TrustSearch: Analyzing the Stance of Media for Promoting Critical Thinking

Andrea Simeri*[1,2], Ivan Grubišić*[2,3], Berta Chulvi*[6,4] and Paolo Rosso[2,5]

[1]DIMES,University of Calabria, Italy
[2]PHRLT. Universitat Politècnica de València, Spain
[3]Ruđer Bošković Institute, Croatia
[4]Social Psychology Department. Universitat de València, Spain
[5]Valencian Graduate School and Research Network of Artificial Intelligence (ValgrAI), Spain
[6]Symanto Research, Spain

Abstract

In the last decade, we have experienced a growing distrust in all sources of information. According to the Edelman Trust Barometer 2024, trust in information is at a record low. This lack of trust could be addressed by a new generation of search engines that promote critical thinking. A way to promote critical thinking is to offer users information about the stance of articles towards important topics and to encourage them to consume more pluralistic information. In this paper, we present the stance detection tool, which is the first result of the TrustSearch project. The goal of the project is to promote critical thinking by providing users with a new experience of browsing news articles by emphasizing the pluralistic nature of information. The tool implements two different use-cases to detect an article's stance on some controversial topics such as climate change, immigration, and vaccination against COVID-19.

Keywords

Critical thinking, stance detection, large language models, prompt engineering, entailment

1. Introduction

Distrust of the media is a serious problem for 21st century democracies. According to Edelman Trust Barometer 2024[1], a survey in 28 countries with more than 32,000 respondents, 64% of people agree with the idea that "journalists and reporters are purposely trying to mislead people by saying things they know are false or gross exaggerations." This percentage has increased by three points compared to 2023.

This widespread mistrust in information sources could be linked to the proliferation of fake news and other disinformation practices. However, there are indications that the problem we face goes beyond the quality of content and may be related to new practices in accessing information. The way we access news has changed radically in recent decades. According to the last edition of the Reuters Institute Digital News Report[2], one of the biggest implications of the shift to online news has been the weakening of the direct relationship between readers and publishers. Across 38 countries, just 29% of users say they prefer to access a website or app directly – down three percentage points from a year ago. Over half of the users (55%) prefer to access news through search engines, social media, or news aggregators, where large tech companies typically use algorithms rather than editors to select and rank stories. These search tools offer a list of news with very little information about their content. Accordingly, the user does not have much useful information to decide what to read and what not to read. At the end, this information selection activity appears to be a random action.

In order to face the problem of mistrust towards the media, we started the TrustSearch project[3]. The goal of the project is to offer a new news search experience that generates greater trust in the search itself. For the first step of the TrustSearch project, we focused on stance detection in news articles from media outlets located in the United Kingdom and Spain. We chose 15 media from each country, and we selected news on three different topics: 1) climate change, 2) vaccination strategy against COVID-19 and 3) immigration. We use stance detection as a way to offer the reader more information about the content of a particular news item. We approach the problem of stance detection from two different per-

SEPLN-CEDI-PD 2024: Seminar of the Spanish Society for Natural Language Processing: Projects and System Demonstrations,June 19-20, 2024, A Coruña, Spain

andrea.simeri@dimes.unical.it (A. Simeri*); grubisic@irb.hr (I. Grubišić*); berta.chulvi@symanto.com (B. Chulvi*); prosso@dsic.upv.es (P. Rosso)
https://it.linkedin.com/in/andrea-simeri (A. Simeri*)
0000-0002-4898-1325 (A. Simeri*)

*These authors contributed equally to this work

[1]https://www.edelman.com/trust/trust-barometer
[2]https://www.digitalnewsreport.org/survey/2019/

[3]Horizon Europe Project funded by Next Generation Internet (NGI) Search: https://www.ngi.eu/ngi-projects/ngi-search/

spectives: 1) the point of view of the general population; and 2) personalized user view. These two approaches were implemented in the interactive graphic tool that we developed as part of the project, and which are presented in the following sections of the paper.

2. Related works

Stance detection is an emerging opinion mining paradigm for various social and political applications [1]. Recently there has been a growing research interest in detecting automatically the stance from opinionated texts, especially if about controversial topics (e.g. vaccines, climate change, etc.). Stance detection is the text classification problem where the stance of the author of the text towards the target comes from this set of labels: {Favor, Against, Neither} [2]. According to Küçük and Can [2], stance detection is closely related to, but distinct from, various NLP tasks, such as: (1) sentiment analysis, (2) emotion recognition, (3) perspective identification, (4) sarcasm/irony detection, (5) controversy detection, (6) argument mining, and (7) biased language detection. The recent survey of ALDayel and Magdy (2021) [3] presents an exhaustive review of stance detection techniques, the different types of targets, the features set used, and the best performing machine learning / deep learning approaches. To solve stance detection problems, traditional machine learning approaches, such as support vector machine (SVM), logistic regression (LR), random forest (RF) and k-nearest neighbors (kNN), were mainly used until 2019 [2]. However, in recent years, deep learning approaches, such as long short-term memory (LSTM), gated recurrent unit (GRU), convolutional neural networks (CNN), graph neural networks (GNN) and transformers, have become increasingly popular [4]. The latest trend is to use large language models (LLMs) as few-shot [5] or zero-shot [6] reasoners to solve this task. LLMs are quite capable of handling the stance detection task, with proper prompt engineering they achieve the state of the art performance in zero-shot mode [7]. Furthermore, reformulating a potential NLP task into an entailment one can lead to performance improvements [8]. Addressing stance detection from an entailment perspective holds promise in capturing the underlying logical structure of arguments and opinions. By considering the entailment relationships between statements, models can potentially discern the stance expressed towards a target more accurately.

The idea of changing the depiction of information to encourage users to be exposed to a plurality of sources and opinions has been directly approached in some research, for instance in Korea and the USA. The NewsCube project [9] was developed in Korea Advanced Institute of Science and Technology with the aim of combating polarization in the media sector. The core of NewsCube was aspect level browsing, a method to provide readers with a classified view of a set of articles with different aspects. Authors also explore the effect of a depiction of information in clusters conducting three user studies. One encouraging result in this research is that by presenting the news in classified clusters according to their content (NewsCube depiction), users read significantly more articles with different points of view compared to the GoogleNews interface (list view).

A similar effort was made at UC Berkeley, which created Opinion Space [10], a self-organizing interactive visualization of an information space to encourage the reading of a greater diversity of opinions in microblogging. The need to provide a multiple perspective about events, especially if controversial, to promote more critical thinking has also captured the interest of the technology industry. In December of 2017, Bing[4] launched several new Intelligent Answers that go beyond the traditional Q&A style of search and offer answers to more complicated questions. Microsoft search engine was interested in providing multiple perspectives to an answer via a tool that offers the opposite points of views about a query. Unfortunately, this functionality is currently not available in the search engine.

To the best of our knowledge, there is not an AI tool that from one user query offers a plurality of news in different European news outlets to contrast the stance of the news. This is the goal of the first part of the TrustSearch project, and the result is the demo tool presented in this paper.

3. Detecting stance of news

In order to face the stance detection problem of news media articles regarding the topics of interest, we experiment with two different approaches. The first one, named as Definition-based approach, is based on a fixed definition of the topic. The latter, named as User-driven approach, is based on the user's previous ideas about a topic. In order to reduce trial costs, in both approaches we make predictions based on short summaries of articles rather than the full text of articles. We obtained these summaries from full-text articles using GPT-3.5-Turbo[5] model. The quality of the generated summaries was checked by three independent evaluators on 30 randomly selected samples of the test corpus and all evaluators agreed that the quality of the selected samples was satisfactory.

[4] https://www.bing.com
[5] https://platform.openai.com/docs/models/gpt-3-5

3.1. Definition-based approach

The Definition-based stance detection approach is designed to provide an assessment of an article's attitude towards a selected topic relative to the general population's view. There are three possible responses to the article's stance:

1. "The article is *in favor* w.r.t. the topic", if the article emphasizes the arguments *in favor* of the topic;
2. "The article is *against* w.r.t. the topic", if the article emphasizes the arguments *against* the topic;
3. "The article is *neutral* w.r.t. the topic", if the article does not emphasize ether of these arguments.

This approach relies on predefined topic definitions and predefined arguments in favor and against the topic. It is based on prompt engineering and utilizes large language models (LLMs) as zero-shot classifiers. According to Zhang et al. [7] and our experiments, LLMs are quite capable in solving this task, and their performance is currently the state of the art. However, there are advantages and disadvantages of using LLMs. The upside is that LLMs can produce good performance with explanations without needing a specific dataset and training for each topic. The downside is that using LLMs is more expensive than using smaller models. There are also issues specific to commercial LLMs. These models are opaque, i.e. the models' parameters are unknown. There is also a risk in terms of service continuity, the company may stop providing access to the model in the future. In the experiments presented in this paper, we use two LLMs, GPT-3.5-Turbo and GPT-4[6], and three different topics: 1) climate change, 2) vaccination strategy against COVID-19 and 3) immigration. First, to ensure that the selected LLMs had the correct conception about these topics, we assessed the quality of topic definitions generated by these LLMs. These are the definitions generated by GPT-4.

> Climate change: *"Climate change refers to long-term shifts in temperature patterns and weather conditions on Earth. It is primarily caused by human activities such as the burning of fossil fuels, deforestation, and industrial processes, which release greenhouse gases into the atmosphere. These gases, like carbon dioxide, trap heat from the sun and result in a warming effect on the planet. Climate change leads to various effects, including rising sea levels, extreme weather events, changes in ecosystems, and threats to human health and livelihoods. Taking action to reduce greenhouse gas emissions and adapt to these changes is crucial in addressing climate change."*

> Covid vaccination: *"The vaccination strategy against COVID-19 involves globally distributing and administering approved vaccines to prioritize high-risk groups such as healthcare workers, the elderly, and people with underlying health conditions. This strategy aims to reduce transmission, severe illnesses, and deaths caused by the virus. Vaccination campaigns usually follow a phased approach, starting with priority groups and gradually expanding to the general population. Additionally, strategies include ensuring vaccine supply, establishing vaccination centers, implementing public education campaigns, and monitoring vaccine safety and efficacy."*

> Immigration: *"Immigration refers to the process by which individuals enter and settle in a foreign country with the intention of residing there permanently or for an extended period. It involves crossing international borders and often requires complying with legal requirements, such as obtaining visas or work permits. Immigration can be motivated by various factors, including economic opportunities, seeking asylum or refuge, family reunification, or pursuing education and better living conditions."*

The definitions produced by the LLMs are consistent with the man-made ones. We also used the LLMs to generate arguments in favor and against these topics and both LLMs generated arguments similar to those generated by humans. Finally, we searched for appropriate prompts and confirmed that the both LLMs provide the well-formed responses needed to successfully extract the models' predictions as zero-shot classifiers.

In performance experiments, we asked LLMs to evaluate an article's stance w.r.t. the selected topic. Together with question we provide the LLMs with the summary of the article body. In addition, to explore the reasoning behind model predictions, we expand the query and make these models provide explanations to elaborate their predictions. The prediction of models were verified by 2 individuals and in most cases the estimation of LLMs were consistent with humans. Next, we explored the possibility of obtaining more comprehensive information about the stance of an article. The idea is to change the prediction type form classification into regression by providing an intensity estimate of an article stance, i.e. how much the article is *against* or *in favor* of the topic.

For convenience, we unified the stance intensity into a single score presented as continuous variable with values within interval *[-1, 1]*, where: 1) value *-1* means that an article is *strongly against* the topic; 2) value *0* means that an article is *neutral* to the topic; and value *1* means that an article is *strongly in favor* of the topic. For this reason, we extended the zero-shot classification approach with LLMs using two different strategies: 1) the sentence-based strategy; and 2) the frequency-based strategy. The sentence-based strategy, is to detect stance on the sentence level, i.e. estimating the stance class for each sentence in article's text separately. Afterwards, we calculate the final score as a weighted sum of all individual sentence stance scores, where the sentence length is used to calculate the weight/importance of the sentence and the individual sentence score is produced by transforming the sentence stance class: 1) *'in favor'* to *1*; 2) *'neutral'* to *0*; and 3) *'against'* to *-1*. The frequency-based strategy is to detect stance on each article's summary multiple times and use the frequency of classes to calculate the final intensity score. In the experiment we use 10 repetitions and calculate the mean value of all individual scores, where each individual score is given by transform-

[6] https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo

ing the stance class: 1) *'in favor'* to *1*; 2) *'neutral'* to *0*; and 3) *'against'* to *-1*. In our experiments, the sentence-based stance intensity strategy gives weak performance[7] for both LLMs, however the frequency-based strategy achieves better results that are significantly correlated with human annotations[8]. This is the reason we utilize the frequency-based strategy to produce stance intensity scores for Definition-based approach within the tool.

3.2. User-driven approach

The user-driven stance detection approach, rooted in entailment task, diverges from the Definition-based stance detection methodology by focusing on contextual nuances in user-generated content. Unlike the latter, which aims to gauge the general population's stance on a chosen topic, the user-driven approach centers around an actual user query and their previous personal perspective (i.e., the stance) on the matter. The process begins with a user query and with a newspaper article relevant to the query. The user expresses their personal stance, indicating their initial perspective on the query. Subsequently, a Large Language Model (i.e., GPT-3.5-Turbo), is employed to generate a sentence that combines the query and the user's stance leveraging on prompt engineering.

The obtained sentence, hereinafter referred to as the *user text* encapsulates the context and user's perspective, forming a cohesive statement. To explore the opposing viewpoint, we utilize the LLM to generate another sentence that represents the negation of the user text hereinafter referred to as the *opposite user text*. This sentence encapsulates the opposite stance, providing a balanced perspective for further analysis. Finally, we employ a Transformer model designed for entailment tasks. This model takes as input the *user text*, the *opposite user text*, and the text of the retrieved newspaper article, in order to assign scores indicating the likelihood of entailment for both perspectives. In this context, the goal is to determine the entailment relation between two texts. In particular, given two paragraphs, named *premise* and *hypothesis* respectively, the goal of the entailment task is to identify whether the hypothesis is true based on the information given in the premise [11]. If we call the newspaper article text as *premise* and the *user text* (or the *opposite user text*) as *hypothesis*, we will be able to obtain in output an answer to the following question: "Is the user's personal stance true based on the information contained in the article?". In this regard, we have used a encoder-decoder Transformer-based model name BART [12] since the promising performance on the entailment tasks. In particular, we have used a checkpoint for bart-large-mnli[9] after being trained on the MultiNLI (MNLI)[10] dataset, then we have applied a NLI-based Zero Shot Text Classification between two classes: the *user text* and the *opposite user text*. The input consists basically in the user's question and the user's stance. Indeed, the topic is not mandatory for this approach and the *opposite user text* can be automatically generated. The output consists of a binary classification between the aforementioned two classes. In this way we are able to obtain a clear view of the relation between the user's stance and the news article's stance promoting the critical thinking of the user. In contrast to traditional methods, this approach empowers users to actively guide the model's understanding, making it particularly adaptable to diverse contexts and individual perspectives. The emphasis on user-driven input allows for a more nuanced and tailored stance detection process. It should be noted that in this way we obtain a topic-independent approach which is totally generalized and applicable to real-world scenarios.

4. Demo

To demonstrate these two stance detection approaches and their utility in searching for relevant news, we implemented them as interactive graphical tool. The demo can be accessed directly by visiting the dedicated huggingface space[11]. The tool is composed of two interactive tabs.

The first tab is based on the Definition-based stance detection approach. When the user selects a topic, the definition of the topic will be displayed on the right side of the screen, and a list of news relevant to that topic will be available via the drop-down menu. Users can select any of these articles by clicking on the article title. Subsequently, a summary of the article will be displayed along with the estimated value of the article's stance w.r.t. the selected topic, as shown in Figure 1.

The second tab is based on the User-driven stance detection approach. In this case, the user selects a question related to the topic. Then, a real user's stance and the opposite stance will be displayed, together with a list of news relevant to the topic as mentioned previously. Eventually, the classification scores between the two classes will be displayed w.r.t. the selected topic, as shown in Figure 2.

5. Conclusions

In this paper we present a tool for discovering news articles with the respect to the stance of the article's author.

[7] Pearson correlation between predictions and human annotations is for: 1) English≈0.12; and 2) Spanish≈0.13.

[8] Pearson correlation between predictions and human annotations is for: 1) English≈0.63; and 2) Spanish≈0.48.

[9] https://huggingface.co/facebook/bart-large-mnli

[10] https://github.com/nyu-mll/multiNLI

[11] https://huggingface.co/spaces/UPV-PRHLT-NLP/TrustSearch

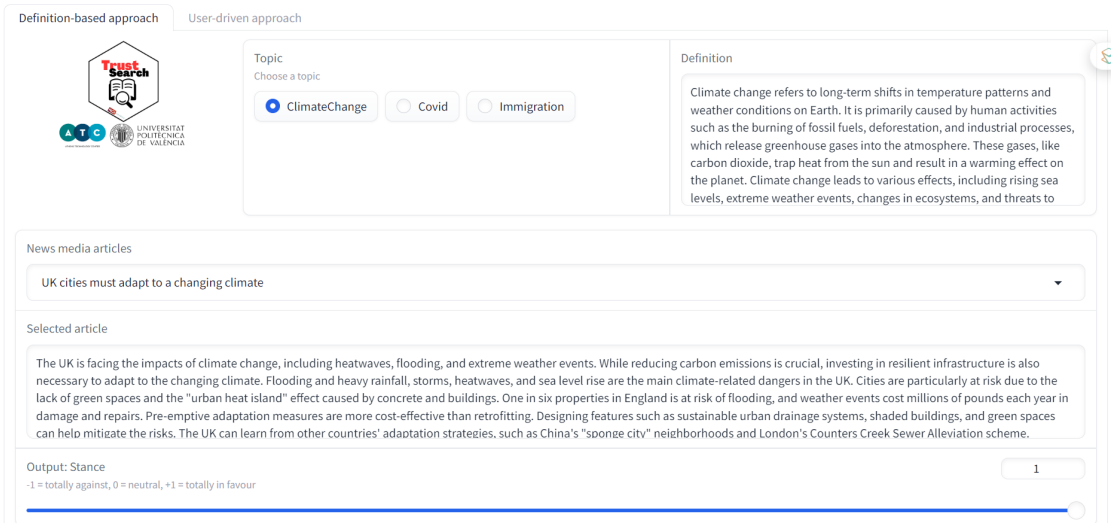

Figure 1: Example of output for the Definition-based approach.

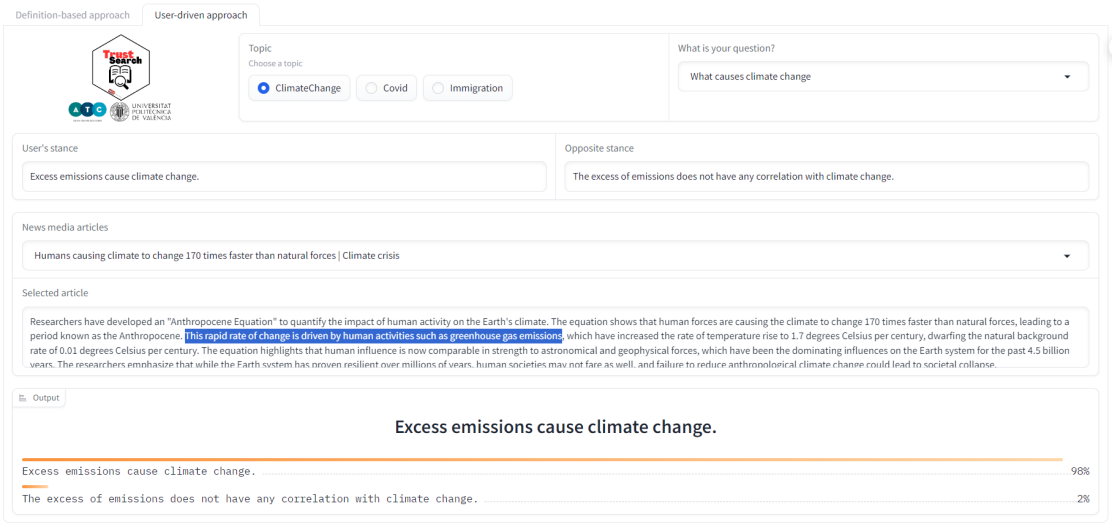

Figure 2: Example of output for the User-driven approach. It should be noted that the system correctly declares “Excess emissions cause climate change” as the most probable class with a confidence of 98%.

The tool implements two different approaches. The first one provides the article’s stance towards selected topics from a general population’s view. This approach is intended to provide users with means to explore information on topics about which they have no prior knowledge or personal views. It is also useful for exploring the views of the general population on a chosen topic. The second one is a user-centric approach which provides information on how well an article aligns with the user’s personal views.

The tool is deployed as a web application which makes it easily accessible by users. All computation is done on the server side and no installation is required on the user side. The only requirement for users is Internet access and a standard web browser installed on a computer. The tool provides a natural way of interaction and presentation of results, which makes it suitable for use by inexperienced users without additional training.

Both approaches are powered by state-of-the-art LLMs utilized in zero-shot mode to detect the stance of articles. The first approach is based on GPT-3.5-Turbo model and prompt engineering, while the second one is based on entailment and utilize BART [12] model. To the best of our knowledge, this is the first time entailment prediction has been used to detect stance of articles.

However, this pilot tool has some limitations. First, the tool currently only provides selection from a limited number of samples for each input field. But, the ultimate goal is to allow all inputs to be in free text format, which will allow for a wider application of the tool. Second, the tool displays the stance detection result only after selecting one of the articles. While this method is useful for obtaining stance information, a more effective way to explore the different views that exist in the media landscape will require offering different news together.

5.1. Future work

To address the limitations of the current version of the tool, in the the second part of TrustSearch project, we plan to apply two major updates to it. First, we plan to change the way of browsing articles from a list view to a customized graphical view, where the physical position of the article in the graph shows the stance of the article according to: 1) the selected topic or 2) personal stance. Second, we plan to combine stance detection with information on the ideological orientation of the media outlet, as well as information on its audience and its age. Finally, we plan to test the hypothesis that combining these two pieces of information, on content and on source characteristics, can help to approach search results with greater confidence and promote more critical thinking.

6. Acknowledgements

This work was carried out in the framework of TrustSearch, a project funded by NGI Search (Horizon Europe Project) Support Programme.

References

[1] M. Lai, M. Tambuscio, V. Patti, G. Ruffo, P. Rosso, Stance polarity in political debates: A diachronic perspective of network homophily and conversations on twitter, Data Knowledge Engineering 124 (2019) 101738.

[2] D. Küçük, F. Can, Stance detection: A survey, ACM Computing Surveys (CSUR) 53 (2020) 1–37. URL: https://doi.org/10.1145/3369026. doi:10.1145/3369026.

[3] A. ALDayel, W. Magdy, Stance detection on social media: State of the art and trends, Information Processing amp; Management 58 (2021) 102597. URL: http://dx.doi.org/10.1016/j.ipm.2021.102597. doi:10.1016/j.ipm.2021.102597.

[4] N. Alturayeif, H. Luqman, M. Ahmed, A systematic review of machine learning techniques for stance detection and its applications, Neural Computing and Applications 35 (2023) 5113–5144. URL: https://doi.org/10.1007/s00521-023-08285-7. doi:10.1007/s00521-023-08285-7.

[5] T. Brown, et al., Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[6] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems, volume 35, Curran Associates, Inc., 2022, pp. 22199–22213.

[7] B. Zhang, D. Ding, L. Jing, How would stance detection techniques evolve after the launch of chatgpt?, 2023. arXiv:2212.14548.

[8] S. Wang, H. Fang, M. Khabsa, H. Mao, H. Ma, Entailment as few-shot learner, CoRR abs/2104.14690 (2021). arXiv:2104.14690.

[9] S. Park, S. Kang, S. Chung, J. Song, Newscube: Delivering multiple aspects of news to mitigate media bias, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '09, Association for Computing Machinery, New York, NY, USA, 2009, p. 443–452. doi:10.1145/1518701.1518772.

[10] S. Faridani, E. Bitton, K. Ryokai, K. Goldberg, Opinion space: A scalable tool for browsing online comments, volume 2, 2010, pp. 1175–1184.

[11] I. Dagan, O. Glickman, B. Magnini, The pascal recognising textual entailment challenge, 2005, pp. 177–190. doi:https://doi.org/10.1007/11736790_9.

[12] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proc. of the 58th Annual Meeting of the Association for Computational Linguistics, ACL, Online, 2020, pp. 7871–7880.