Arguments as Social Good: Good Arguments in Times of Crisis

Johannes Daxenberger and Iryna Gurevych
Ubiquitous Knowledge Processing (UKP) Lab
Technische Universität Darmstadt, Germany
https://www.ukp.tu-darmstadt.de

Abstract

We report on a case study about extracting natural language arguments from news media to support decision-making in crises like the Covid-19 pandemic. In particular, we seek to detect the latest pro- and con-arguments and their trend for crisis relevant topics with the help of a combination of retrieval and machine learning. We present a prototype system that is able to uncover decision critical information about a broad range of topics. Manual analysis shows that the fully automatic system is able to retrieve arguments in real-time and with high quality.

The Covid-19 crisis presents decision-makers in politics, society and business with the challenge of having to make very quick decisions in a completely new situation under conditions that can change daily. Many of these decisions had or have a significant impact on our daily lives, like enforced lockdown of businesses and schools, mandatory face coverings, or travel restrictions. For many of these questions, little or no evidence from previous incidents is available. Consequently, any support to (more) thorough and transparent decision-making is of great use. The aim of this case study is to enable such support by extracting arguments from the broadest possible spectrum of unstructured but up-to-date web sources (in particular, news sources). As a user group, we primarily address decision-makers from politics and business, but also the general public. The result is made available through a publicly accessible web demonstrator.[1]

Our prototype is realized in the form of an argumentative search engine (Wachsmuth et al. 2017), which displays pros and cons (i.e. justified options for action) on a controversial topic or policy making in the context of the Covid-19 pandemic. Trends can be identified with a visualization that reveals the absolute pro- and con-arguments over the last months. To account for a balanced picture and to avoid potential (regional or political) bias, we include sources from all over the world. This submission describes the setup of the system and some preliminary analysis of results.

AAAI Fall 2020 Symposium on AI for Social Good.

[1] https://asg.ukp.informatik.tu-darmstadt.de

	PRO		CON	
Query	**Rel.**	**Stance**	**Rel.**	**Stance**
covid economy	0.80	0.86	0.60	0.83
face masks	1	1	1	0.80
corona tourism	0.70	1	0.70	0.57
social distancing	0.90	1	0.90	0.44
corona party	0.70	0.57	0.80	0.86
covid in schools	0.90	0.89	0.70	0.71
covid vaccination	0.80	0.63	0.90	1
quarantine	1	0.50	0.90	0.56
coronavirus protests	0	0	0.70	1
herd immunity	0.90	1	0.90	1
Avg.	**0.77**	**0.75**	**0.81**	**0.78**

Table 1: Queries and results for manual evaluation. Rel.(evance) gives the percentage of sentences relevant to the search query; Stance is a subset of the latter which is correctly classified as pro- or con-arguments. Numbers are percentages over the total number of sentences assessed.

System Description

Our system consists of three independent components: a) the retrieval component, which—given a query term—searches, downloads, parses and segments articles from the web; b) the connection to the ArgumenText API which classifies the query term and output from a) into pro-, con- or non-arguments; and c) the frontend which displays pro- and con-arguments and their trend. The components are connected through REST interfaces and can be deployed independently.

Retrieval Component

Rather than implementing our own web crawler, we make use of the GDELT project which aggregates news media from all over the world in real-time and in 65 languages.[2] GDELT offers a public full text search API giving access to their collection of news articles and blogs.[3] Given a query term, the GDELT 2.0 DOC API searches in a rolling window of the last three months of their total coverage and returns a

[2] https://www.gdeltproject.org
[3] https://blog.gdeltproject.org/gdelt-doc-2-0-api-debuts/

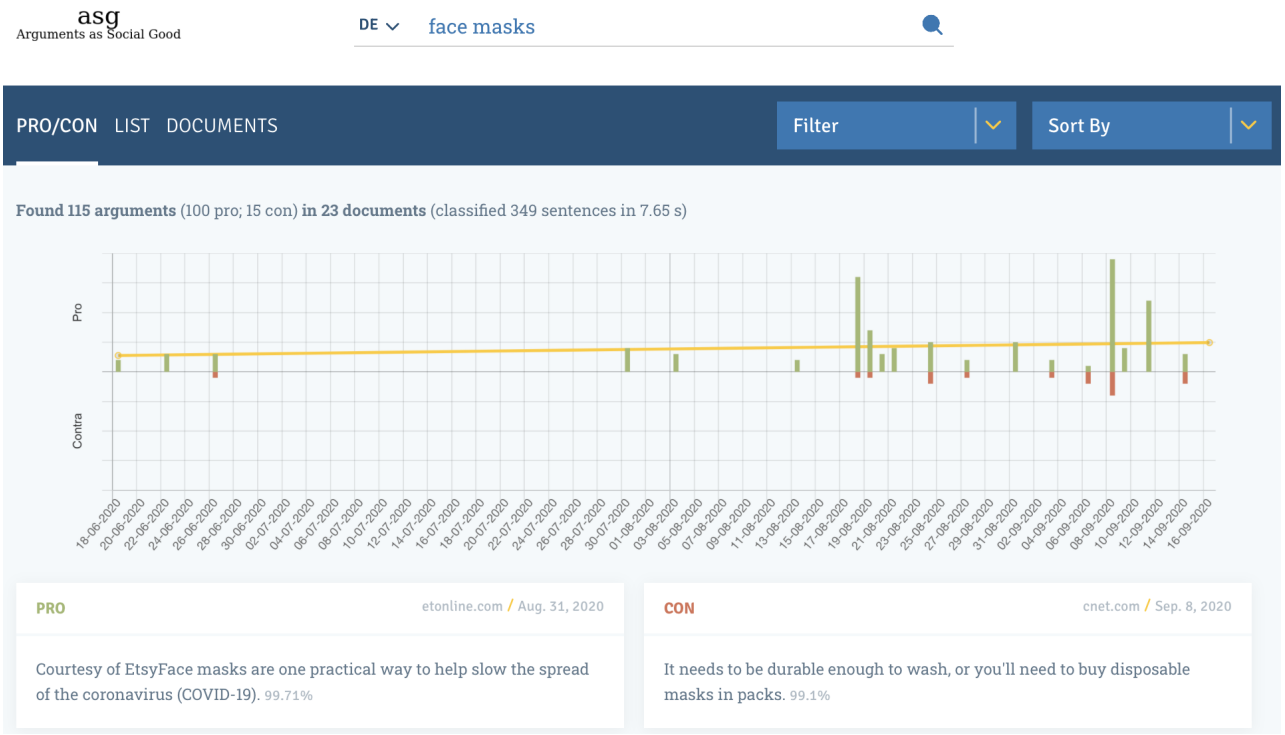

Figure 1: The current user interface of the search engine including the argument trend as bar chart and the first pro- and con-arguments discovered for the query "face masks". Screenshot taken on September 16th, 2020.

list of at most 250 URLs and metadata (e.g. timestamps) of matching web articles.

As we aim to extract arguments from the full text of the articles, we created a pipeline for scraping and parsing HTML content to plain text. Boilerplate removal to clean unwanted text elements is carried out using the Apache Tika toolkit.[4] The processing backbone of this pipeline uses DKPro Core (Eckart de Castilho and Gurevych 2014) for metadata conversion and sentence segmentation. For the sake of interoperability, the retrieval component acts as a proxy, masking the details of the underlying pipeline. It can be queried like any Elasticsearch client and returns responses similar to an Elasticsearch cluster. At the time of submission, endpoints supporting English and German queries (and responses) are available. To minimize the answer delay, articles are scraped and parsed in parallel. As a result, more than hundred pages can typically be processed in less than five seconds. A more exhaustive description of this part of the system is available in Scheunemann et al. (2020).

Argument Classification

Once relevant documents have been identified by the retrieval component, we rely on the ArgumenText API (Daxenberger et al. 2020) to further process all sentences from these documents.[5] The ArgumenText system takes an information-seeking perspective on Argument Mining (Stab et al. 2018b) and classifies any given sentence as a pro-, con- and non-argument with regard to a topic (i.e., the query, in our case). It does so using a transformer-based architecture, where the topic and sentence are jointly embedded using contextualized BERT-large embeddings (Devlin et al. 2019; Reimers et al. 2019). The data used to fine-tune the embeddings spans about 40 different topics from innovation and technology (Stab et al. 2018a) and is extracted from a large web crawl. The resulting model generalizes much better than a model trained on fewer topics. As shown by Stab et al. (2018a), a cross-topic evaluation yields 0.74 macro F1-score compared to 0.66 macro F1-score when trained on only eight topics. The ArgumenText system has been shown to cover 89% of arguments from human experts among the top-ranked results (Stab et al. 2018a). For the sake of this case study, we did not adapt the training data and model architecture. Rather, we seek to analyze the generalization capabilities of the existing model to cope with topics related to the Covid-19 pandemic.

Visualization

The final application can be accessed through a search interface which allows to specify any English or Germany query and explore resulting pro- and con-arguments in multiple ways. The appearance and handling of the frontend is based on the ArgumenText search engine[6], but additionally shows a graph highlighting the occurrence of arguments along a

[4]https://tika.apache.org

[5]https://api.argumentsearch.com/en/doc

[6]https://www.argumentsearch.com

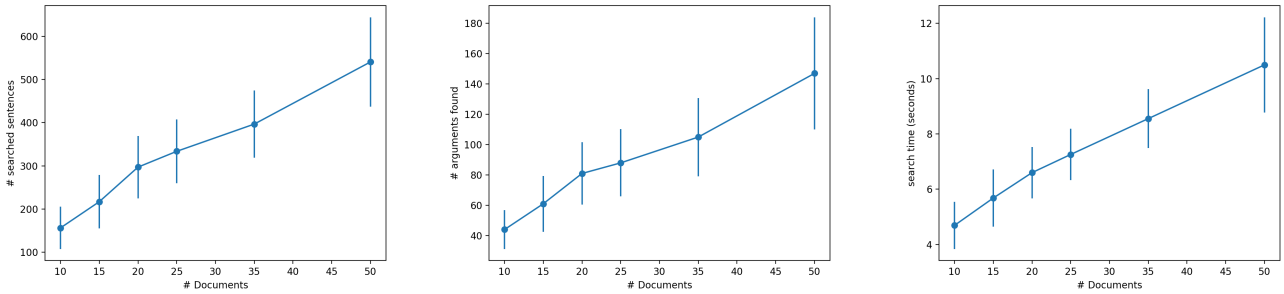

Figure 2: Total search time, number of (pro- and con-) arguments and total number of sentences searched for different input document sizes; mean and standard deviation across 10 query topics (cf. Table 1).

timeline of the last three months (see Figure 1). Absolute counts of pro- (positive) and con-arguments (negative) are shown as a bar chart including a trend line. The trend is calculated by aggregating counts in the first and second half of the full time range. The interface also allows to aggregate or filter arguments by source document.

Evaluation

To evaluate the system for the purpose of supporting decision-making in crises, we defined ten topics around social issues and policy making in the Covid-19 pandemic, as listed in Table 1. We analyzed both the argument coverage and search time as well as the relevance of the top-ranked arguments for different initial sizes of the document collection.

Argument Coverage and Search Time

As the retrieval component searches the GDELT API in real-time, both response duration and the response itself vary over time. To account for this, we repeated each query ten times with a delay of about a minute. Except for the relevancy scores in Table 1, all reported results are averaged across these ten runs. In addition to document retrieval, classification of sentences causes a delay in the overall response time. Both increase with the number of documents/sentences to be processed. Aiming to minimize search response time while covering a broad content range in a period of up to three months, we tested different initial input document sizes. The results are shown in Figure 2.

We report mean and standard deviation across the ten topics for input document sizes between 10 and 50. The number of sentences to be classified as well as the number of detected pro- and con-arguments increases steeper with up to 20 input documents (Figure 2, left and middle), however, this effect is hardly noticeable in the overall response time which increases almost linearly with the number of input documents (Figure 2, right).

Response times vary between 3 and 13 seconds. Among the ten queries considered, “social distancing” and “covid economy” are outliers with considerably more sentences to be searched while “covid in schools” has considerably less. In terms of the detected arguments “coronavirus protests” returned only 3 pro-arguments on average, while “face masks” and “herd immunity” yielded 76 and 71 (average across all topics is 39). Among con-arguments, “face masks” only gave 14 results on average, whereas “herd immunity” gives 92 (average across all topics is 49). We decided to set the default input document size to 35, giving a reasonable trade-off between answer delay and argument coverage.

Argument Relevance

For each query term (topic), we also wanted to know whether the returned sentences were i) relevant to the topic (Potthast et al. 2019) and ii) valid pro- or con-arguments with regard to the topic. In i), as a prerequisite for relevancy, the relation between the result sentence and the topic needed to be comprehensible without any further context. ii) was only assessed among relevant result sentences. To be counted as a valid argument, the sentence had to express evidence or reasoning towards the topic (Stab et al. 2018b) and the stance had to be classified correctly. For the latter, the topic is considered as an implicit claim formed as “`query` is/are not a problem” (pro) or “`query` is/are a problem” (con).

A graduate student with a background in language technology assessed the first 10 pro- and the first 10 con-arguments of the first run for each query according to these prerequisites. The results are given in Table 1 as percentage over all sentences (relevancy) and percentage over relevant sentences (stance). Around 80% of the results are relevant to the input query (with a slight advantage for con-arguments). An exception is “coronavirus protests” for which not a single relevant pro-argument was identified. Similarly, for stance, con-arguments are recognized slightly better (78% as compared to 75% for pro-arguments), but variance among the queries is higher. In most cases, low stance scores only affected either pro- or con-arguments, with the exception of “quarantine”, where many sentences were rather descriptive than argumentative.

To assess the reliability of these judgements, half of the data points were also assessed by the first author of this paper. Fleiss’ Kappa scores (Fleiss 1971) have been calculated over both pro- and con-arguments, but separately for relevancy and stance. The inter-rater agreement for relevancy is $\kappa = 0.79$ and for stance $\kappa = 0.71$. Both values are in the range of substantial agreement (Fleiss 1971), demonstrating

the reliability of the evaluation.

Conclusion and Next Steps

Our application of AI technology showcases how a combination of real-time document retrieval and fully automatic argument classification can support decision-making in crisis situations. We believe that the availability of a balanced and broad range of evidence from all over the world substantially contributes to situational awareness on critical matters, both for policy makers as well as the general public. The system is publicly available for further testing. Results of a preliminary evaluation show that instant retrieval and classification of around 100 arguments is feasible within less than 10 seconds and that the quality of the resulting arguments is high.

Next, we plan to include further document sources. In particular, we want to add scientific literature (e.g. pre-print servers) such that evidence from recent research will also be included among the pro- and con-arguments. Furthermore, we want to integrate the ArgumenText Clustering API[7], to automatically quantify predominant argumentative aspects among similar arguments (e.g. "reusability" for the query "face masks"). This will help to identify important subtopics in the discourse around the query of interest.

Acknowledgments

This research was supported through a grant by the Profile Area "Internet and Digitization" of the Technical University of Darmstadt within their "COVID-19" funding.

References

Daxenberger, J.; Schiller, B.; Stahlhut, C.; Kaiser, E.; and Gurevych, I. 2020. ArgumenText: Argument Classification and Clustering in a Generalized Search Scenario. *Datenbank-Spektrum* 20: 115–121. URL http://tubiblio.ulb.tu-darmstadt.de/121189/.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL'19*, 4171–4186. doi:10.18653/v1/N19-1423.

Eckart de Castilho, R.; and Gurevych, I. 2014. A Broad-coverage Collection of Portable NLP Components for Building Shareable Analysis Pipelines. In *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT*, 1–11. doi:10.3115/v1/W14-5201. URL https://www.aclweb.org/anthology/W14-5201.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76(5): 378–382. ISSN 00332909. doi:10.1037/h0031619. URL http://content.apa.org/journals/bul/76/5/378.

Potthast, M.; Gienapp, L.; Euchner, F.; Heilenkötter, N.; Weidmann, N.; Wachsmuth, H.; Stein, B.; and Hagen, M. 2019. Argument Search: Assessing Argument Relevance. In *SIGIR'19*, 1117–1120. doi:10.1145/3331184.3331327.

[7]https://api.argumentsearch.com/en/doc#cluster-api

Reimers, N.; Schiller, B.; Beck, T.; Daxenberger, J.; Stab, C.; and Gurevych, I. 2019. Classification and Clustering of Arguments with Contextualized Word Embeddings. In *ACL'19*, 567–578. doi:10.18653/v1/P19-1054.

Scheunemann, C.; Naumann, J.; Eichler, M.; Stowe, K.; and Gurevych, I. 2020. Data Collection and Annotation Pipeline for Social Good Projects. In *Proceedings of the AAAI Fall 2020 AI for Social Good Symposium*.

Stab, C.; Daxenberger, J.; Stahlhut, C.; Miller, T.; Schiller, B.; Tauchmann, C.; Eger, S.; and Gurevych, I. 2018a. ArgumenText: Searching for Arguments in Heterogeneous Sources. In *NAACL'18: System Demonstrations*, 21–25. URL http://tubiblio.ulb.tu-darmstadt.de/105466/.

Stab, C.; Miller, T.; Schiller, B.; Rai, P.; and Gurevych, I. 2018b. Cross-topic Argument Mining from Heterogeneous Sources. In *EMNLP'18*, 3664–3674. URL https://www.aclweb.org/anthology/D18-1402.

Wachsmuth, H.; Potthast, M.; Al-Khatib, K.; Ajjour, Y.; Puschmann, J.; Qu, J.; Dorsch, J.; Morari, V.; Bevendorff, J.; and Stein, B. 2017. Building an Argument Search Engine for the Web. In *Proceedings of the 4th Workshop on Argument Mining*, 49–59. doi:10.18653/v1/W17-5106. URL https://www.aclweb.org/anthology/W17-5106.