Pos | Neg | When we fall short of that number , then you put children who ca n’t be immunized at risk . |
| 29 | Pos | Neg | That law passed in part because a little six-year-old boy with leukemia , who could n’t be vaccinated , stood up at a meeting and said , “ What about me ? |
| 33 | Neg | Neg | And often they get bad information from the Internet that makes them afraid of things they do n’t need to fear . |
| 35 | Neg | Neg | There was no biological explanation and virtually no data . |
| 40 | Pos | Neg | A lot of misinformation is based on anecdotal evidence . |
| 41 | Pos | Pos | Why is anecdotal evidence so powerful , even in the face of scientific findings ? |
| 42 | Pos | Neg | Anecdotal evidence is a contradiction in terms . |
| 43 | Pos | Neg | It ’s an emotional , personal experience , but it is n’t evidence . |

Table A: The first polarities extracted according to the NAF file and the VADER sentiment analysis

1. <https://code.google.com/archive/p/word2vec/> [↑](#footnote-ref-0)
2. https://github.com/cjhutto/vaderSentiment [↑](#footnote-ref-1)