Explainable Cross-Topic Stance Detection for Search Results

Draws, Tim; Natesan Ramamurthy, Karthikeyan; Baldini, Ioana; Dhurandhar, Amit; Padhi, Inkit; Timmermans, Benjamin; Tintarev, Nava

DOI
10.1145/3576840.3578296

Publication date
2023

Document Version
Final published version

Published in
CHIIR 2023 - Proceedings of the 2023 Conference on Human Information Interaction and Retrieval

Citation (APA)
Draws, T., Natesan Ramamurthy, K., Baldini, I., Dhurandhar, A., Padhi, I., Timmermans, B., & Tintarev, N. (2023). Explainable Cross-Topic Stance Detection for Search Results. In *CHIIR 2023 - Proceedings of the 2023 Conference on Human Information Interaction and Retrieval* (pp. 221-235). (CHIIR 2023 - Proceedings of the 2023 Conference on Human Information Interaction and Retrieval). ACM. https://doi.org/10.1145/3576840.3578296

Explainable Cross-Topic Stance Detection for Search Results

Tim Draws
Delft University of Technology
Delft, Netherlands
t.a.draws@tudelft.nl

Karthikeyan Natesan Ramamurthy
IBM Research
Yorktown Heights, NY, United States
knatesa@us.ibm.com

Ioana Baldini
IBM Research
Yorktown Heights, NY, United States
ioana@us.ibm.com

Amit Dhurandhar
IBM Research
Yorktown Heights, NY, United States
adhuran@us.ibm.com

Inkit Padhi
IBM Research
Yorktown Heights, NY, United States
inkit.padhi@ibm.com

Benjamin Timmermans
IBM
Amsterdam, Netherlands
b.timmermans@nl.ibm.com

Nava Tintarev
University of Maastricht
Maastricht, Netherlands
n.tintarev@maastrichtuniversity.nl

ABSTRACT

One way to help users navigate debated topics online is to apply *stance detection* in web search. Automatically identifying whether search results are *against*, *neutral*, or *in favor* could facilitate diversification efforts and support interventions that aim to mitigate cognitive biases. To be truly useful in this context, however, stance detection models not only need to make accurate (cross-topic) predictions but also be sufficiently explainable to users when applied to search results – an issue that is currently unclear. This paper presents a study into the feasibility of using current stance detection approaches to assist users in their web search on debated topics. We train and evaluate 10 stance detection models using a stance-annotated data set of 1204 search results. In a preregistered user study ($N = 291$), we then investigate the quality of stance detection explanations created using different explainability methods and explanation visualization techniques. The models we implement predict stances of search results across topics with satisfying quality (i.e., similar to the state-of-the-art for other data types). However, our results reveal stark differences in explanation quality (i.e., as measured by users' ability to simulate model predictions and their attitudes towards the explanations) between different models and explainability methods. A qualitative analysis of textual user feedback further reveals potential application areas, user concerns, and improvement suggestions for such explanations. Our findings have important implications for the development of user-centered solutions surrounding web search on debated topics.

CCS CONCEPTS

• **Information systems** → **Search interfaces**.

CHIIR '23, March 19–23, 2023, Austin, TX, USA

ACM ISBN 979-8-4007-0035-4/23/03.
https://doi.org/10.1145/3576840.3578296

KEYWORDS

stance detection, explainability, web search, viewpoint, bias

ACM Reference Format:
Tim Draws, Karthikeyan Natesan Ramamurthy, Ioana Baldini, Amit Dhurandhar, Inkit Padhi, Benjamin Timmermans, and Nava Tintarev. 2023. Explainable Cross-Topic Stance Detection for Search Results. In *ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '23), March 19–23, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3576840.3578296

1 INTRODUCTION

Stance detection, the task of predicting whether a document is *against*, *neutral*, or *in favor* concerning a debated topic, has received increasing attention in recent years and finds important real-world applications [8, 60, 93, 101, 118]. One such application is situated in web search: users commonly search the web for advice on important decisions surrounding debated topics (e.g., whether to become vegetarian) [15, 39, 74] but may be unaware [40, 89] that this type of interaction can be biased in several different ways [9, 30, 38, 119, 120]. For example, recent research has shown that viewpoint biases on search engine results pages (SERPs) can lead to systematic attitude change in users following whatever viewpoints are most prominent in highly ranked search results [6, 12, 33, 83, 90, 121]. Automatically identifying the stances of search results via stance detection could facilitate search result diversification efforts [78, 116] and support interventions that help users navigate online debates (e.g., by displaying warning labels for viewpoint biases) [16, 34, 97, 99, 125].

Although supplementing SERPs with automatically generated stance labels for search results is a promising step towards boosting users' ability to overcome biased information interaction and attitude change in web search [34, 82, 97], such stance labels may not reach their full potential until users understand the rationale behind them. Adding *explanations* for predictions from stance detection models (e.g., highlighting prediction-relevant words in the search result title and snippet) could assist users in navigating SERPs related to debated topics. For instance, coupling such explanations

with stance labels for search results related to the topic *vegetarianism* could help users quickly identify stance-specific patterns (i.e., what key terms and arguments are commonly brought forward on either side of the vegetarianism debate) and explicitly notice what sort of content they tend to consume. Web search interface applications designed to tackle undesired effects on users indeed often *explain* aspects of the SERP with the aim of supporting users' critical thinking [16, 34, 70, 97] because lack of reasoning is related to biased information interaction [82]. However, to the best of our knowledge, explanations for stance labels have not yet been explored in this context.

Why have stance detection explanations so far not been applied in web search on debated topics? Previous research investigating the mitigation of biased attitude change in web search users has predominantly assigned stance labels via expert annotations, crowdsourced annotations, or proxy measures instead of using automatic stance detection models [32–34, 128]. Furthermore, although there have been attempts to apply stance detection to search results [99], earlier work in this area has so far largely focused on tweets [3, 8, 24, 77], argument sentences [92, 93, 107], news articles [101], microblogs [124, 129], and online forum entries [130], and efforts to explain stance detection models have only recently begun [52]. It has thus been unclear whether and how automatic stance detection models and relevant explainable artificial intelligence (XAI) methods could be applied in the web search context. Search results and the web pages they refer to are much more diverse (e.g., concerning text length and language) and less straightforward compared to the document types typically handled by stance detection models. Moreover, explanations in natural language processing are not always easily interpretable by users [105] and it is currently not known what types of stance label explanations users would exactly require in what situations.

This paper supports the ongoing efforts toward more diverse, transparent, and trustworthy web search. We report on a preregistered[1] user study investigating whether and how explanations for automatic stance detection models can help users in their online information interactions. Two research questions guide our work:

RQ1. Are current stance detection methods sufficiently explainable for users when applied to web search results?

RQ2. What explanation visualization techniques can best explain stance detection for search results?

We address these research questions by first training and evaluating 10 different stance detection models (i.e., using classical machine learning and transformer-based language models) on a data set containing 1204 search results on 11 different debated topics (e.g., *vegetarianism*; see Sections 3 and 4). Our evaluations show satisfying predictive performances from several approaches, with *RoBERTa-base*, *BERT-base*, linear SVM, and logistic regression delivering some of the highest macro-f1 scores. We then investigate the explainability of these four models by asking participants in a preregistered user study to forward-simulate model predictions based on explanations (i.e., generated using different XAI methods and displayed as either salience-based or bar plot explanations; Section 5). We find that some model/XAI method combinations (e.g., LIME for transformer-based language models and coefficients from inherently interpretable models) can produce explanations that are sensible to users most of the time, and significantly more interpretable than randomly generated explanations. A qualitative analysis further reveals potential application areas, challenges, and improvements for such explanations. We discuss the implications and limitations of our findings in Section 7. Supplementary material related to this research (e.g., data, code, and task screenshots) is openly available: https://osf.io/fyvqu.

[1]Preregistering our user study involved openly declaring our hypotheses, experimental setup, and statistical analysis plan before data collection; see https://osf.io/nu28f.

2 RELATED WORK AND HYPOTHESES

Although users typically trust web search engines to deliver accurate and unbiased content [15, 89], search results may in reality be biased toward particular viewpoints or orientations [30, 38, 90, 119, 120]. How much SERP biases can affect users is exemplified in the *search engine manipulation effect* (SEME): users tend to change their attitude in accordance with the most prominent viewpoints among highly-ranked search results [6, 9, 12, 32, 33, 83, 121] without necessarily being aware of it [40]. Recent research has argued that such undesired outcomes root in cognitive user biases that emerge when the cognitive load exceeds users' cognitive capacities [9, 32, 82]. Indeed, *reducing* the cognitive load by re-organizing [34], summarizing [70], or explaining [34, 97, 125] elements of the SERP based on the viewpoints expressed in search results (e.g., re-ranking for greater diversity or displaying warning labels for viewpoint biases) has been shown to help users overcome adverse effects such as SEME. Such interventions have so far largely relied on manual viewpoint annotation of search results but applying them at scale requires reliable and explainable stance detection methods. Moreover, providing users with rich information about search results' stances may generally assist them in navigating debated topics online, even when no particular SERP biases or cognitive biases are at play. The remainder of this section discusses recent advances in stance detection and how it may be explained to users.

2.1 Stance Detection

Stance detection is predominantly applied in a *target-specific* fashion; i.e., a text classifier is trained and evaluated on documents that all refer to a single topic or claim (often referred to as the *target*, e.g., "people should be vegetarian") [5]. For instance, previous work has built models to detect the stance on *atheism* or the *feminist movement* in tweets [25, 60, 62, 77]. Popular stance detection tasks, data sets, and models concern document types such as tweets [2, 22, 69, 77, 110, 114], microblogs [124], online debates [1, 79, 85, 111, 115], and news content [10, 37, 49, 69, 84]; featuring a wide range of topics and several different languages [5, 60, 104]. Due to the multiclass nature of stance detection (i.e., typically classifying documents into *against*, *neutral*, and *in favor*; although sometimes additional classes such as *other/unrelated* are added [44]), predictive performances are most commonly reported in terms of macro-f1 scores [60]. State-of-the-art target-specific stance detection models (e.g., applied to tweets and online forum posts) now regularly achieve macro-f1 scores ranging from .73 to .97 depending on document type and topic [42, 54, 91, 102]. Practical target-specific stance detection applications include handling rumors [18] and *fake news* [20, 45] related to specific topics on social media. However, web search

Intellectual Property Rights and Open Source Software licenses

IPR's are originally created to protect the rights of artists. (music, literature etc.) In case of software a difference between expression and invention ...

Figure 1: Example of a salience-based explanation (using BERT-base and LIME) from our user study.

interventions targeting the mitigation of undesired effects such as SEME require target-agnostic stance detection models to quickly respond to the large variety of debated topics users may search for.

Web search applications need to apply *cross-target* stance detection. In this variant, stance detection models are applied to data sets in which each document may refer to one of a variety of topics [5, 60]. Building models that can detect stances related to *any* topic in such a way usually leads to somewhat weaker predictive performances compared to target-specific models but makes stance detection applicable at scale. Macro-f1 scores for cross-target ternary stance detection (e.g., working with tweets or news articles) have ranged – again depending on document type – roughly from .450 to .750 [4, 7, 8, 46, 93, 123]. Although stance detection has thus far not been applied to openly available search result data, some data sets feature content similar to search results. The *Emergent* data set lends itself well to cross-target stance detection and is comparable to a search result data set: it contains a large number of news articles that have each been expert-annotated as *against*, *observing*, or *in favor* concerning one of 300 rumored claims [37]. Cross-topic stance detection models evaluated at the *Emergent* data set (and its follow-up version, the *2017 Fake News Challenge* data set [84]) have achieved macro-f1 scores of up to .756 [43, 101, 108].

2.2 Explaining Text Classification

Although many methods have been proposed to explain the behavior of *natural language processing* (NLP) models generally (i.e., from abstract global explanations such as *Submodular Pick LIME* [94] and *behavioral probes* [64] to local explanations such as *SHAP* [112], *SEA* [96], or *input reduction* [36]), user-focused solutions often involve explaining specific model predictions. How a particular model prediction came about can be explained in multiple ways, e.g., by adding influential examples [57, 88] or counterfactuals [100]. Jayaram and Allaway [52] recently proposed supplementing *attention weights* with crowdsourced human rationales to explain predictions of stance detection models. Arguably the most common and straightforward way to explain specific text classification predictions, however, is to produce *input feature explanations*. These explanations consist of token-wise importance attributes [72] that can be derived from XAI methods such as *LIME* [94], *anchors* [95], *integrated gradients* [113], or Grad-CAM [106].

2.2.1 Evaluating Explanation Quality. Explanation quality can be measured in numerous ways, from *application-oriented* evaluations that focus on specific use cases (e.g., using human-annotated ground truth data sets) to *functionality-oriented* evaluations that inspect

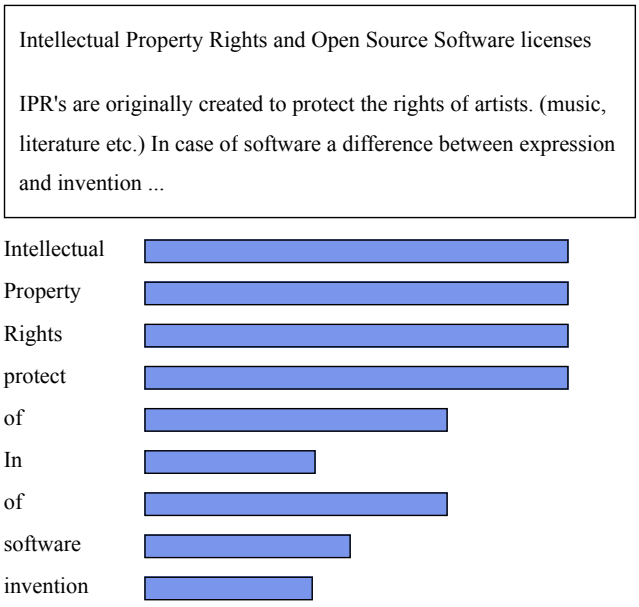

Figure 2: Example of a bar plot explanation below the search result (using BERT-base and LIME) from our user study.

how well explanations reflect a model's technical process (i.e., often referred to as *faithfulness* or *fidelity*) [27, 28, 71, 72, 87]. A commonly chosen path when aiming to evaluate explanations directly with users whilst avoiding the cost of creating a ground truth data set is to conduct *human-oriented* evaluations. These evaluation tasks typically ask users to either choose the best of several models or perform *forward simulation*, i.e., to recreate model predictions based on explanations [28, 51, 72]. Despite some earlier work pointing to a general lack of interpretability among deep learning models [17, 36], it has been demonstrated that explanations can help users simulate the predictions of artificial intelligence (AI) systems [56, 86, 127]. In the NLP domain specifically, earlier work suggests that explanations help users to better understand models [47, 80]. Jayaram and Allaway [52] created explanations for stance detection models based on human-annotated rationales and found users deemed such explanations congruent with model predictions and sufficient. We expect that users will also be able to *simulate* search result stance predictions when provided with automatically generated model explanations with greater accuracy than when provided with pseudo-explanations (i.e., a baseline that looks like a proper explanation but really only highlights words at random).

Hypothesis 1 (H1). Users can simulate the predictions of stance detection models for search results with greater accuracy when provided with a model-specific explanation than a pseudo-explanation that highlights random words.

2.2.2 Explanation Visualization Techniques. Input feature explanations are typically visualized using one of two techniques: as *salience-based explanations* that highlight words or tokens directly in the relevant document depending on their importance [21, 72, 105] (see Figure 1) or bar plots that indicate the token- or word-wise

Topic	N	Stance Distribution Against – Neutral – In Favor
Zoos	48	50% – 6% – 44%
Bottled water	48	46% – 15% – 40%
Vegetarianism	45	38% – 31% – 31%
Homework benefits	45	47% – 18% – 36%
Obesity as a disease	48	33% – 25% – 42%
Milk health benefits	49	29% – 37% – 35%
Social networking sites	50	42% – 26% – 32%
Cell phone radiation safety	50	56% – 20% – 24%
Intellectual property rights	299	13% – 19% – 69%
School uniforms	276	28% – 29% – 43%
Atheism	246	22% – 46% – 32%
Total	1204	27% – 28% – 45%

Table 1: The topic and stance distributions in our data set.

importance individually [105] (see Figure 2). Although salience-based explanations are often seen as an intuitive way to explain text classification models' predictions [21, 72], Schuff et al. [105] recently demonstrated that end users may find those explanations difficult to understand and less intuitive than bar plots. We thus expect that there will be a difference in simulatability for search results stance predictions depending on whether users see salience-based or bar plot explanations.

Hypothesis 2 (H2). Users' ability to simulate stance detection model's decisions differs depending on the way in which the explanation is visualized.

3 DATA

To train, test, and explain stance detection models, we assembled a data set containing search results on 11 debated topics (see Table 1). We obtained these data by combining three different data sets that we had created as part of earlier work [30, 32, 97]. These previously created data sets included URLs, titles, snippets, and stance labels for a total of 1453 search results, which we had retrieved via API or web crawling from two popular web search engines. Stance labels had been assigned on seven-point Likert scales (i.e., ranging from −3 to 3 and thus including three degrees of opposing or supporting a topic) via crowdsourcing in two cases (i.e., taking the median annotation of at least three crowd workers with satisfactory inter-rater reliability; Krippendorff's $\alpha = \{.78, .79\}$) [32, 97] or expert annotation in one case (i.e., mostly single annotations; Krippendorff's $\alpha = .90$) [30]. We mapped these seven-point stance labels into the three categories *against* (−3, −2, −1), *neutral* (0), and *in favor* (1, 2, 3) because automatic stance detection methods typically consider this ternary label taxonomy [60]. Using the provided URLs, we crawled the full web page text bodies (stripped of any HTML tags) for all search results. We here dropped 249 search results from the data as their text bodies could not be retrieved, leaving 1204 search results. Finally, we concatenated each search result's title, snippet, and text body (in this order) into single documents and removed all other information from the data aside from the documents' stance labels.

Table 1 shows the stance distribution per topic in our final data set. These 1204 annotated search results provide a ground truth for stance detection – both for evaluating classification performance (Section 4) and to inform a user study where participants forward simulate stance detection models' predictions based on provided explanations (Section 5).

4 SEARCH RESULT STANCE DETECTION

Explanations for stance detection models' predictions inevitably depend on the models' predictive performance. To ensure a realistic explanation pipeline in the context of search results, we first investigate the performance of current stance detection approaches and determine which methods may work particularly well here. This section thus describes the implementation and evaluation of 10 different stance detection models that we applied to our data (see Section 3). We measured the models' test set macro-accuracy, -precision, -recall, and -f1 scores across different model initializations and data splits, and compared their performance to the state of the art on other data sets (e.g., containing news articles or tweets). Finally, we selected four particularly well-performing models to generate explanations for.

4.1 Stance Detection Models

We implemented two different types of models to perform stance detection on our search result data (see Section 3): *transformer-based language models* and *classical machine learning models*. Although transformer-based language models have recently dominated text classification and other NLP tasks [41], classical machine learning models such as logistic regression continue to demonstrate competitive predictive performances while remaining highly interpretable [75, 98]. It is thus relevant to investigate the performance-explainability trade-offs between these two model types.

4.1.1 Transformer-based Language Models. We implemented five pretrained language models, fine-tuning each of them on our search result data in 10 epochs and using a learning rate of 0.00003.[2] Each model considered the first 512 tokens per document (or 1024 tokens in the case of *Longformer*).

- **BERT**-base [23]: one of the most commonly used pretrained language models [48, 67, 103] and often used for stance detection [5, 44, 46, 53, 93, 102, 104].
- **DistilBERT**-base [103]: a light version of BERT that allows for much faster fine-tuning and inference, yet often with comparable predictive performance [103]. DistilBERT has been used for stance detection before [73] and also performed well on the related task of news classification [14].
- **RoBERTa**-base [67]: an improved version of BERT that has been trained for a longer time and on more data. RoBERTa has also often been used for stance detection [44, 108, 131].
- **DeBERTa**-base [48]: another improved version of BERT that focuses on disentangling attention mechanisms. Although DeBERTa has so far not been used for stance detection, it has been implemented for the related tasks of agreement detection in online debates [85] and fake news detection [109].

[2]We tried different model types (e.g., base and large) and hyperparameter values but observed only marginal improvements beyond these settings.

- **Longformer**-base [11]: an adaptation of RoBERTa to handle long texts and thus potentially better suited for search results and the (often long-form) web pages they refer to. Whereas all above models only considered their maximum of 512 tokens, our Longformer implementation considered the first 1024 tokens per document. Longformer has already been used for rumor stance detection on different kinds of social media posts [55].

4.1.2 Classical Machine Learning Models. We applied five classical machine learning models to a *tfidf* feature matrix we had created from a preprocessed version of our data set.[3] This matrix considered all unigrams with a document frequency between 0.005 and 0.8.[4]

- **Logistic regression**: an inherently interpretable model (i.e., coefficients reflect feature importance) that has often been used for stance detection in previous research [19, 46, 50, 60, 61, 117].
- **Linear support vector machine** (linear SVM): arguably the most common stance detection approach before the advent of transformer-based language models [26, 60, 61, 65, 66, 76, 81, 117, 122]. We used linear rather than kernel SVM because it performed slightly better during testing and is inherently interpretable.
- **Random forest**: a tree-based ensemble model that is often used for stance detection [60, 61, 65, 66, 117, 117].
- **Gradient boosting**: another tree-based model commonly used for stance detection [60, 65, 117].
- **Naive Bayes**: a fully interpretable and highly simple machine learning model that has been used for stance detection in earlier work [60, 61, 66, 76] and lends itself to forming a baseline.

4.2 Evaluation

To enable a thorough and fair comparison between stance detection models, we used different random seeds to create 10 different 80-10-10 (i.e., train, validation, test) splits of our data. We then fine-tuned/trained each of the 10 models we consider (as described in Section 4.1)[5] a total of 100 times (i.e., 10 times using different random seeds that control model randomness on each of the 10 different data splits).[6] Each time we had fine-tuned/trained a model, we produced predictions for the unseen test set and subsequently computed the macro-accuracy, -precision, -recall, and -f1 score for those test set predictions. Table 2 shows each model's performance averaged over the 100 trials. To compare our results with previous research on stance detection (see Section 2.1), we focus on mean macro-f1 scores for the evaluation.

As expected, transformer-based language models (mean macro-f1 = [.647, .703]) performed considerably better than classical machine learning models (mean macro-f1 = [.570, .662]). Pairwise one-sided Wilcoxon signed-rank tests between models show that RoBERTa significantly outperformed all other models aside from DeBERTa (mean macro-f1 = .703; all $p_{\text{adj}} < 0.005$).[7] DeBERTa and Longformer both delivered strong predictive performances in most of the 100 trials but had their average scores greatly reduced by occasional bad runs (see Figure 3). This was especially surprising in the case of Longformer, as Longformer had twice as much training data available compared to the other transformer-based language models (i.e., the first 1024 instead of 512 tokens per document). Linear SVM delivered the best predictions among classical machine learning models, still outperforming all other models of this type (mean macro-f1 = .662; all $p_{\text{adj}} < 0.005$).

Model	Mean Macro- Accuracy	Precision	Recall	F1
RoBERTa	**.770** (±.004)	**.652** (±.005)	**.641** (±.005)	**.703** (±.005)
BERT	.741 (±.004)	.614 (±.005)	.598 (±.006)	.669 (±.004)
Linear SVM	.741 (±.002)	.604 (±.004)	.589 (±.003)	.662 (±.003)
DistilBERT	.737 (±.003)	.602 (±.005)	.591 (±.005)	.660 (±.004)
DeBERTa	.757 (±.007)	.609 (±.016)	.617 (±.011)	.655 (±.016)
Longformer	.747 (±.005)	.598 (±.014)	.598 (±.009)	.647 (±.015)
Logistic Regr.	.719 (±.002)	.584 (±.004)	.542 (±.003)	.642 (±.003)
Random Forest	.687 (±.003)	.551 (±.005)	.477 (±.004)	.607 (±.004)
Grad. Boosting	.668 (±.002)	.569 (±.007)	.434 (±.003)	.606 (±.005)
Naive Bayes	.651 (±.003)	.520 (±.008)	.404 (±.003)	.570 (±.006)

Table 2: Mean test set performances (± standard error) of stance detection models over 100 trials (i.e., using 10 different seeds controlling any model randomness for each of 10 different data splits; best scores in each column are bold).

Our macro-f1 scores ranging from .570 to .703 are comparable to cross-target stance detection conducted on similar (but much larger) data sets, where recent work has achieved macro-f1 scores ranging from around .450 to .750 (see Section 2.1). Moreover, the 6% performance increase from linear SVM to RoBERTa in our experiment aligns with earlier work that has found similar differences between classical machine learning models and transformer-based language models for cross-target stance detection [93].

5 USER STUDY SETUP

To investigate the explainability of stance detection models in the web search context (**RQ1** and **RQ2**), we applied several different XAI methods to four of the best-performing models we had implemented (see Section 4). Specifically, we here considered the two best-performing methods (i.e., in terms of mean macro-f1 score) from each of the two model types; that is, the top two transformer-based language models (i.e., **RoBERTa-base** and **BERT-base**; see Table 2) and the top two classical machine learning models (i.e., **linear SVM** and **logistic regression**). The motivation here was to assemble a group of models that has strong overall predictive performance and represents a broad range of existing methods, yet is small enough to efficiently conduct a meaningful user study without too many different conditions. Furthermore, although we had trained and evaluated the models using 10 different data splits (see Section 4.2), we generated explanations for only one specific scenario, i.e., using the data split where the four selected models performed best overall (see Table 3). The aim here was to reduce the complexity of explanation evaluations while maintaining comparability between stance detection models. For the non-deterministic models RoBERTa and BERT, we chose their respective best-performing

[3] Aside from removing long (>127 characters) and stop words, this preprocessing involved lemmatization and stemming (all using the `nltk` library [68]).
[4] We decided to include only unigrams here as experiments wherein we included bi- and trigrams did not show improved model performances.
[5] For models that do not need validation data for training, we added the 10% validation data to the 80% training data, thus using 90% of the data for training in these cases.
[6] For deterministic models such as naive Bayes or logistic regression, the 10 model initializations for any particular data split were identical.
[7] We Bonferroni-adjusted all p-values reported here to correct for multiple testing.

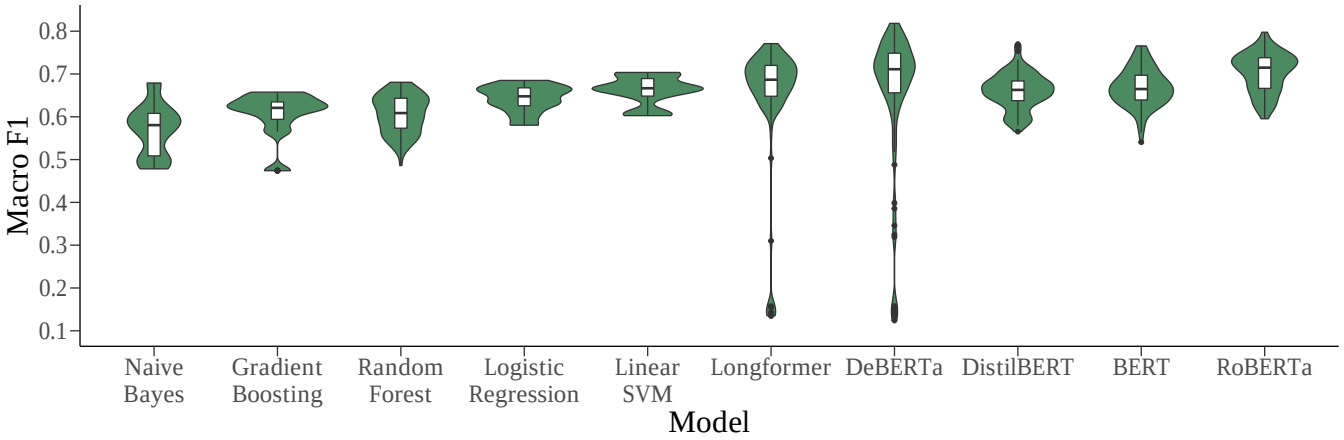

Figure 3: Distributions of macro-f1 scores across stance detection models (see also Table 2). Whereas box plots (in white) show medians and interquartile ranges, violin plots (in green) show how macro-f1 scores were distributed over the 100 runs.

Model	Mean Macro- Accuracy	Precision	Recall	F1
RoBERTa-base	.793(±.004)	.697(±.003)	.676(±.009)	.740(±.003)
BERT-base	.771(±.011)	.654(±.015)	.643(±.015)	.706(±.013)
Linear SVM	.774	.648	.629	.704
Logistic Regr.	.730	.597	.550	.656

Table 3: Mean test set performances (± standard error; except for deterministic models) over 10 random seeds on the data split where the four selected models performed best.

initializations on the selected data split. The remainder of this section describes how we created and visualized input feature explanations and evaluated their quality in a preregistered, online user study.

5.1 Materials

Input Feature Explanations. To enable an explainability comparison between the four stance detection models we selected (i.e., RoBERTa-base, BERT-base, logistic regression, and linear SVM), we created explanations for 20 test set documents (i.e., using the data split where these four models performed best overall) for which all four models made *the same stance prediction* (i.e., 10 correct and 10 incorrect predictions). This allowed us to directly compare the different explanations by looking at how many predictions users could successfully simulate. We obtained feature attributions for specific predictions from transformer-based language models by applying three different XAI methods (i.e., **integrated gradients** [113] and **Grad-CAM** [106]; both using *Captum* [58]; and **LIME** [94]). For the two classical machine learning models we considered (i.e., logistic regression and linear SVM), we obtained feature attributions from the **model coefficients** as these models are inherently interpretable. Moreover, to create a baseline, we also generated one set of **random feature attributions** for each document. Each of the 20 selected test set documents thus received a total of 3 (XAI methods) × 2 (transformer-based language models) + 2 (inherent coefficients of classical machine learning models) + 1 (random feature attributions) = 9 sets of feature attributions.

We mapped feature (token) attributions onto the original text by assigning each word the relevant token attribution (or 0 if there was none). To words that consisted of several tokens, we assigned the maximum attribution among the tokens it consisted of. We finally performed a min-max normalization on the word-wise attributions for each document to bring attributions from all methods to the same scale. This process resulted in nine sets of explanations indicating per-word importance for each of the 20 documents.

Explanation Visualization Techniques. Our aim was to visualize the nine different sets of input feature explanations per document in ways that are (1) intuitively understandable for users and (2) integratable into a search engine user interface. That is why we decided to consider not the full documents but only the title and snippet (thus only the top portion; see Section 3) of each document for the explanation visualizations, as this is what could be shown on a regular SERP. To further limit cognitive load for users and make methods better comparable, we set all negative feature attributions to 0. We created two different visualizations:

(1) **Salience-based explanations over search results** (see Figure 1) highlighted words depending on their attributions. The darker the shade of a word highlight, the greater the word's importance in the model prediction. Words whose (normalized) attributions were below a threshold of 0.25 were not highlighted.
(2) **Bar plot explanations below search results** (see Figure 2) visualized each word's attribution with a bar. The longer the bar next to a word, the greater the word's importance in the model prediction. Words whose (normalized) attributions were below a threshold of 0.25 were not listed in the bar plot.

5.2 Variables

Our study showed each participant the same set of 20 search results for which we had created explanations (see Section 5.1). However, participants saw different explanations for those search results depending on the conditions (i.e., explanation content and explanation

visualization) they had been randomly assigned to. We evaluated participants' proportion of successful simulations and additionally measured several descriptive and exploratory variables.

5.2.1 Independent Variables. These variables were used to test our hypotheses **H1** and **H2** (see Section 2.2).

- **Explanation content** (between-subjects, categorical). Each participant saw explanations stemming from only one of the nine different stance detection model/XAI method combinations we considered (i.e., integrated gradients, GradCam, or LIME explanations from either of the two transformer-based language models, coefficients from either of the two classical machine learning models, or random explanations).
- **Explanation visualization** (between-subjects, categorical). Each participant saw explanation content visualized in one of two ways: either salience-based or as bar plots.

5.2.2 Dependent Variable. Both of our hypotheses **H1** and **H2** had the same dependent variable (see Section 2.2).

- **Simulation proportion** (continuous). We recorded the number of times each participant had correctly identified the stance detection models' predictions and divided that by the total number of documents (20).

5.2.3 Descriptive and Exploratory Variables. We used these measurements to describe our sample and for exploratory analyses, but we did not conduct any conclusive hypothesis tests on them.

- **Demographics** (categorical). We asked participants to state their gender, age group, and level of education from multiple choices. Each of these items included a "prefer not to say" option.
- **Attitudes** (ordinal). We recorded participants' attitudes on each of the debated topics mentioned in the 20 search results they saw (i.e., nine of the eleven topics in Table 1) by asking participants to indicate these attitudes on seven-point Likert scales ranging from "strongly disagree" to "strongly agree".
- **Simulation rationale** (open text). We asked participants to shortly describe their rationale behind each of the 20 simulations.
- **Simulation confidence** (continuous). Participants reported their confidence in each of their simulations on a seven-point Likert scale from "extremely unconfident" to "extremely confident".
- **Explanation quality perceptions** (ordinal). We asked participants to state on seven-point Likert scales the degrees to which they (1) understood what was expected of them in this task, (2) felt that the explanations helped them understand the AI system's decisions, and (3) believe that such explanations (if they have good quality) could make a useful feature in search engines.
- **Textual feedback** (open text). We asked participants to provide feedback on the explanations in three items:
  - "Who would benefit most from stance label explanations for search results? If you don't think such explanations are helpful to anyone, why not?"
  - "In what situations do you think users would benefit from such explanations?" (optional)
  - "What would need to change for such explanations to be (more) useful in web search?" (optional)

5.3 Procedure

Participants of our study went through three subsequent steps. First, after agreeing to an informed consent, participants stated their gender, age group, and level of education. We here also asked participants for their attitudes concerning each debated topic (see Section 5.1; including one attention check where we specifically instructed participants on what option to select from a Likert scale). Second, we randomly assigned participants to one of the nine **explanation content** conditions and one of the two **explanation visualization** conditions, gave them a task introduction, and then presented them – one by one – with the 20 search results. Each search result was accompanied by one of the nine different explanations displayed using one of the two visualization techniques depending on the conditions participants had been assigned to. Below each search result, we asked participants to (1) simulate the stance detection model's prediction, (2) describe their rationale behind the simulation, and (3) state their confidence in the simulation. Third, next to another attention check, we measured participants' perceived explanation quality in three different Likert scale items and asked them to provide textual feedback (see Section 5.2).

5.4 Participants

Prior to the conducting study, we had computed a required sample size of 290 using the software *G*Power* [35] for an ANOVA; specifying the default effect size of 0.25, a significance threshold of $\alpha = \frac{0.05}{2} = 0.025$ (i.e., due to testing multiple hypotheses), a desired power of 0.8, $(9 \times 2) = 18$ groups, and the respective degrees of freedom for the two hypothesis tests (regarding **H1** and **H2**) we aimed to conduct. We eventually recruited 302 participants from *Prolific* (https://prolific.co), who were all above 18 years of age and had high proficiency in English (i.e., as reported by *Prolific*). The task was hosted on *Qualtrics* (https://www.qualtrics.com). Each participant was allowed to participate only once and rewarded $5 for completing the study (i.e., equivalent to an hourly wage of $11.26 considering the median completion time of 26:39 minutes). We excluded observations from 11 participants from data analysis because they had failed at least one of the attention checks in the task, thus leaving 291 observations to be statistically analyzed.

5.5 Statistical Analyses

To test our two hypotheses (see Section 2), we conducted an ANOVA with the two between-subjects-factors *explanation content* (to test **H1**) and *explanation visualization* (to test **H2**) as independent variables and *simulation proportion* as the dependent variable. Because we were testing two hypotheses as part of this study, we applied a Bonferroni correction to our significance threshold, reducing it to $\frac{0.05}{2} = 0.025$. We additionally conducted Tukey posthoc tests to analyze pairwise differences in case there was a main effect in the ANOVA (i.e., here thus adjusting our p-values automatically so that the significance threshold could remain at 0.05). Bayesian hypothesis tests[8] (e.g., to quantify evidence in favor of null hypotheses) and exploratory analyses (e.g., to note any unforeseen trends in the

[8]We denote Bayes Factors as BF_{10} or BF_{01} depending on whether they quantify evidence in favor of the alternative or the null hypothesis, respectively, and interpret them according to the guide proposed by Lee and Wagenmakers [63].

data) further helped us to better understand our results. Using *Atlas.ti* (https://atlasti.com), we finally conducted a *reflexive thematic* (qualitative) analysis [13] of the participants' textual answers to systematically dissect their feedback.

6 RESULTS

This section describes the results of the user study we conducted to evaluate explanations for stance detection models in the web search context (see Section 5; **RQ1** and **RQ2**). We report the results of our preregistered hypothesis tests as well as exploratory and qualitative analyses that may help interpret our findings.

6.1 Descriptive Statistics

Among the 291 recruited participants who passed both attention checks and were thus eligible for statistical analysis (see Section 5.4), 140 (48%) identified as female, 141 (49%) as male, and 9 (3%) as non-binary/third gender, while one participant (< 1%) preferred not to state their gender. Participants were rather young, with most (237; 81%) being under 35 years of age, although there were at least some participants from all age groups until 84 years. There was a diversity of education levels among participants, as only about half of them (146; 50%) had completed a university degree. While seven participants held a doctorate degree, six participants did not hold a high school diploma. Participants' attitudes on the nine debated topics present in the 20 search results they saw were reasonably balanced: across topics, there were always at least 5% who opposed and at least 20% who supported the topic. The average number of highlighted or listed words across *explanation content* conditions was 11.41 (SD = 3.62) and ranged from 8.10 (SD = 6.83, integrated gradients for RoBERTa) to 17.00 (SD = 5.01, random explanations).

Nearly all participants (270; 93%) stated that they understood what was expected from them in this task (i.e., by selecting "somewhat agree", "agree", or "strongly agree" for the relevant item). A majority of participants (216; 74%) at least somewhat agreed that the explanations helped them understand the stance detection model's predictions, with 57(20%) participants strongly agreeing and only 10(3%) participants strongly disagreeing here. Similarly, 217(75%) participants at least somewhat agreed that the explanations they saw (if they have good quality) could make a useful feature in search engines. Participants' overall mean simulation proportion across conditions was .54; slightly above a proportion of .50 that participants would have achieved had they always selected the true instead of (as instructed) the predicted stance label, as half of the shown explanations were for incorrect predictions (see Section 5.1). They reported a mean simulation confidence of 1.11 (i.e., on a scale ranging from −3/extremely unconfident to 3/extremely confident). Examining participants' simulation rationales indicated that participants indeed understood the task and were interpreting the explanations according to the highlighted or listed words (e.g., "*The word help could be a positive meaning for the AI*").

6.2 Hypothesis Tests

Figure 4 shows the mean simulation proportion per explanation content, split by explanation visualization technique. Whereas the difference between explanation types was significant (**H1**; $F = 25.615, p < .001, \eta_p^2 = .42$; see Section 5.5 for our analysis plan), the difference between explanation visualization techniques was not (**H2**; $F = .105, p = .746, \eta_p^2 < .01$). A Bayesian ANOVA further strengthened these findings, revealing extremely strong evidence for a difference between explanation types (**H1**; $BF_{10} = 4.28 \times 10^{26}$) and moderate evidence for the null hypothesis that there is no difference between visualization techniques here (**H2**; $BF_{01} = 6.36$).

Pairwise Tukey posthoc tests between explanation content conditions showed that five explanation types (i.e., coefficients for logistic regression and linear SVM, LIME for RoBERTa and BERT, and integrated gradients for BERT) led to significantly greater simulation proportions (M = [.576, .682], SE = [.019, .028]) than the random explanations (M = .452, SE = .019; $p_{adj} = [< .001, .015]$). However, there were no significant differences among these five best-performing explanation types. We also found no significant differences between the remaining three explanation types (i.e., integrated gradients for RoBERTa and Grad-CAM for both RoBERTa and BERT; M = [.373, .424], SE = [.016, .022]) and the random explanations or each other. Our results thus suggest that explanations generated from logistic regression and Linear SVM coefficients, LIME for RoBERTa and BERT, and integrated gradients for BERT lead to greater simulation proportions among users than other methods or random explanations. Moreover, these five methods all led to median simulation proportions above 0.5 (see Figure 4), indicating that most participants who saw these explanations did better than if they had tried to predict the true stance labels themselves.

6.3 Exploratory Analyses

We conducted exploratory analyses in addition to the hypothesis tests described above to better understand our results. The aim of these additional analyses is to shed light on whether the differences in simulation proportion (see Section 6.2) are reflected in participants' subjective experiences (i.e., whether the explanations were indeed helpful for participants). Note that the analyses below were not preregistered as we conducted them after inspecting the data.

6.3.1 Simulation Proportion Regarding Correct Versus Incorrect Predictions. Although the set of 20 explanations we showed to participants included equal amounts of correct and incorrect predictions and many participants' simulation proportions were greater than if they had tried to predict stance labels themselves (see Sections 6.1 and 6.2), we conducted a separate analysis to test whether participants tended to assign the true instead of (as instructed) the predicted stance label. Had this been case, participants' simulation proportions would be higher for correct than for incorrect model predictions. We thus performed a paired-samples t-test between participants' simulation proportion for the 10 correct versus the 10 incorrect predictions. Participants' mean simulation proportions were .535 and .542 for correct and incorrect predictions, respectively. This difference was not significant (delta = .007, $t = -0.525$, $p = .600$, $d = -0.03$), with a Bayesian t-test suggesting that participants simulation proportions for explanations of correct and incorrect predictions may be the same ($BF_{01} = 13.28$).

6.3.2 Relationship Between Simulation Confidence and Simulation Proportion. Our main analyses (see Section 6.2) measured explanation quality by participants' simulation proportions (i.e., reflecting the degree to which users can understand model predictions based

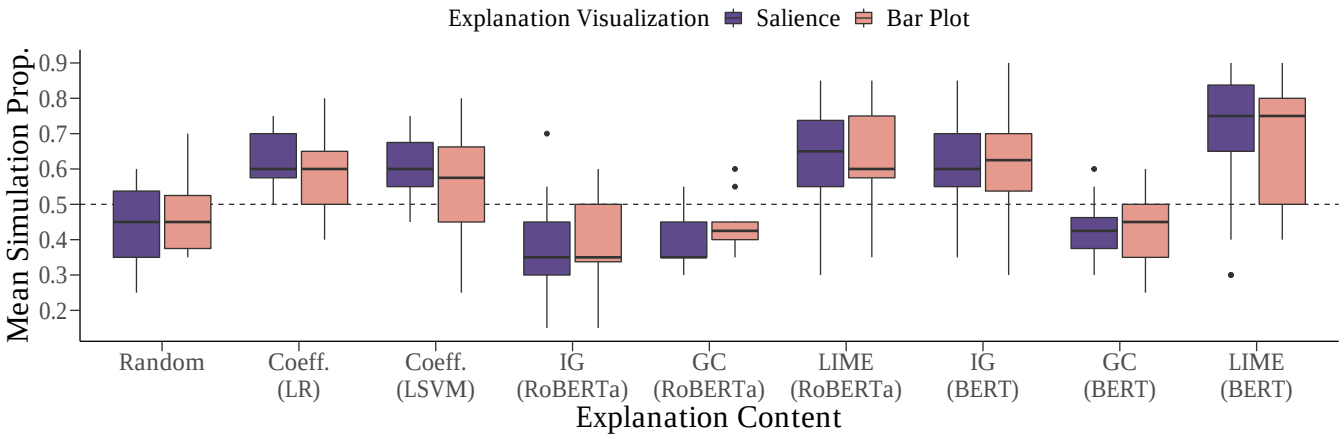

Figure 4: Mean simulation proportion per explanation content, split by explanation visualization (Coeff. = coefficients, LR = logistic regression, LSVM = linear SVM, IG = integrated gradients, GC = Grad-CAM). The dotted line reflects always selecting the true instead of (as instructed) the predicted stance label (i.e., 10 out of 20 explanations were for incorrect predictions).

on explanations), but that does not necessarily mean that participants *realized* when they correctly identified model predictions. To investigate whether participants grasped their ability to simulate model predictions, we looked at the relationship between participants' simulation proportions and their mean confidence (i.e., Likert scale items ranging from −3/extremely unconfident to 3/extremely confident; averaged over 20 items per participant). A Pearson correlation analysis revealed a significant association between these two variables ($r = .17$, $p = 0.003$), suggesting that participants were more confident in their simulations when they had stronger simulation proportions. Users thus may have a sense of their ability to make correct simulations; however, we note that this positive correlation was also rather weak. An ANOVA did not reveal any exploratory evidence for differences in participants' mean confidence across explanations ($F = 0.782$, $p = .619$, $\eta_p^2 = 0.02$) or explanation visualization techniques ($F = 1.462$, $p = .228$, $\eta_p^2 = 0.01$).

6.3.3 Differences in Explanation Quality Perceptions. Simulation proportion and confidence measure participants' ability to correctly simulate stance detection model predictions but do not necessarily speak to participants' *perceived* or *subjective* explanation quality. As with simulation confidence, we found exploratory evidence for a positive relationship between simulation proportion and the degree to which participants felt that the explanations *helped them to understand the model's predictions* ($r = .20$, $p < 0.001$). We did not find any evidence for differences between explanations or explanation visualization techniques regarding participants explanation quality perceptions, though. Given that participants' overall simulation confidence and perceived usefulness was rather high (see Section 6.1), participants across conditions may have felt that the explanations shown to them are useful even when they did not help them to successfully simulate model predictions. There was no sign of a relationship between simulation proportion and participants' perception that *explanations for search results could make a useful feature in search engines if they have a good quality*. Participants may have thus judged the general usefulness of such explanations independently from their experience in the task.

6.4 Qualitative Analyses

We conducted a qualitative, *reflected thematic analysis* [13] on participants' textual feedback to gain insights regarding where participants could see such explanations applied and what improvement suggestions they may have. To perform this analysis, one author generated response codes for participants' textual feedback in an inductive fashion and grouped them into code clusters. This resulted in the identification of **four web search scenarios** where stance label explanations could be especially helpful, **three user groups** who may particularly benefit from stance label explanations in search results, **two concerns** about such explanations, and **two ways** in which stance label explanations for search results could be **improved** according to our participants. We report on these themes below, indicating in brackets how many of our 291 participants mentioned a given theme.

Web Search Scenarios. A common theme among our participants was that explanations for search result stance labels could be used by those who intend to **research** debated topics, i.e., for school or university assignments (13), to prepare for a debate (9), to write an essay (3), or for academic work (29; e.g., "*to facilitate literature reviews*"). Participants also emphasized that stance label explanations for search results could help ordinary users in **forming opinions** by organizing the landscape of arguments on topics (26), enabling users to identify biased search results (3), and offering a diversity of viewpoints (18; e.g., "*I think that this would be a great tool for people to have the option to take a look contrasting perspectives about a subject.*") Related to this, participants believed that such explanations can lead users to **better understand** the topics or viewpoints they are searching about (8) and how search engines work (4; e.g., "[...] *why a result was given to them*"). Participants finally remarked that stance label explanations for search results deliver great **utility** by helping users to save time (46; e.g., "*it helps users to think quickly*")

and teaching them how to search in a more targeted fashion (18; e.g., "[...] *a summary in that sense would make it easier to choose what you want to actually read and spend your time on*").

User Groups. Many participants thought that search result stance label explanations could help web search users in general (54; e.g., "*I think everyone that uses search engines would benefit from these explanations* [...]"). Additionally, participants identified three main user groups for whom stance label explanations may be particularly helpful: **neurodivergent users** who have trouble comprehending complex topics (14; e.g., "*those with learning difficulties*"), **researching users** such as students (33), teachers (5), academics (56), content creators (3), debaters (1), or journalists (6; e.g., "*Journalist or researchers who need to filter a lot of material*"), and **industry users and practitioners** who work directly with stance detection models (14; e.g., "*AI/ML model auditors*") or seek to inform business decisions (7; e.g., "*people who search for quick answers and information, advertising companies and generally the marketing section* [...]").

Concerns. Despite the largely positive feedback (see also Section 6.1), participants' answers contained two themes involving concerns surrounding stance label explanations for search results. The first aspect some participants found problematic was **bad explanation quality**; specifically, participants stated that explanations missed context (1), contained overwhelming amounts of information (2), sometimes highlighted wrong or misleading words (8; e.g., "*i cant see that we can be sure they are accurate based on AI decisions*"), or were just not useful in general (7; e.g., "[...] *they are difficult to understand*"). Although we gathered such feedback from all participants, i.e., including those who saw randomly generated explanations, these comments indicate that explanation quality may be a key concern for web search users. The second problematic aspect participants saw involved the explanations' **influence on users**: they believed that explanations could induce biased behavior in users by providing too much information and thereby discouraging critical thinking (22; e.g., [...] "*it should be up to the individual to make their own mind up rather than be pushed into believing what the author writes*"). Participants were particularly concerned about users' *confirmation bias*, i.e., that stance label explanations would lead more users to just consume content they already agree with (13; e.g., "[...] *If someone is trying to prove their point (whether it is in an everyday discussion, or in science), they could be biased in finding arguments for their point of view because they could easily filter for search results that suit their opinion*"). Concerned participants were distributed across conditions, that is, we did not observe any qualitative differences regarding participants' concerns between explanation content or visualization conditions.

Improvement Suggestions. Partly in line with their concerns surrounding stance label explanations for search results, participants described two main improvement suggestion themes. One of these was rather straightforward: explanations should have **better quality**, i.e., predictions should be highly accurate and explanations should be more consistent in highlighting key terms (20; e.g., "*accuracy must be top notch*" or "*improve the keywords chosen by the AI*"), explanations should highlight words in a smart fashion (4; e.g., "*omit repeating words*" or "*Maybe linking words together* [...]"), stop words and other neutral terms should be ignored (9; e.g., "*Cut out generic words like, the and it etc.*"), and explanations should be simpler and clearer in general (7; e.g., "*just a quick guide, don't get too bogged down in details*"). Some participants, on the other hand, wished for **more extensive explanations**, i.e., supplementing search result stance label explanations with a clear labelling system or description for what makes a stance on the topic at hand (2), confidence scores for stance label predictions (2), more context (4; e.g., "samples could have been a little longer"), or just more information in general (11; e.g., "*Examples of how it works, decisions that were made based on the algorithm*"). We again observed no differences regarding improvement suggestions between conditions. As previous research has pointed out [52], a key issue for the future development of stance label explanations for search results thus seems to be trading off simplicity and clarity with providing information that is extensive enough for users to fully comprehend the stance label predictions.

7 DISCUSSION

This paper has presented a preregistered user study investigating the quality of stance label explanations for web search results. We first applied 10 different stance detection models to search result data and found that several transformer-based language models (e.g., RoBERTa and BERT) significantly outperformed classical machine learning models (e.g., linear SVM and logistic regression) in terms of predictive quality (Section 4.2). Asking user study participants to simulate 20 different stance detection model predictions based on different kinds of explanations (Section 5), we found differences between explanation types regarding participants' proportions of correctly simulated predictions (**RQ1**; Section 6). Several XAI methods (i.e., coefficients from inherently interpretable models, LIME for transformer-based language models, and integrated gradients for BERT) led to significantly higher simulation proportions than other methods or randomly generated explanations. However, we found no evidence for any differences among these best-performing explanations or between explanation visualization techniques (**RQ2**). The remainder of this section pairs these findings with results from our exploratory and qualitative analyses to paint a comprehensive picture of how web search engines could implement stance label explanations to assist their users in navigating debated topics in search results.

7.1 Implications and Recommendations

Can stance label explanations for search results be sufficiently explainable using current methods? Most participants in our user study felt that the explanations helped them understand stance detection model predictions and that such explanations could make a useful feature in web search (Section 6.1). Our hypothesis tests confirm that explanations from at least some XAI methods can lead users to better understand model predictions than randomly generated explanations (Section 6.2). Moreover, participants' simulation proportions were positively related to their simulation confidence ratings and feelings that the explanation helps them understand model predictions (Section 6.3). This suggests that simulation proportion may be a good proxy for explanation quality in the user's eye. Our qualitative analyses underlines the potential usefulness stance label explanations for search results as participants could

imagine a range of potential application areas and user groups who may particularly benefit from such explanations (Section 6.4). Given the stronger predictive performance of transformer-based language models and no apparent explainability differences between stance detection model types in this context, models such as RoBERTa and BERT, coupled with XAI methods such as LIME, may be prime candidates for this endeavor. However, participants also pointed to weaknesses and concerns surrounding search result stance label explanations that need to be dealt with for these explanations to be truly useful.

What would stance label explanations for search results ideally look like? None of our analyses (including a null hypothesis significance test; see Section 6.2) point to any difference in simulation proportion, explanation quality, or preference between the two explanation visualization techniques we had implemented (i.e., salience-based and bar plot explanations). Although our between-subjects user study design meant that we could not show both explanation visualizations to participants for direct comparisons and related research suggested otherwise [105], our findings incline us to assume that there is indeed no difference between these two methods in the web search context. Salience-based explanation visualizations over the search results may, however, still be the better option in this case as they do not require any additional space on the SERP.

Our qualitative analyses send at least two clear messages regarding the future development and implementation of search results (Section 6.4). First, explanations have to be of high quality, i.e., highlight key terms and relate them to each other while ignoring irrelevant terms such as stop words. The number of words that were highlighted in an explanation did not seem to matter to participants as two of the worst-performing explanation types featured the least and most highlighted words on average, respectively (see Section 6.1); indicating that users care primarily about the *quality* of word highlights. This not only means that predictive model performance has to be high but also that the explanation content (i.e., the word attributions) has to clearly describe the model's reasoning in a human-like way [52, 126]. Second, there is a concern that stance label explanations negatively influence user behavior and thereby contribute toward the fragmentation of society. Such concerns could be alleviated by supplementing explanations with information about stance detection and XAI methods, (cognitive) biases in web search [9], or the greater context on the topic at hand.

7.2 Limitations and Future Work

We acknowledge that our work is limited in several important ways. First, in line with most previous work on stance detection (see Section 2), we have considered a simple, ternary taxonomy for stance classification (i.e., *against*, *neutral*, *in favor*). Recent work has represented stances in more comprehensive ways (e.g., on continuous [59] or ordinal [31] scales) and supplemented them with *logics of evaluation* (i.e., reasons behind stances) [29]. Future work could explore how to predict (and subsequently explain) these more nuanced viewpoint representations. Second, for consistency, the topics in our data were all based on claims formulated in a positive direction (e.g., *in favor* on *vegetarianism* meant *supporting* the idea that one should be vegetarian; see Table 1 and Section 3). Users may get confused if they conceptualize topics in other ways (e.g., "vegetarianism is unhealthy") and find that stance labels do not match their preconceived notions (e.g., *in favor* suggesting that vegetarianism is healthy). Aside from further exploring how to explain stance predictions for search results, future work could thus also investigate how to explain debated topics and stances more generally to users and whether external factors (e.g., users' trust in different web page sources) could play a role in this context. Third, we here only looked at two different explanation visualization techniques (i.e., salience-based and bar plot explanations). Future work could explore alternative explanation formats or derive novel explanation styles that lend themselves particularly well to the web search context. Fourth, although our search results came from different search engines and featured 11 topics, we did not have much data at hand and search results had been annotated in part by experts and in part by crowd workers (see Section 3). We recommend that future work creates larger data sets of search results with high-quality human annotations for better performance of stance detection models.

8 CONCLUSION

Recent proposals towards more reliable, bias-free, and trustworthy interactions with debated topics for web search users would greatly benefit from automatic and explainable cross-topic stance detection methods. In this paper, we have presented a preregistered user study investigating the feasibility and ideal implementation of search result stance label explanations. Our findings suggest that automatic stance detection for search results is possible and promisingly show that at least some explainability methods can deliver compelling explanations to users. Moreover, our qualitative analyses reveal potential web search scenarios and user groups where such explanations could be particularly helpful but also uncover important user concerns and improvement suggestions. We hope that this work can meaningfully contribute to the ongoing efforts of understanding and mitigating undesired effects on users in web search.

ACKNOWLEDGMENTS

This activity is financed by IBM and the Allowance for Top Consortia for Knowledge and Innovation (TKI's) of the Dutch ministry of economic affairs.

REFERENCES

[1] Rob Abbott, Brian Ecker, Pranav Anand, and Marilyn Walker. 2016. Internet Argument Corpus 2.0: An SQL schema for Dialogic Social Media and the Corpora to go with it. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. European Language Resources Association (ELRA), Portorož, Slovenia, 4445–4452. https://aclanthology.org/L16-1704

[2] Aseel Addawood, Jodi Schneider, and Masooda Bashir. 2017. Stance Classification of Twitter Debates: The Encryption Debate as A Use Case. In *Proceedings of the 8th International Conference on Social Media & Society* (Toronto, ON, Canada). Association for Computing Machinery, New York, NY, USA, Article 2, 10 pages. https://doi.org/10.1145/3097286.3097288

[3] Abdulrahman I. Al-Ghadir, Aqil M. Azmi, and Amir Hussain. 2021. A novel approach to stance detection in social media tweets by fusing ranked lists and sentiments. *Information Fusion* 67 (March 2021), 29–40. https://doi.org/10.1016/j.inffus.2020.10.003

[4] Abeer Aldayel and Walid Magdy. 2019. Your stance is exposed! analysing possible factors for stance detection on social media. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–20.

[5] Abeer ALDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management* 58, 4 (July 2021), 102597. https://doi.org/10.1016/j.ipm.2021.102597

[6] Ahmed Allam, Peter Johannes Schulz, and Kent Nakamoto. 2014. The Impact of Search Engine Selection and Sorting Criteria on Vaccination Beliefs and Attitudes: Two Experiments Manipulating Google Output. *Journal of Medical Internet Research* 16, 4 (April 2014), e100. https://doi.org/10.2196/jmir.2642

[7] Emily Allaway and Kathleen McKeown. 2020. Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 8913–8931. https://doi.org/10.18653/v1/2020.emnlp-main.717

[8] Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance Detection with Bidirectional Conditional Encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computing Machinery, New York, NY, USA, 876–885.

[9] Leif Azzopardi. 2021. Cognitive Biases in Search: A Review and Reflection of Cognitive Biases in Information Retrieval. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*. ACM, Canberra ACT Australia, 27–37. https://doi.org/10.1145/3406522.3446023

[10] Ramy Baly, Mitra Mohtarami, James Glass, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. 2018. Integrating Stance Detection and Fact Checking in a Unified Corpus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 21–27. https://doi.org/10.18653/v1/N18-2004

[11] Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. https://doi.org/10.48550/ARXIV.2004.05150

[12] Markus Bink, Steven Zimmerman, and David Elsweiler. 2022. Featured Snippets and their Influence on Users' Credibility Judgements. In *ACM SIGIR Conference on Human Information Interaction and Retrieval*. ACM, Regensburg Germany, 113–122. https://doi.org/10.1145/3498366.3505766

[13] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[14] Berfu Büyüköz, Ali Hürriyetoğlu, and Arzucan Özgür. 2020. Analyzing ELMo and DistilBERT on Socio-political News Classification. In *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*. European Language Resources Association (ELRA), Marseille, France, 9–18. https://aclanthology.org/2020.aespen-1.4

[15] Noel Carroll. 2014. In Search We Trust: Exploring How Search Engines are Shaping Society. *International Journal of Knowledge Society Research* 5, 1 (Jan. 2014), 12–27. https://doi.org/10.4018/ijksr.2014010102

[16] Jon Chamberlain, Udo Kruschwitz, and Orland Hoeber. 2018. Scalable visualisation of sentiment and stance. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

[17] Arjun Chandrasekaran, Viraj Prabhu, Deshraj Yadav, Prithvijit Chattopadhyay, and Devi Parikh. 2018. Do explanations make VQA models more predictable to a human?. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 1036–1042. https://doi.org/10.18653/v1/D18-1128

[18] Yi-Chin Chen, Zhao-Yang Liu, and Hung-Yu Kao. 2017. IKM at SemEval-2017 Task 8: Convolutional Neural Networks for stance detection and rumor verification. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada, 465–469. https://doi.org/10.18653/v1/S17-2081

[19] Alessandra Teresa Cignarella, Mirko Lai, Cristina Bosco, Viviana Patti, Rosso Paolo, et al. 2020. Sardistance@ evalita2020: Overview of the task on stance detection in italian tweets. , 10 pages.

[20] Costanza Conforti, Mohammad Taher Pilehvar, and Nigel Collier. 2018. Towards Automatic Fake News Detection: Cross-Level Stance Detection in News Articles. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*. Association for Computational Linguistics, Brussels, Belgium, 40–49. https://doi.org/10.18653/v1/W18-5507

[21] Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A Survey of the State of Explainable AI for Natural Language Processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, Suzhou, China, 447–459. https://aclanthology.org/2020.aacl-main.46

[22] Kareem Darwish, Walid Magdy, and Tahar Zanouda. 2017. Improved Stance Prediction in a User Similarity Feature Space. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017* (Sydney, Australia) *(ASONAM '17)*. Association for Computing Machinery, New York, NY, USA, 145–148. https://doi.org/10.1145/3110025.3110112

[23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[24] Kuntal Dey, Ritvik Shrivastava, and Saroj Kaushik. 2017. Twitter Stance Detection — A Subjectivity and Sentiment Polarity Inspired Two-Phase Approach. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, New Orleans, LA, 365–372. https://doi.org/10.1109/ICDMW.2017.53

[25] Kuntal Dey, Ritvik Shrivastava, and Saroj Kaushik. 2018. Topical stance detection for Twitter: A two-phase LSTM model using attention. In *European Conference on Information Retrieval*. Springer, Cham, 529–536.

[26] Marcelo Dias and Karin Becker. 2016. Inf-ufrgs-opinion-mining at semeval-2016 task 6: Automatic generation of a training corpus for unsupervised identification of stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 378–383.

[27] Shuoyang Ding and Philipp Koehn. 2021. Evaluating Saliency Methods for Neural Language Models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 5034–5052. https://doi.org/10.18653/v1/2021.naacl-main.399

[28] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. https://doi.org/10.48550/ARXIV.1702.08608

[29] Tim Draws, Oana Inel, Nava Tintarev, Christian Baden, and Benjamin Timmermans. 2022. Comprehensive Viewpoint Representations for a Deeper Understanding of User Interactions With Debated Topics. In *ACM SIGIR Conference on Human Information Interaction and Retrieval*. ACM, Regensburg Germany, 135–145. https://doi.org/10.1145/3498366.3505812

[30] Tim Draws, Nirmal Roy, Oana Inel, Alisa Rieger, Rishav Hada, Mehmet Orcun Yalcin, Benjamin Timmermans, and Nava Tintarev. 2023. Viewpoint Diversity in Search Results. In *European Conference on Information Retrieval*. Springer.

[31] Tim Draws, Nava Tintarev, Ujwal Gadiraju, Alessandro Bozzon, and Benjamin Timmermans. 2021. Assessing Viewpoint Diversity in Search Results Using Ranking Fairness Metrics. *ACM SIGKDD Explorations Newsletter* 23, 1 (May 2021), 50–58. https://doi.org/10.1145/3468507.3468515

[32] Tim Draws, Nava Tintarev, Ujwal Gadiraju, Alessandro Bozzon, and Benjamin Timmermans. 2021. This Is Not What We Ordered: Exploring Why Biased Search Result Rankings Affect User Attitudes on Debated Topics. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 295–305. https://doi.org/10.1145/3404835.3462851

[33] Robert Epstein and Ronald E. Robertson. 2015. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences* 112, 33 (Aug. 2015), E4512–E4521. https://doi.org/10.1073/pnas.1419828112

[34] Robert Epstein, Ronald E. Robertson, David Lazer, and Christo Wilson. 2017. Suppressing the Search Engine Manipulation Effect (SEME). *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (Dec. 2017), 1–22. https://doi.org/10.1145/3134677

[35] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior research methods* 41, 4 (2009), 1149–1160.

[36] Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of Neural Models Make Interpretations Difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 3719–3728. https://doi.org/10.18653/v1/D18-1407

[37] William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*. ACL.

[38] Ruoyuan Gao and Chirag Shah. 2020. Toward creating a fairer ranking in search engine results. *Information Processing & Management* 57, 1 (Jan. 2020), 102138. https://doi.org/10.1016/j.ipm.2019.102138

[39] Lisa Gevelber. 2018. *It's all about 'me' —how people are taking search personally.* Technical Report. TechnicalReport.https://www.thinkwithgoogle.com/marketing-strategies

[40] Amira Ghenai, Mark D. Smucker, and Charles L.A. Clarke. 2020. A Think-Aloud Study to Understand Factors Affecting Online Health Search. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*. ACM, Vancouver BC Canada, 273–282. https://doi.org/10.1145/3343413.3377961

[41] Anthony Gillioz, Jacky Casas, Elena Mugellini, and Omar Abou Khaled. 2020. Overview of the Transformer-based Models for NLP Tasks. In *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*. IEEE, 179–183. https://doi.org/10.15439/2020F20

[42] Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. 2021. Stance Detection in COVID-19 Tweets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 1596–1611. https://doi.org/10.18653/v1/2021.acl-long.127

[43] Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M. Meyer, and Iryna Gurevych. 2018. A Retrospective Analysis of the Fake News Challenge Stance-Detection Task. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 1859–1874. https://aclanthology.org/C18-1158

[44] Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2021. Cross-Domain Label-Adaptive Stance Detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 9011–9028. https://doi.org/10.18653/v1/2021.emnlp-main.710

[45] Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2021. A survey on stance detection for mis-and disinformation identification. *arXiv preprint arXiv:2103.00242* (2021).

[46] Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2022. Few-shot cross-lingual stance detection with sentiment-based pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. AAAI, 10729–10737.

[47] Peter Hase, Shiyue Zhang, Harry Xie, and Mohit Bansal. 2020. Leakage-Adjusted Simulatability: Can Models Generate Non-Trivial Explanations of Their Behavior in Natural Language?. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 4351–4367. https://doi.org/10.18653/v1/2020.findings-emnlp.390

[48] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. https://doi.org/10.48550/ARXIV.2006.03654

[49] Tomás Hercig, Peter Krejzl, Barbora Hourová, Josef Steinberger, and Ladislav Lenc. 2017. Detecting Stance in Czech News Commentaries. *ITAT* 176 (2017), 180.

[50] Yuki Igarashi, Hiroya Komatsu, Sosuke Kobayashi, Naoaki Okazaki, and Kentaro Inui. 2016. Tohoku at SemEval-2016 task 6: Feature-based model versus convolutional neural network for stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 401–407.

[51] Alon Jacovi and Yoav Goldberg. 2020. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 4198–4205. https://doi.org/10.18653/v1/2020.acl-main.386

[52] Sahil Jayaram and Emily Allaway. 2021. Human Rationales as Attribution Priors for Explainable Stance Detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 5540–5554. https://doi.org/10.18653/v1/2021.emnlp-main.450

[53] Hema Karande, Rahee Walambe, Victor Benjamin, Ketan Kotecha, and TS Raghu. 2021. Stance detection with BERT embeddings for credibility analysis of information on social media. *PeerJ Computer Science* 7 (2021), e467.

[54] Kornraphop Kawintiranon and Lisa Singh. 2021. Knowledge enhanced masked language model for stance detection. In *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: human language technologies*.

[55] Anant Khandelwal. 2021. Fine-Tune Longformer for Jointly Predicting Rumor Stance and Veracity. In *8th ACM IKDD CODS and 26th COMAD* (Bangalore, India) *(CODS COMAD 2021)*. Association for Computing Machinery, New York, NY, USA, 10–19. https://doi.org/10.1145/3430984.3431007

[56] Been Kim, Rajiv Khanna, and Oluwasanmi Koyejo. 2016. Examples Are Not Enough, Learn to Criticize! Criticism for Interpretability. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (Barcelona, Spain) *(NIPS'16)*. Curran Associates Inc., Red Hook, NY, USA, 2288–2296.

[57] Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *International conference on machine learning*. PMLR, 1885–1894.

[58] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. 2020. Captum: A unified and generic model interpretability library for PyTorch. https://doi.org/10.48550/ARXIV.2009.07896

[59] Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P. Gummadi, and Karrie Karahalios. 2019. Search bias quantification: investigating political bias in social media and web search. *Information Retrieval Journal* 22, 1-2 (April 2019), 188–227. https://doi.org/10.1007/s10791-018-9341-2

[60] Dilek Küçük and Fazli Can. 2021. Stance Detection: A Survey. *Comput. Surveys* 53, 1 (Jan. 2021), 1–37. https://doi.org/10.1145/3369026

[61] Mirko Lai, Alessandra Teresa Cignarella, Delia Irazu Hernandez Farias, et al. 2017. itacos at ibereval2017: Detecting stance in catalan and spanish tweets. In *IberEval 2017*, Vol. 1881. CEUR-WS. org, 185–192.

[62] Mirko Lai, Viviana Patti, Giancarlo Ruffo, and Paolo Rosso. 2018. Stance evolution and twitter interactions in an italian political debate. In *International Conference on Applications of Natural Language to Information Systems*. Springer, 15–27.

[63] Michael D Lee and Eric-Jan Wagenmakers. 2014. *Bayesian cognitive modeling: A practical course*. Cambridge university press.

[64] Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics* 4 (2016), 521–535.

[65] Can Liu, Wen Li, Bradford Demarest, Yue Chen, Sara Couture, Daniel Dakota, Nikita Haduong, Noah Kaufman, Andrew Lamont, Manan Pancholi, et al. 2016. Iucl at semeval-2016 task 6: An ensemble model for stance detection in twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 394–400.

[66] Liran Liu, Shi Feng, Daling Wang, and Yifei Zhang. 2016. An empirical study on Chinese microblog stance detection using supervised and semi-supervised machine learning methods. In *Natural Language Understanding and Intelligent Applications*. Springer, 753–765.

[67] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. https://doi.org/10.48550/ARXIV.1907.11692

[68] Edward Loper and Steven Bird. 2002. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1* (Philadelphia, Pennsylvania) *(ETMTNLP '02)*. Association for Computational Linguistics, USA, 63–70. https://doi.org/10.3115/1118108.1118117

[69] Nikita Lozhnikov, Leon Derczynski, and Manuel Mazzara. 2018. Stance prediction for russian: data and analysis. In *International Conference in Software Engineering for Defence Applications*. Springer, 176–186.

[70] Ramona Ludolph, Ahmed Allam, and Peter J Schulz. 2016. Manipulating Google's Knowledge Graph Box to Counter Biased Information Processing During an Online Search on Vaccination: Application of a Technological Debiasing Strategy. *Journal of Medical Internet Research* 18, 6 (June 2016), e137. https://doi.org/10.2196/jmir.5430

[71] Andreas Madsen, Nicholas Meade, Vaibhav Adlakha, and Siva Reddy. 2021. Evaluating the Faithfulness of Importance Measures in NLP by Recursively Masking Allegedly Important Tokens and Retraining. https://doi.org/10.48550/ARXIV.2110.08412

[72] Andreas Madsen, Siva Reddy, and Sarath Chandar. 2021. Post-hoc Interpretability for Neural NLP: A Survey. https://doi.org/10.48550/ARXIV.2108.04840

[73] Matthew Matero, Nikita Soni, Niranjan Balasubramanian, and H Andrew Schwartz. 2021. MeLT: Message-Level Transformer with Masked Document Representations as Pre-Training for Stance Detection. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. 2959–2966.

[74] Dana Mckay, Stephann Makri, Marisela Gutierrez-Lopez, Andrew MacFarlane, Sondess Missaoui, Colin Porlezza, and Glenda Cooper. 2020. We are the Change that we Seek: Information Interactions During a Change of Viewpoint. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*. ACM, Vancouver BC Canada, 173–182. https://doi.org/10.1145/3343413.3377975

[75] Vincent Menger, Floor Scheepers, and Marco Spruit. 2018. Comparing deep learning and classical machine learning approaches for predicting inpatient violence incidents from clinical text. *Applied Sciences* 8, 6 (2018), 981.

[76] Amita Misra, Brian Ecker, Theodore Handleman, Nicolas Hahn, and Marilyn Walker. 2016. NLDS-UCSC at SemEval-2016 Task 6: A Semi-Supervised Approach to Detecting Stance in Tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 420–427.

[77] Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*. 31–41.

[78] Mats Mulder, Oana Inel, Jasper Oosterman, and Nava Tintarev. 2021. Operationalizing Framing to Support Multiperspective Recommendations of Opinion Pieces. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Virtual Event Canada, 478–488. https://doi.org/10.1145/3442188.3445911

[79] Akiko Murakami and Rudy Raymond. 2010. Support or oppose? Classifying positions in online debates from reply activities and opinion expressions. In *Coling 2010: Posters*. 869–875.

[80] Dong Nguyen. 2018. Comparing Automatic and Human Evaluation of Local Explanations for Text Classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 1069–1078. https://doi.org/10.18653/v1/N18-1097

[81] Braja Gopal Patra, Dipankar Das, and Sivaji Bandyopadhyay. 2016. JU_NLP at SemEval-2016 task 6: detecting stance in tweets using support vector machines. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 440–444.

[82] Gordon Pennycook and David G. Rand. 2019. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated

reasoning. *Cognition* 188 (July 2019), 39–50. https://doi.org/10.1016/j.cognition.2018.06.011

[83] Frances A. Pogacar, Amira Ghenai, Mark D. Smucker, and Charles L.A. Clarke. 2017. The Positive and Negative Influence of Search Results on People's Decisions about the Efficacy of Medical Treatments. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*. ACM, Amsterdam The Netherlands, 209–216. https://doi.org/10.1145/3121050.3121074

[84] Dean Pomerleau and Delip Rao. 2017. Fake news challenge stage 1 (FNC-I): Stance detection. https://fakenewschallenge.org

[85] John Pougué-Biyong, Valentina Semenova, Alexandre Matton, Rachel Han, Aerin Kim, Renaud Lambiotte, and Doyne Farmer. 2021. DEBAGREEMENT: A comment-reply dataset for (dis) agreement detection in online debates. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

[86] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–52.

[87] Danish Pruthi, Rachit Bansal, Bhuwan Dhingra, Livio Baldini Soares, Michael Collins, Zachary C Lipton, Graham Neubig, and William W Cohen. 2022. Evaluating Explanations: How much do explanations from the teacher aid students? *Transactions of the Association for Computational Linguistics* 10 (2022), 359–375.

[88] Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. 2020. Estimating training data influence by tracing gradient descent. *Advances in Neural Information Processing Systems* 33 (2020), 19920–19930.

[89] Kristen Purcell, Lee Rainie, and Joanna Brenner. 2012. Search engine use 2012. (2012).

[90] Cornelius Puschmann. 2019. Beyond the Bubble: Assessing the Diversity of Political Search Results. *Digital Journalism* 7, 6 (July 2019), 824–843. https://doi.org/10.1080/21670811.2018.1539626

[91] Pavani Rajula, Chia-Chien Hung, and Simone Paolo Ponzetto. 2022. Stacked Model based Argument Extraction and Stance Detection using Embedded LSTM model. *Working Notes Papers of the CLEF* (2022).

[92] Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and Clustering of Arguments with Contextualized Word Embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 567–578. https://doi.org/10.18653/v1/P19-1054

[93] Myrthe Reuver, Suzan Verberne, Roser Morante, and Antske Fokkens. 2021. Is Stance Detection Topic-Independent and Cross-topic Generalizable? - A Reproduction Study. In *Proceedings of the 8th Workshop on Argument Mining*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 46–56. https://doi.org/10.18653/v1/2021.argmining-1.5

[94] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[95] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[96] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging NLP models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

[97] Alisa Rieger, Tim Draws, Nava Tintarev, and Mariet Theune. 2021. This Item Might Reinforce Your Opinion: Obfuscation and Labeling of Search Results to Mitigate Confirmation Bias. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media (HT '21)*. Association for Computing Machinery, New York, NY, USA, 189–199. https://doi.org/10.1145/3465336.3475101

[98] Kevin Roitero, Cristian Bozzato, Vincenzo Della Mea, Stefano Mizzaro, and Giuseppe Serra. 2020. Twitter goes to the Doctor: Detecting Medical Tweets using Machine Learning and BERT.. In *SIIRH@ ECIR*.

[99] Mir Rosenberg. 2018. Toward a More Intelligent Search: Bing Multi-Perspective Answers. https://blogs.bing.com/search-quality-insights/february-2018/Toward-a-More-Intelligent-Search-Bing-Multi-Perspective-Answers

[100] Alexis Ross, Ana Marasović, and Matthew E Peters. 2021. Explaining NLP Models via Minimal Contrastive Editing (MiCE). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 3840–3852.

[101] Arjun Roy, Pavlos Fafalios, Asif Ekbal, Xiaofei Zhu, and Stefan Dietze. 2022. Exploiting Stance Hierarchies for Cost-Sensitive Stance Detection of Web Documents. *J. Intell. Inf. Syst.* 58, 1 (feb 2022), 1–19. https://doi.org/10.1007/s10844-021-00642-z

[102] Younes Samih and Kareem Darwish. 2021. A Few Topical Tweets are Enough for Effective User Stance Detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, Online, 2637–2646. https://doi.org/10.18653/v1/2021.eacl-main.227

[103] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. https://doi.org/10.48550/ARXIV.1910.01108

[104] Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Stance detection benchmark: How robust is your stance detection? *KI-Künstliche Intelligenz* 35, 3 (2021), 329–341.

[105] Hendrik Schuff, Alon Jacovi, Heike Adel, Yoav Goldberg, and Ngoc Thang Vu. 2022. Human Interpretation of Saliency-based Explanation Over Text. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 611–636. https://doi.org/10.1145/3531146.3533127

[106] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.

[107] Anirban Sen, Manjira Sinha, Sandya Mannarswamy, and Shourya Roy. 2018. Stance classification of multi-perspective consumer health information. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*. ACM, Goa India, 273–281. https://doi.org/10.1145/3152494.3152518

[108] Robiert Sepúlveda-Torres, Marta Vicente, Estela Saquete, Elena Lloret, and Manuel Palomar. 2021. Exploring summarization to enhance headline stance detection. In *International Conference on Applications of Natural Language to Information Systems*. Springer, 243–254.

[109] S. M. Sadiq-Ur-Rahman Shifath, Mohammad Faiyaz Khan, and Md. Saiful Islam. 2021. A transformer based approach for fighting COVID-19 fake news. https://doi.org/10.48550/ARXIV.2101.12027

[110] Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. 2017. A dataset for multi-target stance detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. 551–557.

[111] Swapna Somasundaran and Janyce Wiebe. 2010. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*. 116–124.

[112] Mukund Sundararajan and Amir Najmi. 2020. The many Shapley values for model explanation. In *International conference on machine learning*. PMLR, 9269–9278.

[113] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*. PMLR, 3319–3328.

[114] Mariona Taulé, M Antonia Martí, Francisco M Rangel, Paolo Rosso, Cristina Bosco, Viviana Patti, et al. 2017. Overview of the task on stance and gender detection in tweets on Catalan independence at IberEval 2017. In *2nd Workshop on Evaluation of Human Language Technologies for Iberian Languages, IberEval 2017*, Vol. 1881. CEUR-WS, 157–177.

[115] Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the Vote: Determining Support or Opposition from Congressional Floor-Debate Transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (Sydney, Australia) *(EMNLP '06)*. Association for Computational Linguistics, USA, 327–335.

[116] Nava Tintarev, Emily Sullivan, Dror Guldin, Sihang Qiu, and Daan Odjik. 2018. Same, Same, but Different: Algorithmic Diversification of Viewpoints in News. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization*. ACM, Singapore Singapore, 7–13. https://doi.org/10.1145/3213586.3226203

[117] Martin Tutek, Ivan Sekulić, Paula Gombar, Ivan Paljak, Filip Čulinović, Filip Boltužić, Mladen Karan, Domagoj Alagić, and Jan Šnajder. 2016. Takelab at semeval-2016 task 6: Stance classification in tweets using a genetic algorithm based ensemble. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*. 464–468.

[118] Rui Wang, Deyu Zhou, Mingmin Jiang, Jiasheng Si, and Yang Yang. 2019. A Survey on Opinion Mining: From Stance to Product Aspect. *IEEE Access* 7 (2019), 41101–41124. https://doi.org/10.1109/ACCESS.2019.2906754

[119] Ryen White. 2013. Beliefs and biases in web search. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, Dublin Ireland, 3–12. https://doi.org/10.1145/2484028.2484053

[120] Ryen W. White and Ahmed Hassan. 2014. Content Bias in Online Health Search. *ACM Transactions on the Web* 8, 4 (Nov. 2014), 1–33. https://doi.org/10.1145/2663355

[121] Ryen W. White and Eric Horvitz. 2015. Belief Dynamics and Biases in Web Search. *ACM Transactions on Information Systems* 33, 4 (May 2015), 1–46. https://doi.org/10.1145/2746229

[122] Michael Wojatzki and Torsten Zesch. 2016. ltl. uni-due at semeval-2016 task 6: Stance detection in social media using stacked classifiers. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 428–433.

[123] Chang Xu, Cécile Paris, Surya Nepal, and Ross Sparks. 2018. Cross-Target Stance Classification with Self-Attention Networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Melbourne, Australia, 778–783. https://doi.org/10.18653/v1/P18-2123

[124] Ruifeng Xu, Yu Zhou, Dongyin Wu, Lin Gui, Jiachen Du, and Yun Xue. 2016. Overview of nlpcc shared task 4: Stance detection in chinese microblogs. In *Natural language understanding and intelligent applications*. Springer, 907–916.

[125] Yusuke Yamamoto and Satoshi Shimada. 2016. Can Disputed Topic Suggestion Enhance User Consideration of Information Credibility in Web Search?. In *Proceedings of the 27th ACM Conference on Hypertext and Social Media*. ACM, Halifax Nova Scotia Canada, 169–177. https://doi.org/10.1145/2914586.2914592

[126] Scott Cheng-Hsin Yang, Nils Erik Tomas Folke, and Patrick Shafto. 2022. A psychological theory of explainability. In *International Conference on Machine Learning*. PMLR, 25007–25021.

[127] Arnold Yeung, Shalmali Joshi, Joseph Jay Williams, and Frank Rudzicz. 2020. Sequential explanations with mental model-based policies.

[128] Elad Yom-Tov, Susan Dumais, and Qi Guo. 2014. Promoting Civil Discourse Through Search Engine Diversity. *Social Science Computer Review* 32, 2 (April 2014), 145–154. https://doi.org/10.1177/0894439313506838

[129] Nan Yu, Da Pan, Meishan Zhang, and Guohong Fu. 2016. Stance detection in Chinese microblogs with neural networks. In *Natural Language Understanding and Intelligent Applications*. Springer, 893–900.

[130] Shaodian Zhang, Lin Qiu, Frank Chen, Weinan Zhang, Yong Yu, and Noémie Elhadad. 2017. We Make Choices We Think Are Going to Save Us: Debate and Stance Identification for Online Breast Cancer CAM Discussions. In *Proceedings of the 26th International Conference on World Wide Web Companion* (Perth, Australia) *(WWW '17 Companion)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1073–1081. https://doi.org/10.1145/3041021.3055134

[131] Guangzhen Zhao and Peng Yang. 2020. Pretrained Embeddings for Stance Detection with Hierarchical Capsule Network on Social Media. *ACM Trans. Inf. Syst.* 39, 1, Article 1 (sep 2020), 32 pages. https://doi.org/10.1145/3412362