Identifying Argumentative Questions in Web Search Logs

Yamen Ajjour
Leipzig University
Bauhaus-Universität Weimar
Leipzig, Germany

Pavel Braslavski
Ural Federal University, Yekaterinburg
HSE University, Moscow
Russia

Alexander Bondarenko
Martin-Luther-Universität Halle-Wittenberg
Halle, Germany

Benno Stein
Bauhaus-Universität Weimar
Weimar, Germany

ABSTRACT

We present an approach to identify argumentative questions among web search queries. Argumentative questions ask for reasons to support a certain stance on a controversial topic, such as "Should marijuana be legalized?" Controversial topics entail *opposing* stances, and hence can be supported or opposed by various arguments. Argumentative questions pose a challenge for search engines since they should be answered with both pro and con arguments in order to not bias a user toward a certain stance.

To further analyze the problem, we sampled questions about 19 controversial topics from a large Yandex search log and let human annotators label them as one of *factual*, *method*, or *argumentative*. The result is a collection of 39,340 labeled questions, 28% of which are argumentative, demonstrating the need to develop dedicated systems for this type of questions. A comparative analysis of the three question types shows that asking for reasons and predictions are among the most important features of argumentative questions. To demonstrate the feasibility of the classification task, we developed a BERT-based classifier to map questions to the question types, reaching a promising macro-averaged F_1-score of 0.78.

CCS CONCEPTS

• **Information systems → Query log analysis**.

KEYWORDS

Argumentative Questions; Web Search Log; Crowdsourcing

ACM Reference Format:
Yamen Ajjour, Pavel Braslavski, Alexander Bondarenko, and Benno Stein. 2022. Identifying Argumentative Questions in Web Search Logs. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3477495.3531864

SIGIR '22, July 11–15, 2022, Madrid, Spain

ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3531864

1 INTRODUCTION

Search engine users may take opposite stances on controversial topics, such as, whether to ban or legalize marijuana. To satisfy their corresponding information needs, they ask questions ranging from those that look for facts to those that look for reasons to support a certain stance, e.g., "Should marijuana be banned?" Search engines are less effective in answering the latter, non-factual questions, compared to factual ones [10]. Moreover, search results for queries related to controversial topics tend to be a source of bias [19].

Argument retrieval aims at building systems that help users form informed and unbiased opinions about controversial topics [14, 18, 31]. A typical argument retrieval system integrates methods for argument mining [2, 15], stance classification [5], argument clustering [17, 30], as well as a dedicated search results interface (e.g., side-by-side pro and con arguments [1]). Extracted arguments are ranked by their relevance to the input query and persuasiveness [29]. Classifying the stances of retrieved arguments into pro and con gives the user an opportunity to analyze conflicting opinions and formulate their own opinion on the topic in an unbiased way. Retrieving arguments on a controversial topic promotes transparency, since arguments not only support a position on such a topic, but also include the justification for this position.

Equipping search engines with bias-aware technologies became a necessity in times of fake news and misinformation, especially for controversial topics. Integrating argument retrieval systems in web search requires identifying questions that look for arguments (*argumentative questions*) in the query stream. Existing question taxonomies for non-factual questions include different question types among which is *opinion*—questions that ask for personal experiences and judgments [25, 27]. Even though argumentative questions are close to opinion questions, they differ in many aspects, e.g., they demand reasons and evidence and benefit from structured result presentation (pro vs. con arguments).

In this paper, we address the task of identifying questions that look for arguments—argumentative questions—using a two-step annotation scheme. Our annotation scheme simplifies the task by first detecting whether the context of a question is controversial or not, and, if a question is controversial, then classifying the question as one of *factual*, *method*, or *argumentative*. We perform a crowdsourcing annotation of a sample of the Yandex query log from 2012. This results in a dataset of 39,340 questions about 19 controversial topics, 28% of which are argumentative.[1] Additionally, we analyze

[1]The anonymity of the questions' askers is preserved by sampling only frequent questions from the logs. An exemplary data sample can be found here:
https://files.webis.de/data-in-production/data-research/arguana/webis-arg-questions/dataset.csv

the questions in the dataset with regard to their form and structure. The main insight drawn from the analysis is that argumentative questions are characterized mainly by asking for reasons and predictions. Based on this dataset, we build classifiers that identify argumentative, factual, and method questions. Our experiments show that a BERT-based model can classify the three question types with a reasonable macro F_1-score of 0.78, outperforming a strong baseline at identifying non-factual questions on unseen topics.

2 RELATED WORK

Bias in search engines. Starting from a pre-defined list of controversial topics, Gezici et al. [19] observed a tendency by major search engines to rank liberal content higher than conservative content. Kulshrestha et al. [22] evaluated the contribution of search system components such as source documents and ranking algorithms to political bias in search results. Yom-Tov et al. [36] showed that most users are more likely to read opinions that match their own, and that diversification of search results can only be successful if documents with opposite views are lexically similar to the user's queries. Azzopardi [3] surveyed different sources of bias in search on socio-political topics. We expect users who ask questions about controversial topics and especially those that are argumentative to be prone to bias toward one of the stances. This work sheds light on the magnitude and characteristics of argumentative questions in search engine query logs and proposes an approach to identify them.

Argument retrieval. The goal of argument retrieval is to support users in forming an unbiased opinion on a controversial topic by retrieving pro and con arguments for a given query [9, 14, 18, 24, 31, 34]. Argument retrieval builds on a decade of research in mining arguments [2, 15], classifying their stances into pro and con [5, 32], and clustering them according to the aspect they emphasize [30]. Recently, Potthast et al. [29] and Bondarenko et al. [8] developed benchmarks to assess the effectiveness of argument retrieval systems in terms of relevance and argument quality. The proposed systems assume a query to be a controversial topic, a statement, or a question. Integrating argument mining and retrieval systems in a conversational answering system or search engine requires identifying questions that look for arguments.

Opinion questions in CQA/QA. Researchers studied subjective and opinion questions mainly in the context of community question answering (CQA). The motivation for this classification was to fact-check answers to factual questions [27], suggest questions with the same intent [13, 25], or automatically answer opinion questions [4, 21]. These latter approaches classify the polarity of an opinion question into negative or positive and return answers with the matching polarity. Since this might lead to undesired bias by amplifying the view of the question, Moghaddam and Ester [28] proposed a mechanism to retrieve both negative and positive answers. Even though opinion questions seem to be close to argumentative questions, argumentative questions differ conceptually by targeting controversial topics, which are characterized by disagreement and require reasoned arguments as answers.

Query analysis on controversial topics. Gyllstrom and Moens [20] identified controversial topics in web search queries by checking whether Google auto-complete suggests positive or negative words for a given concept. Weber et al. [35] filtered queries on controversial topics and labeled them with "left" or "right" by checking whether the clicked URL for a query is a left or right political blog. Chelaru et al. [12] extracted queries from the AOL query log using templates and labeled them as positive, negative, or objective. Topics that occur in both positive and negative queries more than a specific threshold were considered to be controversial. While this work is closest in spirit to ours, our work focuses on questions-like queries which are less ambiguous with regard to their intent than short queries. Cambazoglu et al. [11] annotated a sample of 1,000 web search questions with a taxonomy of 16 question types which include *opinion* and *reason* questions. The study shows that opinion and reason questions amount to about 1% of web search questions. The low prevalence of opinion questions in search query streams is a challenge for such holistic taxonomies and hence for developing answering systems for them. In this paper, we concentrate on argumentative web search questions by sampling questions on controversial topics and classifying them into factual, argumentative, and method questions.

Table 1: Controversial topics used in the study.

Debate Portals	Russian News
Abortion	2011–2013 Russian protests
Death penalty	Alexei Navalny
Euthanasia	Anatoliy Serdyukov
Evolution	European debt crisis
Gay marriage	Floods in Krymsk
God exists	Magnitsky act
In vitro fertilization	Nordstream
Legalize marijuana	Presidential elections
	Putin
	Pussy riot trial
	Yukos

3 DATASET CONSTRUCTION

Given that no available question datasets exist on controversial topics, we conducted an annotation task to create one starting from 2 billion archived Yandex queries in Russian from 2012.[2]

Following Völske et al. [33], we first extracted queries from the Yandex log starting with a question word that resulted in 1.5 billion Russian questions. To find questions asking about controversial topics, we created a list of such topics (cf. Table 1) by: (1) Selecting eight debate topics from the args.me corpus [1] with the highest number of arguments.

(2) To cover local issues, we also selected 11 debate topics from the list of the most important events in 2012 according to the Russian RIA news agency.[3] Since question topic classification is not the focus of our study, we opted for the following simple approach. We manually expanded each topic with synonymous phrases, e.g., "gay marriage" → "same-sex marriage" (on average, five phrases for each topic). A question was then considered on a topic if its lemmas

[2]Original Russian questions were used in the annotation task as well as the next steps in the study.
[3]https://ria.ru/20121221/915705250.html

Table 2: The absolute and relative count of the questions per label in the dataset.

Label	Abs.	Rel.	Label	Abs.	Rel.
Topic Aboutness			**Question Types**		
On topic	40,689	73%	Factual	25,332	64%
Not on topic	11,665	24%	Argumentative	10,982	28%
Ill-formed	1,477	3%	Method	3,026	8%

contained all of the lemmas of one of the topic phrases. Filtering the questions using the expanded phrases for the 19 topics resulted in 4.5 million questions.

We then sampled 54,850 questions and annotated them in two subsequent labeling tasks on the crowdsourcing platform Toloka:[4] (1) *topic aboutness* to label questions with whether they are about the controversial topic (*on topic*), contain the topic's lemmas but do not ask about the topic (*not on topic*), or are not grammatically correct questions (*ill-formed*). An example of a not-on-topic question is "What is evolution of marketing?" which does not ask about Darwin's theory of evolution (our controversial topic). And (2) *question type labeling* for on-topic questions into:

Factual questions asking about information that most people agree on (facts), e.g., "Which countries legalized marijuana?"

Argumentative questions seeking arguments or opinions for or against a topic or a statement in a question—an answer would ideally contain reasoned evidence which people might accept, reject, or doubt, e.g., "Should marijuana be legalized?"

Method questions seeking a list of instructions or a description of a method to reach a goal, e.g., "How to hold a referendum on legalizing marijuana?"

These three question types differ in how widely acceptable their answers are. An answer to factual questions is a single fact that can be verified. On the other hand, a multitude of opposing and acceptable arguments exist to argumentative questions. Similarly, different lists of instructions exist for achieving a goal, which in turn differ in the required effort to follow them and their outcomes.

We conducted both annotation tasks in two steps: a pilot study to test the annotation tasks and collect gold labels and a main study. The annotation instructions for both tasks included the description of the labels and an example for each label. The annotation interface for topic aboutness included also an excerpt of the corresponding Wikipedia article that describes the topic. We split questions belonging to a topic into batches of 10 items, one of which was a quality check. We assigned three workers to each task and allocated a new worker in case one of the workers got suspended due to low annotation quality. To guarantee the quality of annotations, the tasks were conducted with a qualification test and quality checks.

Pilot Study. We randomly sampled 120 questions from the dataset on each of the 19 topics (cf. Table 1) resulting in 2,280 questions. From these questions, 25% were used as gold labels for the quality checks and the qualification tests, while the rest 75% of the questions were labeled by crowd workers. The gold labels were annotated by two experts who are native Russian speakers. The workers' inter-annotator agreement for the topic aboutness annotations was a Krippendorff's α = 0.55 and for the question type labeling α = 0.45.

Main Study. The main annotation phase covered the 52,570 questions that remained after excluding the questions used in the pilot study. Questions with perfect agreement in the pilot study were added to the quality checks and used in subsequent batches. During the annotation, we expanded the set of quality checks with annotations with perfect agreement. We used quality checks only once to ensure that workers do not memorize them.

The workers achieved an α of 0.55 on the topic aboutness task and an α of 0.49 (a slight improvement over the pilot) on question type labeling. The questions on which the crowd workers achieved majority agreement in both tasks amounted to 50,832 questions (97% of all questions). The final dataset includes these questions in addition to the pilot questions on which the crowd workers achieved majority agreement. Notice that the dataset which we use in the next steps includes 39,340 questions which are those labeled as factual, argumentative, or method in the final dataset.

Table 2 shows the distribution of the annotated questions over the topic aboutness labels and question types. The statistics show the merit of conducting the topic filtering step as 24% of the sampled questions are not on the 19 controversial topics. The majority (64%) of the questions on controversial topics in our study look for facts, while 28% of the questions look for arguments. This indicates that people use search engines more often to look for some background information about controversial topics like factual evidence. Still, the statistics demonstrate that the share of argumentative questions is substantial.

4 QUANTITATIVE QUESTION ANALYSIS

Having crowdsourced a question dataset on controversial topics, we analyze what distinguishes argumentative questions from factual and method ones. Our analysis mainly targets four characteristics of questions that we assume to set apart argumentative questions from the other question types: question words, predictions, comparisons, and personal pronouns. To capture the characteristics in a question, we develop patterns that use surface features of questions (e.g., lemmas, POS tags, and tense information), which we extract using the *mystem* tagger.[5]

Question Words. Question words are a strong indicator of the answer type for a question. Early research on question answering considered factual questions to start with wh-words [26] and mapped each wh-word to an entity type (e.g., "Time" for `when`). Compared to wh-questions which seek short answers, yes/no questions are statements which are converted into questions. In the context of controversial topics, we expect yes/no questions to be claims that the users have and would like to collect evidence for. On the other hand, we anticipate that question starting with wh-words to look for background knowledge about a controversial topic.

Personal Pronouns. We expect search engine users to refer to themselves or to an imaginary audience while formulating an argumentative question. To capture such questions, we extracted

[4] https://toloka.ai/

[5] https://yandex.ru/dev/mystem/

Table 3: The absolute and relative count of factual, argumentative, and method questions in the dataset for each characteristic in the analysis and question examples (English translations of original Russian questions).

Characteristic	Factual		Argument.		Method		All		Example	Type
	Rel.	Abs.	Rel.	Abs.	Rel.	Abs.	Rel.	Abs.		
Yes/No	7.2%	1,743	**13.8%**	**1,501**	0.4%	13	8.3%	3,262	Is marijuana legalization possible?	Arg.
Predictions	3.8%	921	**8.2%**	**892**	0.6%	19	4.8%	1,875	Will marijuana be legalized in Russia?	Arg.
Comparisons	3.2%	777	**5.7%**	**625**	4.4%	130	4.0%	1,559	Should we have partial or full marijuana legalization?	Arg.
Personal Pronouns	0.3%	83	**3.8%**	**412**	0.43%	13	1.3%	508	Do you think the president will legalize marijuana?	Arg.
Wh-words	67.0%	16,980	62.0%	6,807	**94.3%**	**2,852**	67.7%	26,639		
`why`	1.3%	325	**20.7%**	**2,253**	0.0%	0	6.6%	2,605	Why are people in favor of legalizing marijuana?	Arg.
`how`	7.0%	1,704	7.6 %	833	**87.5%**	**2,590**	13.4%	5,274	How to fill an amendment for marijuana?	Meth.
`how much/many`	**10.5%**	**2,553**	2.1%	228	0.2%	5	7.4%	2,914		
• `money`	**3.4%**	**819**	0.4%	43	0.1%	2	2.3%	907	How much does marijuana cost?	Fact.
• `people`	**2.4%**	**585**	0.7%	81	0.0%	0	1.8%	691	How many people consume marijuana?	Fact.
• `time`	**1.3%**	**322**	0.1%	14	0.0%	0	0.9%	354	How many hours can one detect marijuana in the body?	Fact.

all questions whose subject is a first-person or a second-person pronoun.

Predictions. One way of approaching a controversial topic is deliberation, where people try to argue for a possible course of action by predicting its consequences. We expect a subset of argumentative questions to ask for predictions that pertain to the controversial topic (e.g., "Will legalizing marijuana reduce crime?"). To extract prediction questions, we developed a pattern that looks up whether the first verb is `will` or whether it is in the future tense.

Comparisons. Controversial (or argumentative) topics can also be formulated as a comparison between at least two options (e.g., death penalty vs. life imprisonment). A recent study on comparative questions asked on the Web shows that more than 50% of such questions are argumentative, not factual [7]. To identify comparative questions in our dataset we apply 8 regular expressions that were proposed in [6] and were shown to classify comparative questions with a precision of 1.0.

The distribution of factual, argumentative, and method questions in the extracted questions for each characteristic are shown in Table 3. We also list examples of the extracted question for each characteristic, together with the question type most associated with it. By comparing the relative counts for factual and argumentative yes/no questions, we observe that they are almost twice more likely to be argumentative than factual. Wh-questions, on the other hand, cover almost the same proportion (two-thirds) of factual and argumentative questions and the majority of method questions. By analyzing the distributions for the single question words we notice clear associations of some of them with the question types that we report in the table. As illustrated in the table, 20% of argumentative questions start with `why`, which shows that users ask explicitly for reasons when they look for arguments. A stronger association can be seen for method questions which are dominated by `how` with 87.5%. Interestingly, about 10% of factual questions look for quantities using `how much/many`. We customized the regular expression to capture different types of quantities people ask for (`money`, `people`, and `time`) by specifying synonymous verbs or nouns to the quantity type after `how much/many`. It turns out that a third of the questions that look for quantities ask about money, while about 20% ask for the count of people.

A closer look at the question types distribution for predictions shows that 8.2% of argumentative questions are written in the future tense in comparison to 4.4% for the other question types. These numbers confirm our assumption that asking for predictions is a strong indicator of argumentative questions. Personal pronouns match almost only argumentative questions, which renders personal pronouns a strong indicator of argumentative questions. Still, the very low percentage of matched argumentative questions (3.8%) shows that users formulate argumentative questions more objectively. Comparisons fall short at distinguishing argumentative questions since the relative count of method questions is quite close to that of argumentative (5.7% vs 4.4%).

5 EXPERIMENTS

In this section, we assess the effectiveness of automatic classifiers that map the questions in our dataset to their question types (factual, method, or argumentative). Since new controversial topics emerge all the time, a key challenge lies in generalizing beyond the 19 topics contained in the dataset.

To assess this, we conduct in-topic and cross-topic experiments on the dataset where we control for the topic differently. We use only questions that are labeled on topic in our dataset in both experiments. The in-topic experiments are conducted in a 5-fold cross-validation fashion. While sampling the folds, the dataset questions are stratified by their topics, making each fold equally cover all the topics. The cross-topic experiments, on the other hand, are conducted in a leave-one-out cross-validation fashion. Here, we use all the questions on one topic as a test set while taking the remaining questions as a training set. As evaluation metrics, we use F_1-score for each of the three question types and their macro average.

Our classifier is based on RuBERT [23] which is a BERT [16] model trained on the Russian Wikipedia and news articles. We feed the question to RuBERT as `[CLS]question[SEP]` and fine-tune it for two epochs with a learning rate of 2×10^{-5}.

In addition, we use four baselines: random baseline, majority baseline, a rule-based classifier, and logistic regression (LR). The

Table 4: F_1-score for classifying questions on controversial topics into factual, method, and argumentative; in-topic and cross-topic settings.

	In-topic				Cross-topic			
Classifier	**Fact.**	**Method**	**Arg.**	**Macro**	**Fact.**	**Method**	**Arg.**	**Macro**
Random	0.44	0.13	0.31	0.29	0.43	0.12	0.30	0.28
Majority	0.78	0.00	0.00	0.26	0.78	0.00	0.00	0.26
Rule-based	0.71	0.62	0.48	0.60	0.69	0.56	0.46	0.57
LR	0.86	0.70	0.67	0.74	0.80	0.52	0.61	0.65
RuBERT	0.90	0.83	0.78	0.84	0.85	0.74	0.74	0.78

rule-based classifier relies on the insights gained from the analysis in Section 4, which shows a strong association between the wh-words and the three question types. The rule-based classifier categorizes a question into factual if it starts with one of the wh-words, except for `how` and `why` for which the classifier predicts the question types method and argumentative, respectively. In case the question starts with any other word, it is classified as argumentative. The logistic regression classifier takes the count of 1-3-grams and the count of part-of-speech 1-3-grams in the question as features.

Table 4 shows the classification results in the in-topic and cross-topic experiments. The rule-based classifier reaches a comparable macro F_1-score of 0.57 in both experiments, showing that question words are a strong indicator of the question type regardless of the topic. RuBERT is more robust across topics than logistic regression and suffers only a drop of 0.06 macro F_1-score between the two experiments in comparison to 0.09 for logistic regression. Whilst RuBERT and logistic regression perform very well on factual questions, RuBERT performs substantially better on non-factual questions.

5.1 Error Analysis

The results of the experiments show promising results in classifying questions on controversial topics into factual, method, and argumentative. Still, the effectiveness of RuBERT in the cross-topic setting (F_1-score of 0.78) indicates a large potential to improve the classifier. To this end, we conduct an error analysis which aims at detecting systematic errors that provide insights into how to improve the approach. In the error analysis, we manually check questions in the test sets of the cross-topic experiments for which RuBERT predicts the wrong question type.

Overall, we find that the most confused question types are factual and argumentative, with 2,995 factual questions classified as argumentative and 2,683 argumentative questions classified as factual. We notice that the cause of some errors is keywords or the question tense which are correlated with factual or argumentative questions. Table 5 shows examples of these errors. Some keywords are often used in factual questions in the dataset (e.g., "allowed" or "approve"). RuBERT seems to rely extensively on such keywords, causing argumentative questions that use them to be classified as factual (e.g., Question 1 in Table 5). A similar case can be observed for questions in the past tense, which is more used in factual questions. Because of this, RuBERT tends to classify method questions in the past tense as factual (e.g., Question 2 in Table 5). The analysis shows that RuBERT tends to rely on surface features to predict the question type. This can be explained by the scarce context provided in the question and hints at the need to expand the question with more information about the topic.

Table 5: Examples of questions in the test datasets of the cross-topic experiments which RuBERT classified wrongly.

Question	**Label**	**Pred.**
Should gays be allowed to marry?	Arg.	Fact.
How was death penalty done in the USSR?	Meth.	Fact.

6 CONCLUSION

We suggest that future search engines should support users in forming unbiased opinions on controversial topics. To foster the related research, we annotated a questions dataset that is sampled from the Yandex query log and that covers 19 controversial topics. Each of the questions is labeled as to whether it relates to one of the 19 controversial topics, and if so, whether it is looking for a fact, a method, or arguments. The crowdsourcing study shows that the percentage of argumentative questions is high (28%), which underlines the importance of developing customized answering systems for this question type. A comparative analysis of argumentative questions with the other question types gives first insights into their structure and properties: argumentative questions tend to ask for reasons and predictions. Experiments with the new dataset show high effectiveness (F_1-score of 0.78) in automatically classifying questions into argumentative, factual, or method, even on unseen topics. The promising classification performance opens a way to properly handle questions on controversial topics by the search engines and to adapt argument retrieval systems to answer those questions that are argumentative. A direct improvement of the question classification approach can be achieved by incorporating the retrieved documents for a question. Such a classification approach can extract arguments from the retrieved documents or learn from simpler features (e.g., document structure or meta-information). The next research steps include extending our work by applying advanced question classification methods, and analyzing argumentative questions in search engine logs using our method on a larger scale.

ACKNOWLEDGMENTS

This work was partially funded by the German Federal Ministry of Education and Research within the project Competence Center for Scalable Data Services and Solutions (ScaDS) Dresden/Leipzig (BMBF 01IS18026B). Alexander Bondarenko is supported by the DFG (projects "ACQuA" and "ACQuA 2.0": Answering Comparative Questions with Arguments; grants HA 5851/2-1 and HA 5851/2-2) as part of the priority program "RATIO: Robust Argumentation Machines" (SPP 1999). Pavel Braslavski acknowledges funding from the Ministry of Science and Higher Education of the Russian Federation (project 075-02-2022-877). We are especially grateful to Yandex for granting access to the data.

REFERENCES

[1] Yamen Ajjour, Henning Wachsmuth, Johannes Kiesel, Martin Potthast, Matthias Hagen, and Benno Stein. 2019. Data Acquisition for Argument Search: The args.me corpus. In *Proceedings of the 42nd German Conference on Artificial Intelligence (KI 2019)*, Christoph Benzmüller and Heiner Stuckenschmidt (Eds.). Springer, 48–59. https://doi.org/10.1007/978-3-030-30179-8_4

[2] Khalid Al-Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. 2016. A News OPTeditorial Corpus for Mining Argumentation Strategies. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING 2016)*, Yuji Matsumoto and Rashmi Prasad (Eds.). ACL, 3433–3443. http://aclweb.org/anthology/C16-1324

[3] Leif Azzopardi. 2021. Cognitive Biases in Search: A Review and Reflection of Cognitive Biases in Information Retrieval. In *Proceedings of the Conference on Human Information Interaction and Retrieval (CHIIR 2021)*, Falk Scholer, Paul Thomas, David Elsweiler, Hideo Joho, Noriko Kando, and Catherine Smith (Eds.). ACM, 27–37. https://doi.org/10.1145/3406522.3446023

[4] Alexandra Balahur, Ester Boldrini, Andrés Montoyo, and Patricio Martínez-Barco. 2009. A Comparative Study of Open Domain and Opinion Question Answering Systems for Factual and Opinionated Queries. In *Proceedings of the Recent Advances in Natural Language Processing (RANLP 2009)*, Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, Nicolas Nicolov, and Nikolai Nikolov (Eds.). ACL, 18–22. https://aclanthology.org/R09-1004/

[5] Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. 2017. Stance Classification of Context-Dependent Claims. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*, Mirella Lapata, Phil Blunsom, and Alexander Koller (Eds.). ACL, 251–261. https://doi.org/10.18653/v1/e17-1024

[6] Alexander Bondarenko, Yamen Ajjour, Valentin Dittmar, Niklas Homann, Pavel Braslavski, and Matthias Hagen. 2022. Towards Understanding and Answering Comparative Questions. In *Proceedings of the 15th ACM International Conference on Web Search and Data Mining (WSDM 2022)*, K. Selcuk Candan, Huan Liu, Leman Akoglu, Xin Luna Dong, and Jiliang Tang (Eds.). ACM, 66–74. https://doi.org/10.1145/3488560.3498534

[7] Alexander Bondarenko, Pavel Braslavski, Michael Völske, Rami Aly, Maik Fröbe, Alexander Panchenko, Chris Biemann, Benno Stein, and Matthias Hagen. 2020. Comparative Web Search Questions. In *Proceedings of the 13th ACM International Conference on Web Search and Data Mining (WSDM 2020)*, James Caverlee, Xia (Ben) Hu, Mounia Lalmas, and Wei Wang (Eds.). ACM, 52–60. https://dl.acm.org/doi/abs/10.1145/3336191.3371848

[8] Alexander Bondarenko, Maik Fröbe, Meriem Beloucif, Lukas Gienapp, Yamen Ajjour, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. 2020. Overview of Touché 2020: Argument Retrieval. In *Working Notes Papers of the CLEF 2020 Evaluation Labs (CEUR Workshop Proceedings, Vol. 2696)*, Linda Cappellato, Carsten Eickhoff, Nicola Ferro, and Aurélie Névéol (Eds.). CEUR-WS.org, 22 pages. http://ceur-ws.org/Vol-2696/

[9] Alexander Bondarenko, Lukas Gienapp, Maik Fröbe, Meriem Beloucif, Yamen Ajjour, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. 2021. Overview of Touché 2021: Argument Retrieval. In *Proceedings of the 12th International Conference of the CLEF Association (CLEF 2021) (Lecture Notes in Computer Science, Vol. 12880)*, K. Selçuk Candan, Bogdan Ionescu, Lorraine Goeuriot, Henning Müller, Alexis Joly, Maria Maistro, Florina Piroi, Guglielmo Faggioli, and Nicola Ferro (Eds.). Springer, 450–467. https://doi.org/10.1007/978-3-030-85251-1_28

[10] Berkant Barla Cambazoglu, Valeria Bolotova-Baranova, Falk Scholer, Mark Sanderson, Leila Tavakoli, and W. Bruce Croft. 2021. Quantifying Human-Perceived Answer Utility in Non-factoid Question Answering. In *Proceedings of the Conference on Human Information Interaction and Retrieval (CHIIR 2021)*, Falk Scholer, Paul Thomas, David Elsweiler, Hideo Joho, Noriko Kando, and Catherine Smith (Eds.). ACM, 75–84. https://doi.org/10.1145/3406522.3446028

[11] Berkant Barla Cambazoglu, Leila Tavakoli, Falk Scholer, Mark Sanderson, and W. Bruce Croft. 2021. An Intent Taxonomy for Questions Asked in Web Search. In *Proceedings of the Conference on Human Information Interaction and Retrieval (CHIIR 2021)*, Falk Scholer, Paul Thomas, David Elsweiler, Hideo Joho, Noriko Kando, and Catherine Smith (Eds.). ACM, 85–94. https://doi.org/10.1145/3406522.3446027

[12] Sergiu Chelaru, Ismail Sengör Altingövde, Stefan Siersdorfer, and Wolfgang Nejdl. 2013. Analyzing, Detecting, and Exploiting Sentiment in Web Queries. *ACM Trans. Web* 8, 1 (2013), 6:1–6:28. https://doi.org/10.1145/2535525

[13] Long Chen, Dell Zhang, and Mark Levene. 2012. Understanding User Intent in Community Question Answering. In *Proceedings of the 21st World Wide Web Conference (WWW 2012)*, Alain Mille, Fabien Gandon, Jacques Misselis, Michael Rabinovich, and Steffen Staab (Eds.). ACM, 823–828. https://doi.org/10.1145/2187980.2188206

[14] Sihao Chen, Daniel Khashabi, Chris Callison-Burch, and Dan Roth. 2019. PerspectroScope: A Window to the World of Diverse Perspectives. In *Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL 2019)*, Marta R. Costa-jussà and Enrique Alfonseca (Eds.). ACL, 129–134. https://doi.org/10.18653/v1/p19-3022

[15] Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. 2017. What is the Essence of a Claim? Cross-Domain Claim Identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, Martha Palmer, Rebecca Hwa, and Sebastian Riedel (Eds.). ACL, 2055–2066. https://doi.org/10.18653/v1/d17-1218

[16] 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 ACL: Human Language Technologies (NAACL-HLT 2019)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). ACL, 4171–4186. https://doi.org/10.18653/v1/n19-1423

[17] Lorik Dumani, Patrick J. Neumann, and Ralf Schenkel. 2020. A Framework for Argument Retrieval - Ranking Argument Clusters by Frequency and Specificity. In *Proceedings of the 42nd European Conference on IR Research (ECIR 2020) (Lecture Notes in Computer Science, Vol. 12035)*, Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins (Eds.). Springer, 431–445. https://doi.org/10.1007/978-3-030-45439-5_29

[18] Lorik Dumani and Ralf Schenkel. 2019. A Systematic Comparison of Methods for Finding Good Premises for Claims. In *Proceedings of the 42nd International Conference on Research and Development in Information Retrieval (SIGIR 2019)*, Benjamin Piwowarski, Max Chevalier, Éric Gaussier, Yoelle Maarek, Jian-Yun Nie, and Falk Scholer (Eds.). ACM, 957–960. https://doi.org/10.1145/3331184.3331282

[19] Gizem Gezici, Aldo Lipani, Yücel Saygin, and Emine Yilmaz. 2021. Evaluation Metrics for Measuring Bias in Search Engine Results. *Inf. Retr. J.* 24, 2 (2021), 85–113. https://doi.org/10.1007/s10791-020-09386-w

[20] Karl Gyllstrom and Marie-Francine Moens. 2011. Clash of the Typings - Finding Controversies and Children's Topics Within Queries. In *Proceedings of the 33rd European Conference on IR Research (ECIR 2011) (Lecture Notes in Computer Science, Vol. 6611)*, Paul D. Clough, Colum Foley, Cathal Gurrin, Gareth J. F. Jones, Wessel Kraaij, Hyowon Lee, and Vanessa Murdock (Eds.). Springer, 80–91. https://doi.org/10.1007/978-3-642-20161-5_10

[21] Lun-Wei Ku, Yu-Ting Liang, and Hsin-Hsi Chen. 2008. Question Analysis and Answer Passage Retrieval for Opinion Question Answering Systems. *Int. J. Comput. Linguistics Chin. Lang. Process.* 13, 3 (2008), 307–326. http://www.aclclp.org.tw/clclp/v13n3/v13n3a3.pdf

[22] 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. *Inf. Retr. J.* 22, 1-2 (2019), 188–227. https://doi.org/10.1007/s10791-018-9341-2

[23] Yuri Kuratov and Mikhail Y. Arkhipov. 2019. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. *CoRR* abs/1905.07213 (2019). arXiv:1905.07213 http://arxiv.org/abs/1905.07213

[24] Ran Levy, Ben Bogin, Shai Gretz, Ranit Aharonov, and Noam Slonim. 2018. Towards an Argumentative Content Search Engine using Weak Supervision. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, Emily M. Bender, Leon Derczynski, and Pierre Isabelle (Eds.). ACL, 2066–2081. https://aclanthology.org/C18-1176/

[25] Baoli Li, Yandong Liu, Ashwin Ram, Ernest V. Garcia, and Eugene Agichtein. 2008. Exploring Question Subjectivity Prediction in Community QA. In *Proceedings of the 31st Annual International Conference on Research and Development in Information Retrieval (SIGIR 2008)*, Sung-Hyon Myaeng, Douglas W. Oard, Fabrizio Sebastiani, Tat-Seng Chua, and Mun-Kew Leong (Eds.). ACM, 735–736. https://doi.org/10.1145/1390334.1390477

[26] Xin Li and Dan Roth. 2006. Learning Question Classifiers: The Role of Semantic Information. *Nat. Lang. Eng.* 12, 3 (2006), 229–249. https://doi.org/10.1017/S1351324905003955

[27] Tsvetomila Mihaylova, Georgi Karadzhov, Pepa Atanasova, Ramy Baly, Mitra Mohtarami, and Preslav Nakov. 2019. SemEval-2019 Task 8: Fact Checking in Community Question Answering Forums. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval@NAACL-HLT 2019)*, Jonathan May, Ekaterina Shutova, Aurélie Herbelot, Xiaodan Zhu, Marianna Apidianaki, and Saif M. Mohammad (Eds.). ACL, 860–869. https://doi.org/10.18653/v1/s19-2149

[28] Samaneh Moghaddam and Martin Ester. 2011. AQA: Aspect-based Opinion Question Answering. In *Proceedings of the 11th International Conference on Data Mining Workshops (ICDMW 2011)*, Myra Spiliopoulou, Haixun Wang, Diane J. Cook, Jian Pei, Wei Wang, Osmar R. Zaïane, and Xindong Wu (Eds.). IEEE Computer Society, 89–96. https://doi.org/10.1109/ICDMW.2011.34

[29] Martin Potthast, Lukas Gienapp, Florian Euchner, Nick Heilenkötter, Nico Weidmann, Henning Wachsmuth, Benno Stein, and Matthias Hagen. 2019. Argument Search: Assessing Argument Relevance. In *Proceedings of the 42nd International Conference on Research and Development in Information Retrieval (SIGIR 2019)*, Benjamin Piwowarski, Max Chevalier, Éric Gaussier, Yoelle Maarek, Jian-Yun Nie, and Falk Scholer (Eds.). ACM, 1117–1120. https://doi.org/10.1145/3331184.3331327

[30] 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 Conference of the Association for Computational Linguistics (ACL 2019)*, Anna Korhonen, David R.

Traum, and Lluís Màrquez (Eds.). ACL, 567–578. https://doi.org/10.18653/v1/p19-1054

[31] Christian Stab, Johannes Daxenberger, Chris Stahlhut, Tristan Miller, Benjamin Schiller, Christopher Tauchmann, Steffen Eger, and Iryna Gurevych. 2018. ArgumenText: Searching for Arguments in Heterogeneous Sources. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2018)*, Yang Liu, Tim Paek, and Manasi S. Patwardhan (Eds.). ACL, 21–25. https://doi.org/10.18653/v1/n18-5005

[32] Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018. Cross-topic Argument Mining from Heterogeneous Sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). ACL, 3664–3674. https://doi.org/10.18653/v1/d18-1402

[33] Michael Völske, Pavel Braslavski, Matthias Hagen, Galina Lezina, and Benno Stein. 2015. What Users Ask a Search Engine: Analyzing One Billion Russian Question Queries. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management (CIKM 2015)*, James Bailey, Alistair Moffat, Charu C. Aggarwal, Maarten de Rijke, Ravi Kumar, Vanessa Murdock, Timos K. Sellis, and Jeffrey Xu Yu (Eds.). ACM, 1571–1580.

[34] Henning Wachsmuth, Martin Potthast, Khalid Al Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. 2017. Building an Argument Search Engine for the Web. In *Proceedings of the 4th Workshop on Argument Mining (ArgMining@EMNLP 2017)*, Ivan Habernal, Iryna Gurevych, Kevin D. Ashley, Claire Cardie, Nancy L. Green, Diane J. Litman, Georgios Petasis, Chris Reed, Noam Slonim, and Vern R. Walker (Eds.). ACL, 49–59. https://doi.org/10.18653/v1/w17-5106

[35] Ingmar Weber, Venkata Rama Kiran Garimella, and Erik Borra. 2012. Mining Web Query Logs to Analyze Political Issues. In *Proceedings of the 2012 Web Science Conference (WebSci 2012)*, Noshir S. Contractor, Brian Uzzi, Michael W. Macy, and Wolfgang Nejdl (Eds.). ACM, 330–334. https://doi.org/10.1145/2380718.2380761

[36] Elad Yom-Tov, Susan Dumais, and Qi Guo. 2014. Promoting Civil Discourse through Search Engine Diversity. *Social Science Computer Review* 32, 2 (2014), 145–154.