Analyzing, Detecting and Exploiting Sentiment in Web Queries

SERGIU CHELARU, L3S Research Center, Hannover, Germany
ISMAIL SENGOR ALTINGOVDE, Middle East Technical University, Ankara, Turkey
STEFAN SIERSDORFER, L3S Research Center, Hannover, Germany
WOLFGANG NEJDL, L3S Research Center, Hannover, Germany

The Web contains an increasing amount of biased and opinionated documents on politics, products and polarizing events. In this paper, we present an in-depth analysis of Web search queries for controversial topics, focusing on query sentiment. To this end, we conduct extensive user assessments and discriminative term analyses, as well as a sentiment analysis using the SentiWordNet thesaurus, a lexical resource containing sentiment annotations. Furthermore, in order to detect the sentiment expressed in queries, we build different classifiers based on query texts, query result titles, and snippets. We demonstrate the virtue of query sentiment detection in two different use cases. First, we define a query recommendation scenario that employs sentiment detection of results to recommend additional queries for polarized queries issued by search engine users. The second application scenario is controversial topic discovery, where query sentiment classifiers are employed to discover previously unknown topics that trigger both highly positive and negative opinions among the users of a search engine. For both use cases, the results of our evaluations on real-world data are promising, and show the viability and potential of query sentiment analysis in practical scenarios.

Categories and Subject Descriptors: H.4 [**Information Systems Applications**]: *Miscellaneous*

General Terms: Algorithms, Experimentation

Additional Key Words and Phrases: Opinionated queries, sentiment analysis, Web search

1. INTRODUCTION

In recent years we have witnessed an increasing amount of opinionated text appearing on the Web, with people discussing ideas and political issues, criticizing movies, reviewing books, or elaborating on features of their newly-bought camera. Not surprisingly, this content is not only appreciated by ordinary end users but also by professionals ranging from marketing and advertisement specialists to political strategists. The growing interest and demand for automatic analysis techniques and tools for opin-

A preliminary version of this paper appeared as a poster in the Proceedings of the 34th European Conference on IR Research (ECIR 2012) [Chelaru et al. 2012].

DOI 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

ionated digital texts have also fueled research and led to various approaches in opinion mining and sentiment analysis [Pang and Lee 2008].

While there exists a considerable body of literature on mining opinions from product reviews, blogs, Web search results, news articles and microblogs [Pang and Lee 2008; Pan et al. 2010; Thomas et al. 2006], another rich source of information, namely, Web search queries, has largely been overlooked (an exception being the recent work [Gyllstrom and Moens 2011], which we discuss in Section 8). We anticipate that a non-trivial amount of Web queries that explicitly reflect opinions is issued to search engines, especially on controversial/popular topics in the society. For instance, when searching for the topic "abortion" using a major search engine, we are suggested not only a number of neutral queries such as "abortion facts" or "abortion statistics" but also queries that are in support of or against abortion (e.g., "abortion is right" vs. "abortion is morally wrong"). As these suggestions are usually based on real (and frequent) queries by other users, this provides clear evidence that opinionated queries are not exceptional on the Web[1]. However, to the best of our knowledge, no previous work has attempted to characterize opinionated queries or detect sentiment in queries.

In this paper, our goal is to analyze and exploit Web queries as a new, rich and mostly unexplored source of user-generated content that can convey community views and opinions on a multitude of topics. We motivate the need for detecting the sentiment in Web queries through several use case scenarios:

Query recommendation. While search engines can already generate effective recommendations for popular queries by exploiting the abundance of information captured in query logs (from past interactions of the users with the search system), generating useful recommendations for the queries in the long tail (accounting for more than 50% of the query volume [Broccolo et al. 2012]) is more challenging and poorly addressed by the current techniques. Especially for such queries, the sentiment reflected in a query can be used as a novel and additional cue for query recommendation. Two potential use cases for a query sentiment classifier in this scenario can be: (i) suggesting queries that are aligned with the user's opinion in the original query (to improve relevance of recommendations), and (ii) suggesting queries in the directions other than the user's opinion (to improve diversity of recommendations, as in [Song et al. 2011]).

Query result aggregation. This can be considered as a follow-up scenario for the above use case in the sense that once the system decides on the queries that have similar (or, opposite) opinion to the initial query, an aggregated result from such queries can be constructed, to provide an overview of the required viewpoint on the issue. For instance, for the query "human cloning is good", it is possible to construct an aggregate result using the results of like-minded queries, such as "human cloning is right", "human cloning is necessary", or present alternative directions reflected in queries like "human cloning is ethically wrong". We emphasize that for both query recommendation and result aggregation use cases, we essentially target *opinionated* queries (possi-

[1]Note that an opinionated query may not necessarily express the personal view of the user who submitted it.

bly on some controversial issues), rather than considering all possible types of queries submitted to a search engine.

Trend analysis. Many recent studies attempt to infer real world trends by mining opinionated texts, such as blogs, Twitter messages, etc.(see e.g. [O'Connor et al. 2010]). We anticipate that our sentiment detectors can be employed to detect top-N opinionated queries submitted to a search engine within different time periods, which might be processed afterwards to detect emerging controversial topics/trends among search engine users.

Targeted advertising. Sentiment analysis of queries can be further employed to improve targeted ads displayed for opinionated queries. For instance, assume that the user submits the query "Product X sucks". Although the query includes the product name as a keyword, it seems logical not to display this particular product's ad but maybe a competitor's ad in the result page, instead. Note that such issues might have non-negligible implications on the profits of a search engine, as also discussed in [Broder et al. 2007].

Our key contributions in this work are as follows:

— This work is the first to provide a detailed *analysis of sentiment in Web queries* on controversial topics. To this end, we employ a number of different query templates on the query suggestion service of a major search engine as well as a publicly available query log to obtain a large and representative sample of real user queries. Using this dataset, we conduct manual and lexicon-based analyses of sentiments in the queries, and provide answers to various research questions: To what extent can Web queries include opinions (this may or may not reflect the query issuer's own opinion)? To what extent is sentiment in the queries mirrored in retrieved results and user clicks? Is sentiment in the queries correlated with the geographical locations of users?
— Secondly, we study the applicability of state-of-the-art sentiment analysis methods (including both lexicon-based and machine learning based methods) for *detecting the sentiment of the queries*. Query texts exhibit inherently different characteristics in comparison to classical corpora used for sentiment analysis (i.e., news stories, blogs, product reviews, comments, and even tweets). In this work, we use features obtained from the top-ranked result titles and snippets, as well as the pure query text, while applying and evaluating the current sentiment detection techniques for this new source of data with its unique characteristics. The performance is evaluated on more than 7,651 human annotated queries for 50 controversial topics.
— As a final contribution of this paper, we employ our query sentiment detectors in two of the scenarios discussed above, namely, *query recommendation* and *controversial topic discovery* (for trend analysis). In extensive user studies including both in-house participants and workers from a crowdsourcing platform we show the viability of sentiment detection for both applications.

Outline. The rest of this paper is organized as follows. In Section 2 we describe our methodology for collecting queries on controversial topics. Section 3 provides an in-depth characterization of opinionated queries based on manual and automated anno-

tations. We analyze the performance of query sentiment classification in Section 4. In Sections 5 and 6 we present and evaluate the application of query sentiment detectors in two practical scenarios, namely, query recommendation and controversial topic discovery. In Section 7 we discuss the importance and implications of our analyses and application scenarios for real world search systems as well as other fields of research like social and political sciences. Section 8 describes related work in the fields of opinion mining, query classification, and query recommendation. Finally, we conclude and point to future research directions in Section 9.

2. DATA COLLECTION

In order to obtain opinionated queries on controversial topics, we gathered a set of 50 such topics from three different resources. First, we used all of the 14 controversial topics that were employed in [Demartini and Siersdorfer 2010] in the (different yet related) context of investigating the sentiment in the *search results* retrieved by different search engines. In addition, we sampled 36 topics from the Web sites `http://www.procon.org/` and `http://en.wikipedia.org/wiki/Wikipedia:List_of_controversial_issues`. The former site includes the list of controversial topics that caused a large number of edit conflicts in the corresponding Wikipedia articles, and has been employed in other studies as well (e.g., see [Gyllstrom and Moens 2011]). The latter is a non-profit and popular Web site that is referred to by various educational institutions as an online resource. From these two Web sites, we discarded topic names that are unlikely to be issued as a keyword query (e.g. the topic name "prescription drug ads to consumers" from `procon.org`). For all selected topics, we used the topic name (as-is) as initial query. The resulting list of topics is shown in Table I.

Ideally, we would consider all queries submitted to a search engine on a particular topic, in order to determine the fraction of opinionated queries and analyze sentiment in them. As query logs are precious assets of search engines and usually kept confidential, instead, we opted for exploiting publicly available resources to sample an adequate number of queries. To this end, for each topic, we gathered a set of queries using a major search engine's query suggestion (auto-completion) service and the AOL query log [Pass et al. 2006].

For obtaining queries from the auto-completion service, we created 5 different templates, as listed in Table II. With the first, most general, template we collected all query auto-completions as ⟨*topic*⟩ followed by any letter in the English alphabet (e.g., when typing the query "abortion a", suggestions are "abortion articles", "abortion arguments", etc.). These instant suggestions are usually constructed from the popular and related queries submitted by other users [Shokouhi and Radinsky 2012; Bar-Yossef and Kraus 2011].[2] The Google web search help page[3], for instance, states that these

[2]If the prefix terms in a query do not fully match to any of the queries in the suggestion database, some search engines still auto-complete only the last term being typed. We did not collect this type of partial suggestions. This guarantees that our dataset includes only *full* queries that were actually issued by users.
[3]`http://support.google.com/websearch/answer/106230?hl=en`

Table I: List of controversial topics (along with the number of manually annotated queries per topic)

Topic	Queries	Topic	Queries	Topic	Queries
abortion	189	euthanasia	188	john mccain	124
anorexia	185	fidel castro	117	judaism	179
barack obama	129	gaddafi	116	marijuana	189
bill clinton	172	gay marriage	129	marriage	185
bit torrent	79	genetic engineering	118	nato	179
britney spears	186	george bush	188	nuclear energy	127
bullfighting	114	global warming	123	nuclear power	126
christianity	187	gypsies	157	obesity	185
circumcision	107	hamas	120	patriotism	130
climate change	129	hillary clinton	183	prostitution	184
cloning	186	hippies	118	ronald reagan	119
communism	187	homosexuality	186	sarah palin	123
cyprus	183	hugo chavez	105	scientology	187
death penalty	187	human cloning	117	stem cell research	120
drinking age	124	immigration	188	terrorism	188
economy	189	iphone	128	vegetarianism	119
employment	181	islam	192		

auto-completions are "a reflection of the search activity of all web users and the content of web pages indexed by Google."

Our second template just prefixes the topic name with one of the six interrogative pronouns (who, where, what, when, why, how) and an appropriate auxiliary verb (e.g., typing "why is abortion" returns suggestions like "why is abortion bad/good/legal", etc.). The remaining three templates use auxiliary verbs "is/are" (with negations) and, as also discussed in [Gyllstrom and Moens 2011], are more biased towards opinionated queries. However, via template 5, we again obtain all queries that are in the form of "⟨*topic*⟩ is [*letter*]", yielding a more general set than the one in [Gyllstrom and Moens 2011]. Finally, we also selected all queries from the AOL query log containing the topic name. The AOL log includes around 20M queries submitted to the AOL search engine in 2006. Since almost half of the topics match very few or no queries in this log (due to limited size and older date of this log), we only gathered queries for 26 of the topics. This case is denoted as template 6 in Table II. Note that, we intentionally chose generic templates (rather than introducing additional bias in templates to find a larger number of opinionated queries). We did this in order to get a better idea of the role of sentiment in real-world query streams. The overall process yielded a total of 31,053 queries for our 50 topics. For each of these queries, the top-10 query result titles, URLs, and snippets were gathered using the Yahoo! Search API (in June 2011).

While we attempted to use as much of this dataset as possible in our studies, for some analyses or experiments the amount of annotated data and number of redundant annotations we could gather varied due to the relatively high costs and varying avail-

Table II: Templates for gathering queries (along with the number of manually annotated queries per template): queries for templates 1-5 are obtained using the query suggestion service, and those for template 6 are extracted from the AOL log.

Id	Template	Queries
1	`<topic> [letter]`	2,664
2	`{what, why, how, where, who, when} (is|are) <topic>`	300
3	`<topic> (is|are) [blank]`	345
4	`<topic> (is|are) not [blank]`	349
5	`<topic> (is|are) [letter]`	2,458
6	`[letter]* <topic> [letter]*`	1,535

ability of human judges. However, we describe setup and number of judges involved in each individual experiment, and we made sure that data samples were chosen uniformly at random (unless specified otherwise).

3. CHARACTERISTICS OF OPINIONATED QUERIES

3.1. Sentiment in Web Search Queries

We first investigated the sentiment expressed in queries by conducting an annotation study.

Setup. We randomly sampled sets of queries proportional to the total number of queries in the templates. For each topic, we asked users to annotate around 130 queries obtained from search engine suggestions, and an additional set of 60 queries from the AOL log (if available). The annotator pool included 14 undergraduates, PhD students, and Post-docs in the area of computer science from the authors' current and former affiliations, in addition to the authors themselves. Each user was assigned at most three different topics, and asked to label a given query as positive, negative or neutral, depending on the opinion expressed in the query text. Furthermore, the queries for five randomly chosen topics (corresponding to 929 queries) were separately annotated by two of the authors in order to get an idea of the inter-user agreement. Fleiss' Kappa [Gwet 2010] was found to be 0.7[4]. Note that according to Fleiss' definition, $\kappa < 0$ corresponds to no agreement, $\kappa = 0$ to agreement by chance, and $0 < \kappa \leq 1$ to agreement beyond chance. Due to the high inter-agreement and the large number of queries to be annotated, queries for each topic were assigned to only one annotator.

Results. Overall we obtained a set of 7,651 annotated queries (see Tables I and II for the breakdown of annotated queries per topic and template, respectively), with 890 and 1,490 of them annotated as positive and negative, respectively, and the rest as objective. Fig. 1 shows the distribution of queries from each template to one of the three sentiment classes. As might be expected, queries from the first (most general) and last (AOL log) templates represent a random sample for a given topic, and are

[4]As we have only two annotators in this case, we also computed Cohen's Kappa, which is found to be 0.71.

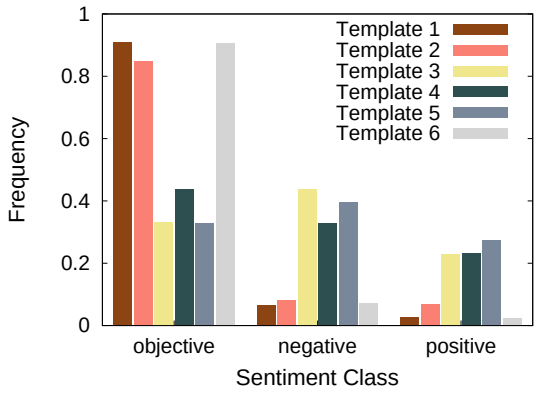


Fig. 1: Distribution of queries over the sentiment classes for different templates.

Table III: Queries and sentiment categories for the topic "George Bush"

Objective	Positive	Negative
mr george bush birthday	george bush is smart	george bush is the worst president
george bush is from texas	george bush is my hero	george bush is a lizard
george bush oil	george bush is awesome	george bush is incompetent
pre iraqi war speech george bush	george bush is not that bad	george bush doesn't care about black people

Table IV: Top-20 (stemmed) query terms w.r.t. MI values for objective vs. subjective category (left) and positive vs. negative category (right)

Terms for Objective Queries		Terms for Subjective Queries		Terms for Positive Queries		Terms for Negative Queries	
com	issu	bad	moral	good	life	bad	child
new	use	good	gay	right	awesom	wrong	wors
vs	state	wrong	crime	legal	posit	kill	problem
www	doe	right	idiot	import	best	evil	sin
countri	pictur	kill	racist	great	hero	gay	failur
lyric	question	evil	diseas	better	pro	danger	dumb
statist	japan	stupid	uneth	cool	funni	dead	retard
fact	journal	problem	great	moral	world	worst	old
2011	yahoo	hero	safe	healthi	futur	uneth	stupid
histori	york	danger	worst	hot	religion	racist	hitler

dominated by objective queries (around 90%). Still, there remain a non-negligible 10% of opinionated queries. The queries in the form of questions exhibit a slightly larger fraction of polarization, with the percentage of objective queries dropping to 85%. The templates of the form "⟨*topic*⟩ is/are ..." naturally reveal the highest subjectivity, with 67% of queries in one of these forms involving an opinion expressed in the query text. Table III shows example queries for each sentiment class for the topic "George Bush".

Term Analysis. We also conducted a term analysis on the query texts to compare the objective vs. subjective and positive vs. negative classes. For each case, we ranked the query terms (after stemming) using the Mutual Information (MI) measure [Manning et al. 2008], which essentially quantifies how much the joint distribution of terms deviates from a hypothetical distribution in which terms and sentiment classes are independent of each other. In the literature, Pointwise Mutual Information is also employed in an unsupervised method for detecting sentimental words [Turney 2002; Turney and Littman 2002]; however in this work we use MI only for identifying the most distinctive terms of the queries for each sentiment class. Table IV shows the top-20 (stemmed) terms with highest MI scores for the objective vs. subjective (left) and positive vs. negative classes (right). Note that term lists are quite intuitive in that the objective class involves mainly general query terms (e.g. *fact*, *question*, *journal*) whereas most of the subjective terms express some opinion (e.g. *racist*, *worst*, *stupid*). A clear distinction between positive and negative query terms can also be observed in Table IV.

Recall that in the user study, a considerably larger number of queries was labeled as negative rather than positive (i.e., 1,490 vs. 890). This is also reflected by the terms shown in the second column of Table IV, as most of the subjective terms seem to convey a negative feeling. Interestingly, our recent work on user comment analysis reports that users in social communities tend to cast more positive than negative votes (see [Siersdorfer et al. 2010]). A possible reason for this (rather contradictory) finding might be that, in case of the Web search activity (which is an individual act rather than a social one), users express more negative feelings, maybe for the purposes of finding like-minded people complaining or providing solutions for the same issue.

3.2. Analysis of Query Volumes

We also conducted an analysis of the volume of queries containing sentiment; in particular we studied the frequency of queries from the different sentiment classes, using Google's Keyword Tool[5]. This service was created to help choosing appropriate ad words and provides the local and global monthly average search volume of a query over the last 12 months for the selected countries, languages and devices (e.g. desktops, laptops, or mobile devices).[6]

Setup. Since Google Keyword does not allow for submitting a large number of queries automatically, we employed a crowdsourcing solution. To this end, for each query, we created a Human Intelligence Task (HIT) in Amazon Mechanical Turk (AMT) that asked workers to submit a given query to the Google Keyword Tool and to provide the returned volume (or, "-1" if no volume information was found). To study the agreement among workers for this task, we assigned all of the 378 queries for two of our topics, namely "abortion" and "euthanasia" to two different AMT workers.

[5]`https://adwords.google.com/o/KeywordTool`

[6]We chose not to use Google Trends[7], because, although employed in recent works (e.g. [Preis et al. 2013]), it only provides relative volume information and, thus, cannot be used for comparing and aggregating volumes across different queries.

Crowdsourcing provided very reliable results for this task: the volume values entered to HITs differed only for two of the queries. Due to the very high overlap among the annotators, the remaining queries were assigned to only one AMT worker.

Results. Among the 7,651 annotated queries in out dataset, the Keyword Tool did not provide a volume value for 3,256 (42.54%) of the queries. This might be due to the time gap between constructing our query set and using the Keyword Tool for collecting volumes, or differences in the procedure used for generating query auto-completions and volume values by the search engine. Nevertheless, there is no significant bias towards a particular sentiment class; we found that 41%, 54%, and 43% of the queries labeled as objective, positive, and negative, respectively, had no associated volume. For the remaining 4,392 queries, we observed a typical heavy-tailed distribution of query frequencies. The total volume adds up to 257,751,783, with objective queries amounting to 97% of the volume. However, a non-negligible amount of 3% of the query volume (i.e., around 7.5 million queries per month) for our topics are opinionated. The imbalance between negative and positive queries (as shown in the previous section) is still apparent yet less strong, with each class containing 4,319,345 and 3,208,531 queries, respectively.

3.3. Sentiment in Query Results

In addition to analyzing the sentiment in query texts, we also investigated the traces of bias in the query results. In their previous study, Demartini and Siersdorfer employed an automatic approach to investigate the opinions expressed in top-ranked results of controversial topics [Demartini and Siersdorfer 2010]. In that study, different from our work, the initial query (i.e., simply the topic name) is not opinionated, and the goal was to analyze the search result lists for queries on controversial issues. The authors report that on average, the top results returned from three different search engines do not express extreme opinions. We complement and extend that study by providing a manual analysis on the results of opinionated queries in the rest of this section. We discuss the lexicon-based analysis of opinions in queries in Section 3.5.

Setup. For our study, we randomly selected three queries labeled as objective, positive and negative for 20 of our topics (again chosen uniformly at random) during the annotation process. For each query, we retrieved the top-10 query result titles and snippets via the Yahoo! API. Next, we shuffled these query results and annotated each title and snippet as positive, negative, or objective. In this way we obtained 600 annotated titles and the same number of annotated snippets.

Results. Figs. 2(a) and (b) show the distribution of sentiments in result titles and snippets, respectively, for each query class. Our experiment reveals that, regardless of the opinion in the query, most search results do not express a considerable bias towards an opinion. Titles, which are shorter, are found to be more objective than snippets. On the other hand, the fraction of positive (negative) results is larger for positive (negative) queries than for the other queries. For instance, the fraction of negatively labeled snippets retrieved for negative queries is up to 50% higher than negative snip-

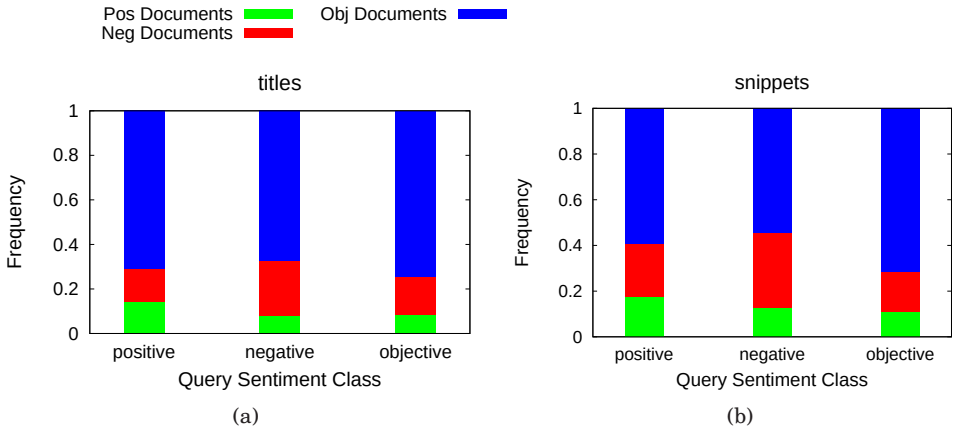


Fig. 2: Sentiment distribution of (a) query result titles, and (b) query result snippets for the queries from each sentiment class.

pets retrieved for positive or objective queries. This implies that, although the majority of search results are objective, the sentiment of a query is also reflected by the results to some extent.

3.4. Post-Retrieval Analysis

As mentioned in Section 1, we do not assume that an opinionated query does necessarily express the personal view of the user issuing it. For instance, a person may submit the query "abortion is a sin" to see the arguments of the people holding that opinion, or just to see whether such an opinion exists for this topic. Thus, we cannot guarantee that the existence of the query corresponds to the opinion of the user who submitted it, but we can identify that this particular opinion exists for the topic in the query and furthermore the opinion was searched for by a large number of users, as justified by the aforementioned query volume analysis. However, it is still worthwhile and illustrative to analyze the post-retrieval behavior of the user, i.e., *how* she behaves after the results are displayed. Although this cannot perfectly explain why she submitted the query, it can help to verify whether she was really looking for a particular opinion.

Setup. We conducted a small-scale analysis on an MSN query log excerpt (the RFP 2006 dataset) that contains 15 million queries along with the full URL information for the clicked results. (The AOL log used in the other parts of the paper turned out to be useless as for the clicked results only the top-level domain of the URL is provided.) We chose 5 topics ("abortion", "euthanasia", "genetic engineering", "marijuana", and "stem cell research") out of the 50 used in our paper, for which some related queries in the log could be found. For each of these topics, we annotated the queries as objective, positive or negative, yielding 79 (5%) opinionated queries among a set of 1,583 queries. For these 79 queries, there existed a total of 222 clicked URLs. For each of the clicked

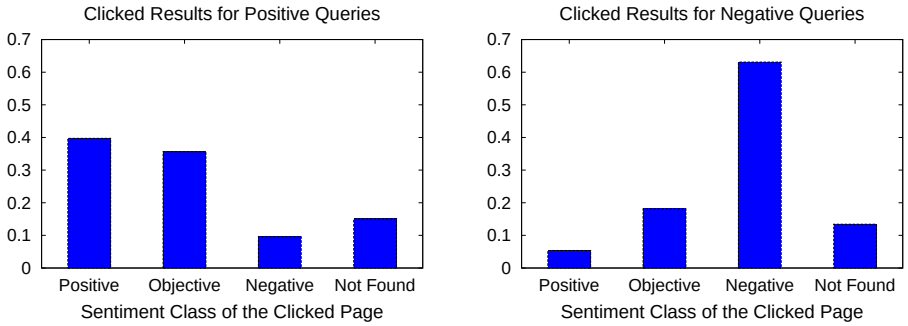


Fig. 3: Sentiment distribution of the clicked results for (a) positive queries, and (b) negative queries. We also show the fraction of the pages that are not found, i.e., not accessible online anymore.

pages, we used the Way Back Machine[8] to get the version back from 2006 (if available) and annotated them as objective, positive or negative. In this way, we gathered and annotated 191 clicked pages.

Results. Figs. 3(a) and (b) show the percentage of clicked pages per sentiment class for the queries labeled as positive and negative. For the positive queries, the majority of the clicked pages were either positive or objective, with each class containing approx. 40% of the clicks. (The high number of objective queries is consistent with our findings above that the majority of the retrieved results by the search engines are objective, regardless of the sentiment in the query.) For the negative queries, more than 60% of the clicked pages are negative. Therefore, our results provide evidence that users who submit opinionated queries (especially negative ones) are likely to click on opinionated results in the same direction (i.e., these queries really serve as a mechanism for accessing opinionated material on the topic in question).

3.5. Lexicon-Based Sentiment Analysis

3.5.1. Sentiment in Queries. We also investigated whether the human judged labels for the 7,651 queries match to automatically obtained sentiment scores using the SentiWordNet thesaurus. SentiWordNet [Esuli and Sebastiani 2006] is a lexical resource built on top of WordNet [Fellbaum 1998] that assigns a triple of sentivalues (pos, neg, obj) to each concept in WordNet. Each value in a triple reflects the tendency of words or concepts to be in the corresponding class, with values in a triple summing up to 1.

Using SentiWordNet we first assigned a sentivalue to each query by computing the averages of positivity, negativity and objectivity values over the adjectives extracted from the query text that have an entry in the SentiWordNet thesaurus. If an adjective appears in more than one WordNet concept (synset), the sentivalues for each occur-

[8] `web.archive.org`

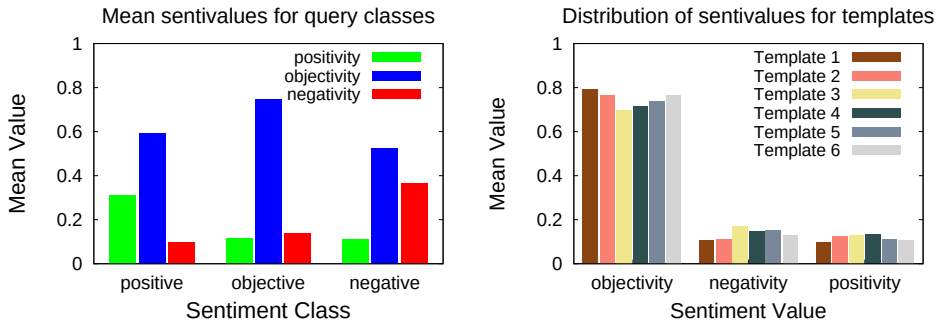


Fig. 4: (a) Mean sentivalue scores (from SentiWordNet) in each query class, (b) Distribution of average sentivalue scores of queries (from SentiWordNet) obtained from each template.

rence are averaged to obtain the triple for this term. We used only adjectives because our experiments with additional term types (i.e., nouns and verbs) yielded less accurate results; a similar observation is also reported for the short user comments in YouTube [Siersdorfer et al. 2010]. At the end, an overall number of 2,517 queries (i.e., 31% of the manually annotated queries) was found to contain adjectives covered by SentiWordNet.

Fig. 4(a) shows the means of positivity, negativity and objectivity scores over all queries in each class as labeled by human judges. For all three classes, we observe that the objectivity scores are rather high. This might be due to the fact that some of the positive/negative terms in the queries do not appear in the thesaurus. Nevertheless, the positively (negatively) labeled queries yield a considerably higher positivity (negativity) than negativity (positivity) score. Fig. 4(b) shows the distribution of average sentivalues for queries from each template in Table II. A comparison with Fig. 1 reveals that automatically derived sentivalues for each template follow a similar distribution as human annotated class labels. However, the positivity and negativity scores are lower, as already discussed for Fig. 4(a). To remedy this problem, we applied a machine learning based approach that will be discussed in Section 4.

3.5.2. Sentiment in Query Results. Finally, we studied the sentiment of query results as in the previous section. In particular, for all queries in each sentiment class, we computed the sentivalues for the adjectives in the query text and the top-10 result snippets gathered via the Yahoo! API.

Fig. 5 shows the distribution of human labeled queries from three sentiment classes across the sentivalue bins for neutrality, positivity, and negativity, respectively. The histogram in Fig. 5(a) can be regarded as further evidence supporting the trends observed in the previous section and in [Demartini and Siersdorfer 2010]: the results for objective queries also yield the highest objectivity scores, especially when the score is

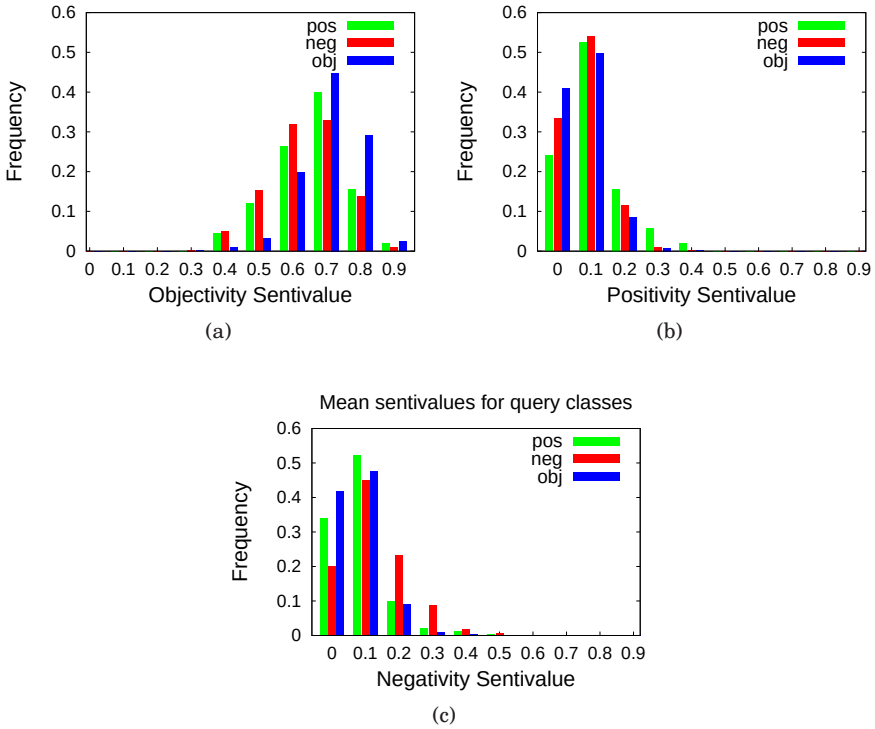


Fig. 5: Distribution of query snippets' (a) objectivity, (b) positivity, and (c) negativity sentivalue scores (from SentiWordNet) in each query sentiment class.

larger than 0.7 (i.e., indicating higher confidence). In contrast, for opinionated queries, the snippets also reflect the opinion to some extent (cf. Figs. 5(b) and 5(c)).

3.5.3. Summary. Our findings in this section reveal that even a limited-vocabulary based sentiment analysis strategy serves well in our framework and yields results quite consistent with manual annotations. In the following sections, we employ machine learning techniques for automatic sentiment analysis to facilitate the adaptation to the rapidly changing vocabulary of Web users.

3.6. Regional Analysis

Opinions regarding a controversial topic can vary considerably with respect to location of the searcher and time of the search. For instance, a controversial topic such as "gay marriage" might be perceived more positively in Europe than in the Middle East. Similarly, for the same example topic, opinions have become less negative over time as social tolerance on such issues has increased. In this section, we provide an analysis of the impact of region on the sentiments expressed in queries for controversial topics.

Setup. We used the query collection strategy outlined in Section 2. In order to avoid issuing excessive numbers of requests to search engines, we focused on template 5

Table V: Topics and the number of manually annotated queries (obtained via template 5) in each of the three languages (English, German and Spanish)

Topic	**Queries**	**Topic**	**Queries**	**Topic**	**Queries**
abortion	223	abtreibung	31	aborto	95
barack obama	233	barack obama	32	barack obama	58
climate change	205	klimawandel	41	cambio climtico	29
communism	208	kommunismus	27	comunismo	62
economy	227	wirtschaft	45	economa	29
homosexuality	208	homosexualitt	53	homosexualidad	62
iphone	260	iphone	130	iphone	141
islam	243	islam	128	islam	52
marijuana	226	marihuana	25	marihuana	110
marriage	240	ehe	190	matrimonio	86

(i.e.,"⟨*topic*⟩ is [*letter*]", as shown in Table II), which turned out to yield the largest fraction of opinionated queries. To obtain region-specific queries, we sent the search requests in English[9], German, and Spanish to the corresponding search front ends with domains extensions .com, .de and .es, respectively. For German and Spanish, we revised template 5 with the auxiliary verbs in the corresponding language. Note that the scenario of English queries submitted to the main front end of the search engine represents a rather global case, whereas queries in German or Spanish might reveal more region-specific opinions of the web users (assuming that the majority of queries in German and Spanish are submitted from the corresponding countries).

We observed that the number of opinionated queries gathered with template 5 is smaller for German and Spanish than that for English (see Table V). This might be due to the unequal volumes of queries, as English queries constitute the largest query stream for most search engines. We emphasize that our choice of a fixed template for collecting opinionated queries is essentially caused by the lack of very large publicly available query logs. We aimed to gather queries that are more probable to be opinionated with the least possible burden to the suggestion system of the search engine.

Among our initial set of 50 topics, we identified 10 topics (listed in Table V) for which all three front ends returned at least 25 queries. Next, all of the queries retrieved for these topics were manually annotated by native speakers of the corresponding language using the guidelines of Section 3.1. In this way we obtained 724, 702, and 2,272 annotated queries for the Spanish, German, and English front end, respectively.

Results. Figs. 6(a), (b) and (c) show the fraction of queries that are labeled as positive, negative and objective for each topic and region/language. A comparison among the figures reveals that, as expected, various topics are perceived differently among the users that submitted their queries to distinct front ends. For instance, the queries submitted in Spanish and German for the topic "climate change" exhibit a considerably

[9]Note that, we repeated the data collection for English to obtain queries for all three languages at the same point in time.

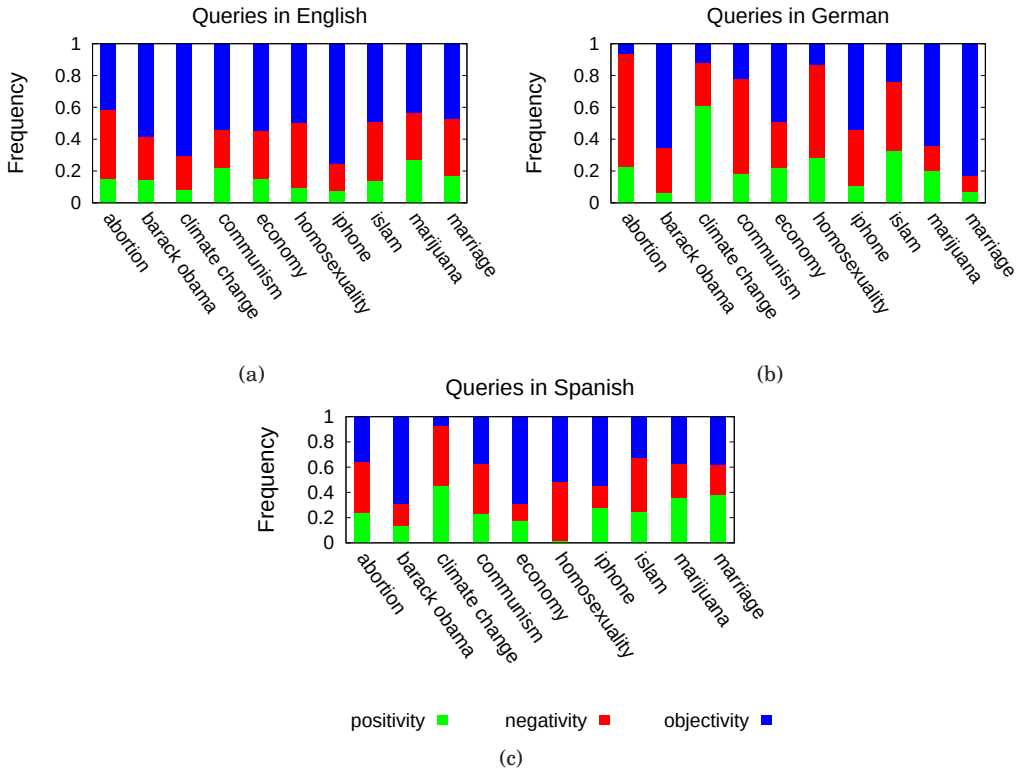


Fig. 6: Distribution of sentiment class annotations for each topic using queries submitted in (a) English, (b) German, and (c) Spanish.

higher positivity in comparison to those submitted in English to the main front-end. The higher positivity in Germany and Spain might be explained by the EU countries' leading role in developing policies related to climate change as well as the existence of supportive political groups in these countries (such as the Greens in Germany). On the other hand, the United States, from where most of the English queries are possibly submitted, are known to be rather reluctant to related legislation on the issue (see e.g. [Wilkinson 2012]). While Fig. 6 indicates noticeable differences in the perception of the topics, analyzing the underlying reasons for such differences is beyond the scope of this study and the authors' expertise. However, we believe that our findings unleash the potential of analyzing opinionated queries, which is a rather overlooked source of information up to now.

We emphasize that our findings in this section regarding the regional dynamics of opinionated queries are not comprehensive, as it is difficult to obtain datasets and ground truth from different regions and at many points in time (that is why we leave the temporal dimension as a future work). However, our examples motivate the investigation of opinionated queries and methods for automatic sentiment analysis of

queries, and imply interesting applications, such as trend analysis and detection of controversial topics.

4. DETECTING QUERY SENTIMENT

In this section, we study the application of various state-of-the-art classifiers to detect the sentiment class of a given query.

4.1. Setup

For our classification experiments, we constructed feature vectors that based on the top-10 result titles and snippets, in addition to query text itself. We considered 4 different representations for a given query: (i) query text only (denoted as *QText*), (ii) query text + titles for top-10 query results (*QTextTitle*), (iii) query text + snippets top-10 query results (*QTextSnippet*), and (iv) query text + titles + snippets for top-10 query results (*QTextTitleSnippet*). We constructed multi-dimensional feature vectors using tf-idf weights of the terms involved in each possible representation. While doing this, we also accounted for negations (i.e., if a negation, say, "not", immediately precedes another term t, we created a virtual term *not_t* in a similar way as described in [Pak and Paroubek 2010]).

We used five state-of-the-art text classification approaches: simple logistic regression (SLR), multinomial Naive Bayes (mNB), SVM (SMO variant) and SVM (L2-loss linear) as implemented in the well-known Weka library [Hall et al. 2009], and the ν-Support Vector Classification (ν-SVC) formulation of SVM from LIBSVM [Chang and Lin 2011]. We built three types of binary classifiers to separate each sentiment class from the other two classes, i.e. we applied a "one vs. all" (OVA) strategy. We build four different versions of each classifier based on the query representations discussed above.

For training the classifiers, we randomly split the instances from the target class into two sets reserved for training and testing, and randomly selected an equal number of instances from the remaining two classes for training as well as for test sets. In this way, we created balanced training and test sets for each classifier.[10] We repeated the experiments by switching training and test sets and computed the averages for the evaluation metrics. We chose the number of training queries in such a way that the maximum number of available annotated queries could be used during the training and testing. For instance, as around 800 queries are annotated as positive, the positive vs. all classifier was trained with 400 queries from the positive class and 200 queries selected from each of the negative and objective classes. The test set was created analogously. For the negative vs. all and subjective vs. all classifiers, it was possible to use more training queries as there exist a larger number of annotated queries for these scenarios; therefore, we trained and evaluated them with 1,200 and 1,600 queries, respectively.

Table VI: Classification accuracy and AUC for the subjective vs. all classifiers trained with four different representations of the queries (*QAll* stands for *QTextTitleSnippet*).

	Accuracy				AUC			
	QText	QTextTitle	QTextSnippet	QAll	QText	QTextTitle	QTextSnippet	QAll
mNB	0.76	0.73	0.72	0.72	0.86	0.81	0.79	0.79
SLR	0.80	0.79	0.73	0.73	0.85	0.84	0.80	0.80
SVM (L2-LL)	0.81	0.80	0.74	0.75	0.81	0.80	0.74	0.75
SVM (SMO)	0.80	0.77	0.71	0.70	0.81	0.77	0.71	0.71
SVM (ν-SVC)	0.80	0.80	0.74	0.75	0.86	0.85	0.82	0.82

Table VII: Classification accuracy and AUC for the positive vs. all classifiers trained with four different representations of the queries (*QAll* stands for *QTextTitleSnippet*).

	Accuracy				AUC			
	QText	QTextTitle	QTextSnippet	QAll	QText	QTextTitle	QTextSnippet	QAll
mNB	0.68	0.63	0.62	0.61	0.75	0.68	0.66	0.65
SLR	0.71	0.66	0.62	0.62	0.76	0.70	0.65	0.66
SVM (L2-LL)	0.73	0.66	0.64	0.63	0.73	0.66	0.64	0.63
SVM (SMO)	0.72	0.68	0.61	0.61	0.72	0.68	0.62	0.61
SVM (ν-SVC)	0.73	0.70	0.64	0.64	0.81	0.76	0.70	0.71

Table VIII: Classification accuracy and AUC for the negative vs. all classifiers trained with four different representations of the queries (*QAll* stands for *QTextTitleSnippet*).

	Accuracy				AUC			
	QText	QTextTitle	QTextSnippet	QAll	QText	QTextTitle	QTextSnippet	QAll
mNB	0.72	0.67	0.66	0.66	0.80	0.73	0.71	0.71
SLR	0.73	0.67	0.63	0.62	0.79	0.72	0.66	0.67
SVM (L2-LL)	0.76	0.69	0.67	0.66	0.76	0.69	0.67	0.66
SVM (SMO)	0.77	0.69	0.63	0.63	0.77	0.69	0.63	0.63
SVM (ν-SVC)	0.76	0.71	0.64	0.67	0.84	0.78	0.70	0.73

4.2. Results

We first evaluated each classifier in terms of the classification accuracy and area under the curve (AUC) (for the Receiver Operating Characteristic (ROC) curve). The evaluation results in Tables VI, VII, and VIII reveal that using richer query representations (i.e., with titles and snippets) does not result in additional gains compared to simply using the query text itself. Indeed, query text alone is adequate to decide on the sentiment of a query with good accuracy, especially for the positive (negative) vs. all classifiers. This is not surprising, as users usually try to convey their information need clearly and concisely in their keyword queries. In contrast, longer texts, such as blog entries or reviews, may usually involve more sophisticated use of language (e.g., id-

[10]This is similar to the approach employed by [Bermingham and Smeaton 2010] to eliminate the effect of any underlying bias for a particular sentiment class in the data.

ioms, metaphors, or irony), which can make sentiment analysis more difficult [Ahmad 2011]. A similar observation is also reported for sentiment classification in microblogs, where brevity turned out to be an advantage [Bermingham and Smeaton 2010].

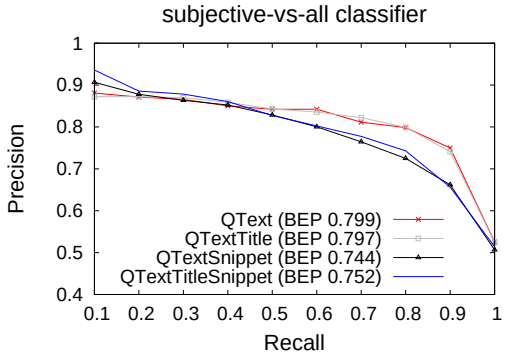
The results in Tables VI, VII, and VIII show that mNB and SLR are usually inferior to SVM classifiers for the query sentiment detection task, and among the latter group of classifiers, ν-SVC (from LIBSVM) performs the best. Using only query texts, binary ν-SVC classifiers *positive vs. all*, *negative vs. all* and *subjective vs. all* yield accuracy values of 0.74, 0.76 and 0.80, respectively. Fig. 7 shows the performance of ν-SVC classifiers for each query representation in terms of precision-recall curves and break-even points (BEPs) for these curves (i.e., precision/recall at the point where precision is equal to recall, which is also equal to F1 in that case). The major trends are similar to previous findings; classifiers based on the query text are superior to those that make use of additional information. Result snippets seem to be slightly useful for distinguishing subjective queries from the objective ones at low recall values (i.e., up to 0.40). Actually, scenarios that allow trading recall against precision are perfectly supported by all classifiers; for instance, for the positive vs. all and negative vs. all classifiers based on the query text, precision values remain over 0.9 up to a recall level of 0.4. This can be useful for finding specifically strong candidates of opinionated queries in large query logs.

For the best performing query representation, namely *QText*, we also applied two well-known lexicon based methods from the literature: SentiWordNet (SWN) [Esuli and Sebastiani 2006] and SentiStrength [Thelwall et al. 2010]. We used the test queries employed in each of the classification tasks above. SWN yields accuracy values of 0.65, 0.63, and 0.65 for the test sets employed in positive-, negative- and subjective-vs-all classification experiments, respectively. SentiStrength yields slightly lower accuracy (0.62) than SWN, for positive vs. all, but is superior to the latter method at detecting negative and subjective queries, resulting in accuracy values of 0.72 and 0.68, respectively. Nevertheless, these figures are considerably lower than those for the machine learning based strategies (a finding that confirms Figure 1 in [Bermingham and Smeaton 2010]); therefore, we do not discuss the lexicon based methods for query sentiment detection in the rest of this section.

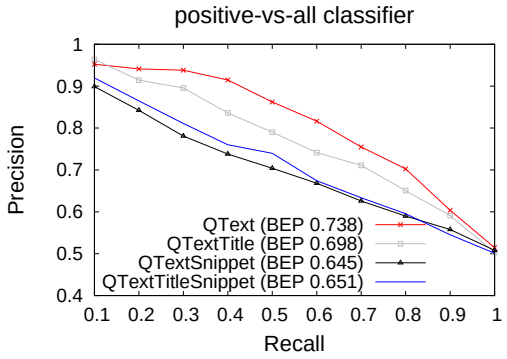
4.3. Do classification models generalize to new topics?

Setup. We repeated the entire experimental procedure by applying a topic-wise split of training and test sets. To this end, queries from the first 25 topics were used for training and those from the second half were used for testing (again, by selecting equal number of queries and also taking into account the number of annotated queries from each sentiment class) and vice versa. We employed the best-performing classifier in the above experiments, namely, ν-SVC.

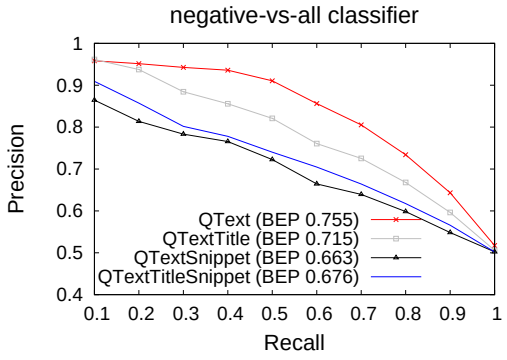
Results. The trends observed are similar as for our previous classification experiments. We obtain prec=0.89 for recall=0.2, and prec=0.83 for recall=0.4 for the positive vs. all classification task using query terms (i.e., *QText*), indicating that it is possible to trade recall against precision for better applicability. (We discarded PR-curves for



(a)



(b)



(c)

Fig. 7: Precision-recall curves and BEPs for (a) subjective vs. all, (b) positive vs. all, and (c) negative vs. all classifiers.

these experiments for brevity.) Our findings show that even for previously unseen topics, our classifiers perform well at detecting the sentiment in queries. Furthermore, even if new contexts that require annotating additional queries may arise in time, an-

notating the sentiment in queries would be a less labor-intensive task than annotating full-length documents. This is another reason for exploiting sentiment in queries as proposed in this paper.

5. QUERY RECOMMENDATION

As described in Section 1, there are various interesting scenarios that can benefit from detection of sentiment in Web queries. In this section we focus on the task of recommending additional queries for opinionated queries as a use case. More specifically, we investigate the potential of improving the relevance of suggested queries by analyzing the sentiment in the submitted query, and suggesting queries in the same direction.

5.1. Recommender Methods

For the query recommendation scenario, we first trained a positive (negative) vs. all sentiment classifier in a leave-one-topic-out manner, i.e., by using 49 topics for training and one for testing. We used balanced sets with equal number of randomly selected instances from each class. Next, we ranked queries for each topic based on the distance from the separating SVM hyperplane, and used the query classified as positive (negative) with the highest confidence as seed query for the topic. Then, we generated query suggestions for each seed query in two ways: 1) As a baseline, we issued the seed query to a major search engine (i.e., the same one used in the other parts of this paper), and collected all suggestions that were obtained through auto-completions, or recommended on the result page under "related queries" (only the top-10 were selected). We name this set "search engine suggestions". 2) We selected the same number of queries from the distance ranked list of classified queries for the same topic (except for the seed query itself). We refer to this set as "opinionated suggestions".

5.2. Evaluation Setup

For evaluating the query recommendations, we shuffled both suggestion sets and conducted a user study where subjects were asked to label each query as relevant/irrelevant/undecided with respect to the seed query. To reduce the manual workload of the participants, we decided to consider only 15 out of the 50 topics with highest polarity with respect to the ground truth annotations discussed in Section 3. We gathered the manual assessments using two sets of annotators: in-house annotators and annotators from a crowdsourcing platform.

For the in-house annotations, five computer science researchers/students were involved who were not aware of the final of goal of our evaluation. Each of the human judges was randomly assigned 6 seed queries along with the suggestions. We made sure that the seed queries from the same topic were assigned to different judges. On average, the participants annotated a combined list of around 20 suggestions per seed query and topic. We considered two seed queries (i.e., the most positive and negative ones) for each of the 15 topics, yielding 600 annotated suggestions for 30 seed queries.

For crowdscourced annotations, we created a Human Intelligence Task (HIT) at Amazon Mechanical Turk (AMT) for each pair of a seed query and a suggested query,

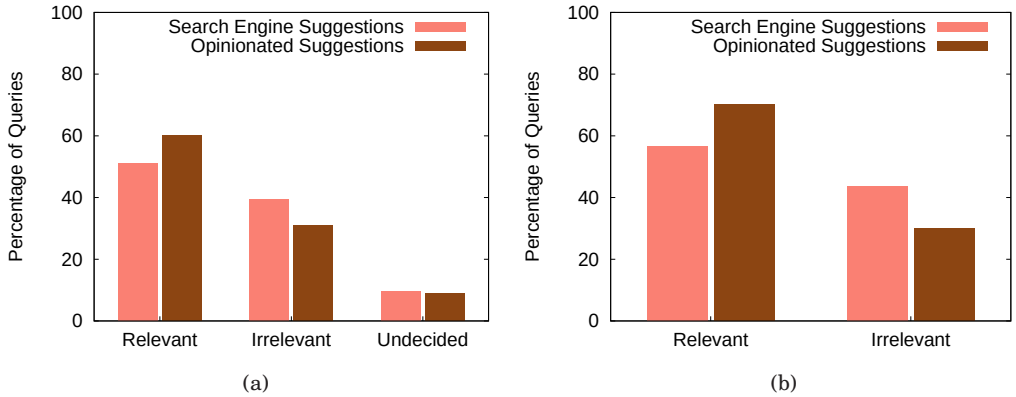


Fig. 8: Query recommendation performance based on (a) in-house annotations, and (b) AMT annotations.

where we asked the workers to label the suggestion as relevant, irrelevant or undecided. Each HIT was assigned to five different workers and the final decision was computed based on majority voting.

5.3. Results

Figs. 8(a) and (b) show the average percentage of "search engine suggestions" and "opinionated suggestions" that are labeled as relevant, irrelevant or undecided by the in-house and AMT annotators, respectively. In contrast to the in-house study where each suggestion is labeled by one annotator, the AMT evaluation provides five annotations per suggestion and hence, no query suggestion is labeled as "undecided" after the majority voting (see Fig. 8(b)). Still, 67.3% of the AMT annotations agree with those of the in-house annotations. The Fleiss's Kappa for the inter-user agreement among the AMT workers was found to be 0.68. Figs. 8(a) and (b) reveal that the opinionated suggestions are more relevant than the original recommendations from the search system. We applied an unpaired t-test to compare whether the difference between the mean relevance scores (based on the in-house annotations) of two recommendation lists is significant (assuming that all undecided queries are also irrelevant), and found that our improvements are statistically significant on a 95% confidence level (with df=628, $|t|=2.01$). We verified the statistical significance also for the results based on the annotations of AMT workers (with df=595, $|t|=3.56$). Furthermore, we observed that there is only little overlap between the two recommendation lists, indicating that result set merging can further improve the relevance of query recommendations.

A detailed inspection of the results provided further interesting insights. We noticed that for most of the seed queries, there are no auto-completions provided, which are usually very accurate; instead, only a list of, more error prone, "related queries" is shown. In practice, our sentiment based recommendation mechanism can be applied in situations where no or few auto-completions are available (see Table IX). We also

Table IX: Search engine's suggestions (provided as "related queries" and "auto-completions", the latter are shown in italics) vs. opinionated suggestions for the query "economy is really bad"

Search Engine Suggestions	Opinionated Suggestions
economy is really bads	economy is bad
economy is really bad right now	why is economy bad
economy is really bad 2009	economy is still bad
economy is really band	economy is very bad now
economy is really good	economy is getting worse
economy is really funny	economy is obama's fault
economy is really bag	economy is worse than divorce
economy is really dirty	economy is killing people
gdp is really bad	economy is destroyed
mileage is really bad	economy is going to get worse

observed that the top-10 search results vary largely for the different recommended queries listed in Table IX, confirming that this type of query reformulation can provide additional information and perspectives. While our primary goal here is to provide more relevant recommendations that are aligned with the sentiment of the seed query, our approach can also be employed for improving the diversity of recommendations (as in [Song et al. 2011]) by suggesting queries in the directions other than the user's opinion. We plan to explore the combination of our sentiment-based approach with original search engine suggestions and other query recommendation approaches/scenarios in our future work.

6. CONTROVERSIAL TOPIC DISCOVERY

Trend analysis on opinionated digital data is an emerging area that has drawn substantial attention over the last years [Goorha and Ungar 2010; Allan 2002]. We anticipate that controversial topic discovery can be an important stage of trend analysis studies, as it sheds light on the issues on which Web users have diverse opinions. For instance, the drift of controversial topics discussed in a society over time may indicate underlying changes in its value system.

Controversial topics can have regional and temporal aspects (as it is also implied in our discussion in Section 3). Mining opinionated text on the Web such as blogs or reviews in order to discover controversial topics for a particular region of the world and for a specific time period would require sophisticated and expensive techniques to accurately detect these spatio-temporal features as well as the sentiment expressed in relatively long and complicated articles. First, one has to deal with the traditional and hard problem of capturing the sentiment from a natural language text (e.g. a blog post discussing a popular product such as the iPhone) with metaphors and ironical statements. Second, there is the issue of capturing the time and region the post is intended for (for instance, the post might be discussing an older or future version of

the product, or identifying problems specific for a particular country which could even differ from the host country where the blog is published).

We envision that the huge volume of queries submitted to Web search engines can be employed for opinion mining with considerably less effort. It is easy to associate queries with a particular region, as search engines already keep track of the search front ends to which a query is submitted for localization and personalization purposes. Furthermore, queries can often be seen as short and concise statements about the topics in question. Therefore, query logs accumulated over a sufficiently long period from different search front ends can serve as an invaluable resource for discovering controversial topics in a desired region and time period. In this section, we show the applicability of our query sentiment classifiers in this context.

6.1. Topic Discovery Method

In an idealistic setup, it would be sufficient to classify the sentiment of each query in the query logs of a search engine to infer potentially controversial topics. Since such large-scale query logs are not publicly available we devised a two-stage selection and filtering strategy (cf. Fig. 9) to provide a proof-of-concept for this scenario, instead. In the first stage, the *candidate topic generation* step, we formed a set of queries that were prefixed by any combination of three letters from the English alphabet and followed by the term "is" (e.g., "mil is"). Next, we trained a machine learning model as discussed in Section 4 to distinguish queries with positive and negative sentiments (we dismissed the objective class as it is not useful for our purposes in this section). The queries in this set were sorted with respect to the classifier scores, from the most positive instances to the most negative ones. Finally, we selected the queries from the top and bottom $P\%$ of the list formed in the previous step, removed the part starting with "is", and grouped with respect to the remaining topic names. Those topic names that appeared more than K times in our set were identified as *candidate topics*.

We observed that while a too high K yields only few topics, a too low value of K produces rather noisy topics. For instance, for the template "lef is", suggestions include "**lef**t is right" and "**lef**t is seldom right", which clearly state an opinion about the political left. However, also the query "**lef**t ventricular hypertrophy is reversible" is suggested, which might be classified as rather positive, but probably does not refer to a controversial topic. Therefore, in the second (*controversy discovery*) stage, we collected all query suggestions using template 5 (i.e., <topic> is [letter], as shown in Table II) for the candidate topics and again classified them using our classifier, to verify whether a large number of opinionated queries existed for the given topic. As might be expected, for the above example, the candidate topic "left" yields a large number of query suggestions that are opinionated, whereas "left ventricular hypertrophy" yields none. In this step, we chose topics that had at least N opinionated queries, along with the classifier scores for these queries. Finally, we ranked the topics according to the *variance* of the classifier scores, envisioning that topics with a higher variance in query sentiments would be more controversial. The entire process is illustrated in Fig. 9.

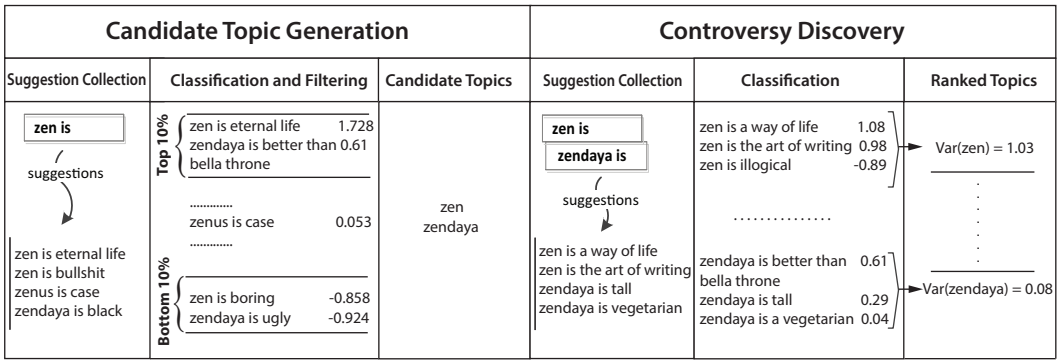


Fig. 9: A toy example illustrating controversial topic detection: the procedure will output only "zen" as being controversial, as it yields very high variance in query sentiment scores and filter "zendaya", as its queries have less variance.

In this paper, we set parameter P to 10%, K to 2, and N to 50, in an ad hoc manner. The initial query suggestion set includes 98,359 queries, and almost one third of them could be classified by our detector (for the rest, none of the terms apart from the topic name appeared in the trained model). After applying these steps, we ended up with a ranked list of 273 topics from which we also removed stopwords and adjectives, resulting in an overall number of 263 topics. Note that although adjectives are very important in the sentiment detection step they usually do not correspond to actual topic *labels*. The variance scores in this list starts from 1.0, representing a probability of high-controversy and drops to 0.3 at the end of the list, corresponding to a potentially non-controversial topic.

6.2. Example Results

The top-20 (controversial) and bottom-20 (non-controversial) topics are shown in Table X. The most controversial topic, *dairy*, reflects a popular and hot debate on whether food products produced from the milk of mammals are healthy or not. *Wicca* is defined as a modern pagan religion in Wikipedia that gives rise to contradicting opinions, as some people apply attributes like *fake*, *evil*, or *stupid*, whereas others think that it is *cult*, *good* and *right*. *Splenda* is an artificial sweetener and *msg* is short for monosodium glutamate used as a flavor enhancer in food, both of which seem to trigger highly polarized views. Notice that, except for the topic *left*, which has been a controversial issue in politics for centuries, the other 4 topics among the top-5 discussed above can not be easily detected, and reflect the Web searchers' popular discussion issues at the time of this experiment. In this sense, we believe that our methodology serves well to discover topics otherwise unrevealed that cause controversy within the Web community. In contrast, the bottom-20 topics seem to be less-controversial, as topics in this group are more likely to attract either mostly negative or mostly positive attitude (if they ever cause any polarity). For instance, *wood* is mostly viewed neu-

Table X: Topics ranked with respect to the variance in sentiment scores of their queries

Top-20 topics		Bottom-20 topics	
dairy	ritalin	wood	sitting
wicca	lie	ignorance	yesterday
left	oatmeal	jesus	egg
splenda	vanity	icp	danger
msg	lsd	insanity	justice
hunting	acid	ncis	hell
euthanasia	lying	africa	weird
losing	skateboarding	registry	pakistan
lust	liz	beauty	all i do
abortion	living	pope	truth

trally whereas *ignorance* is mostly perceived negatively. Similarly, *ICP*, a rap band in the US, seems to have a mostly negative reputation.

6.3. Quantitative Evaluation

Setup. For a systematic evaluation of our strategy for controversial topic discovery, we selected the top- and bottom-50 topics and conducted a user study, where each annotator was given a shuffled list of the resulting 100 topics and asked to label the topics as either controversial (denoted with 1) or non-controversial (denoted with 0). We examined if the top-50 topics were significantly more controversial than those in the bottom-50. In order to avoid personal bias, we emphasized that the annotators should not rely on their own perception of a topic, but rather decide on the possibility of existence of large groups of people that would have opposing views on the topic. As in the previous sections, we employed both in-house annotators (4 computer science researchers) and AMT workers (5 Turkers).

Results. For each topic, we decided on the final label (controversial or not) using majority voting, and then computed the grand averages for the top-50 and bottom-50 topics. For the in-house annotations, we found the scores of 0.62 and 0.36 for the top-50 and bottom-50 topics, respectively. Similarly, AMT annotations yielded a score of 0.58 (0.32) for the top-50 (bottom-50) topics. An unpaired t-test showed the statistical significance of the difference of the population mean values on a 95% confidence level for both sets of annotations, with df = 98 and $|t|=2.67$ for the in-house participants and with df = 98 and $|t|=2.68$ for the AMT workers. We observed Fleiss' kappa coefficients [Gwet 2010] of 0.34 and 0.42 for inter-user agreement among the in-house annotators and AMT workers, respectively. (Note that according to Fleiss' definition, $\kappa < 0$ corresponds to no agreement, $\kappa = 0$ to agreement by chance, and $0 < \kappa \leq 1$ to agreement beyond chance.) Also note that, 80% of the labels assigned by AMT workers overlap with those assigned by the in-house annotators. The result of this study provides further evidence for the potential of our controversial topic discovery strategy in a real-life setup.

7. DISCUSSION

This work is the first to analyze and exploit Web queries as a new, rich and mostly unexplored source of user-generated content that can convey community views and opinions on a multitude of topics. To this end, we made a best-effort to provide insightful analyses and interesting and useful applications of opinionated Web search queries using publicly available data sources and crowdsourcing. While using public resources allows others to reproduce and build on our findings, the setup inevitably involves some limitations. In what follows, we describe some aspects of our work that can be enriched and extended - especially by research groups or commercial search engines that have the privilege of having access to query logs or internal mechanisms of large-scale information systems.

— *Quantifying the volume of opinionated queries*: In our study, we proposed exploiting query suggestions via various templates for constructing a large sample of real user queries on a set of controversial topics. In a real search engine, queries mentioning a topic can be more directly identified (for instance, based on the query text or other clues like result page URLs, titles, or snippets) and the volume of the opinionated queries related to a topic can be determined accurately. Furthermore, mining the entire query stream of the search engine allows for quantifying the volume of opinionated queries, which are not necessarily associated with a predefined controversial topic. This can enable the discovery and opinion-oriented analysis of an even broader range of topics.

— *Inferring user intent*: As our query dataset includes only individual queries without information about user sessions, it is difficult to infer the user's actual intent for opinionated queries. While we provide a small-scale experiment to analyze the post-retrieval behavior of the users over a small query log (in Section 2), a search engine can exploit both the click behavior and the query re-formulations following a query for gaining deeper insights into the user's intent and adapting the system accordingly for various application scenarios mentioned in this paper.

  An orthogonal yet related issue is the ambiguity in queries. This does not require special treatment in our setup as the underlying intent is fairly clear for the topics chosen for the analysis in Section 2 (see topic names in Table I). In a broader scenario where opinionated queries are identified from the query logs, typical techniques to disambiguate different intents (again, keeping, for instance, track of the reformulations and user clicks [Radlinski et al. 2010]) can be applied orthogonally to sentiment detection approaches. Thus, using query logs, it would be possible to detect alternative interpretations (e.g. for the topic "apple") and their association with opinionated queries.

— *Developing and evaluating application scenarios*: Recent work shows that the competition among search engines is becoming more focused on queries in the long tail, e.g., for generating better rankings [Zaragoza et al. 2010; Aktolga and Allan 2011]) and recommendations [Szpektor et al. 2011; Broccolo et al. 2012]. Opinionated queries constitute a tiny, yet non-negligible, fraction of the query volume and search engines

can adapt, combine and enhance our application scenarios for better handling and/or exploiting such queries in their systems.
For instance, a search engine can first apply sentiment detectors over a query log to identify potentially opinionated queries and then extract controversial topics that frequently appear within these queries, without requiring the use of templates described in Section 5. Next, while recommending queries for the controversial topics determined in the previous stage, the sentiment in the input query can also be exploited, as we outline in a restricted context in Section 4. The overall performance of the recommendation scenario can be evaluated by alternative means, for instance, by analyzing the follow-up queries in query sessions, or even integrating sentiment-oriented recommendation in the system and observing user behavior (i.e., A/B testing). Finally, search engines can integrate additional features such as sentiment-aware re-ranking of search results for the opinionated queries, result aggregation and targeted advertising, some of which are discussed in the introduction.

To sum up, we believe that our study accomplishes its objective of identifying Web search queries as a new and rich source of information for detecting and exploiting sentiment. In addition to the aforementioned directions for future work, our findings can inspire and trigger further research in social and political sciences. Many recent studies in the social sciences exploit search information for various purposes, such as identifying political tendencies [Weber et al. 2012], detecting public attention [Ripberger 2011] and predicting stock market moves[11] [Preis et al. 2013]. These works indicate that exploiting sentiment in queries can help finding interesting connections between topics and community opinions along with the time and space dimensions and open new avenues for research in the social and political sciences.

8. RELATED WORK

There is a plethora of work on *sentiment classification, opinion mining, and opinion retrieval* [Pang and Lee 2008]. Sentiment classification (described, for instance, in [Pang et al. 2002; Thomas et al. 2006]) deals with the problem of automatically assigning opinion values (e.g. “positive” vs. “negative” vs. “neutral”) to documents or topics using various text-oriented and linguistic features. Recent work in this area makes also use of annotated lexical resources such as SentiWordNet [Esuli and Sebastiani 2006] or SentiStrength [Thelwall et al. 2010] to improve classification performance (e.g., the former thesaurus is employed in [Denecke 2009]). Cross-domain sentiment classification was studied, for instance, in [Pan et al. 2010] where spectral graph analysis is used to infer links between domain-independent and domain-specific terms. In the position paper [Orimaye et al. 2011] the authors provide an overview of challenges in opinion retrieval that arise due to the highly context-dependent character of opinions expressed in Web pages. The authors propose a grammar-based approach to account for opinion-related contexts on a sentence level. There are several works that make use of sentiment thesauri for exploratory studies. For instance, in [Siersdorfer et al.

[11] http://www.bbc.co.uk/news/science-environment-22293693

2010] we use SentiWordNet to analyze sentiment in YouTube comments and the relationship between sentiment and comment ratings. In [Kucuktunc et al. 2012] another sentiment thesaurus is leveraged for studying sentiment in Yahoo! Answers with respect to temporal and demographic aspects. Vural et al. [2012] employs a sentiment thesaurus to guide a focused crawler for discovering opinionated web content. In this work we apply sentiment analysis in what is a novel context, i.e., Web search queries.

In their recent work [Demartini and Siersdorfer 2010] the authors make use of sentiment analysis to compare the sentiment expressed in query *results* (in contrast to the queries themselves) for different search engines. Pera et al. [2011] suggest an approach for summarizing query results with respect to sentiments and facets. [Weber et al. 2012] analyze aspects like topics, trends, and opposition in search results for left and right leaning queries related to US politics. The political polarity of queries (left or right) is determined by click behavior on left or right wing blogs. To our knowledge, the work closest to our study is a recent paper by Gyllstrom and Moens [2011]. In this study, the authors also point out that a number of Web queries represent an opinion; however, they use such queries to detect controversy of a *given* topic in a search engine for children, so that additional protective mechanisms can be triggered if children search for such topics. To decide on the controversy of a topic, they obtain a set of suggestions for a given topic from a major search engine, and then create the negations of these suggestions (using antonyms and negating terms). If such queries with negations (i.e., anti-queries) also appear in the suggestion list of the search engine, the given topic is considered as controversial.

Our work presented in this paper differs from that previous work in several aspects. First of all, our goal in this study is beyond detecting whether a given topic is controversial, or not. Instead, after providing enough evidence on the existence of opinionated queries in real Web search scenarios, we essentially focus on mining and exploiting the opinion expressed in such queries. Second, we build classifiers to detect the sentiment of a query submitted to a search engine. Note that this approach is more general than using query/anti-query pairs as proposed in [Gyllstrom and Moens 2011]. Third, we describe several use cases where such sentiment classifiers can be employed, and provide experimental results for two of these scenarios, namely, query recommendation and controversial topic discovery.

Our work also has some connections to the well-known concept of *semantic markedness* in the linguistics literature, which suggests that for a certain pair of related words, one can be unmarked whereas the other can have a semantic orientation/implication. As exemplified in [Hatzivassiloglou and McKeown 1995], for the adjective pair tall-short, the term “tall” is the semantically unmarked one as there is no implication in the question “How tall is Jack?”, whereas replacing “tall” with “short” in the question would imply that the speaker thinks that Jack is indeed short. We anticipate that findings from this area (e.g., see [Hatzivassiloglou and McKeown 1995; 1997]) can be applied for analyzing and/or detecting the sentiment in queries, which is an interesting future work direction.

Controversy has also been studied for data sources other than queries. [Kittur et al. 2007] analyze conflicts in Wikipedia updates, and apply machine learning techniques using characteristics of the update history as features in order to predict articles containing controversies. In [Vuong et al. 2008] articles in Wikipedia are ranked by controversy using information about the conflicting interaction between contributors. In [Awadallah et al. 2012] the OpinionNetIt system is proposed for extracting opinion holders (e.g. politicians) and their opinions about different topics and facets; the resulting information can be leveraged to detect controversies. Data sources used for information extraction include Google News, Wikipedia, and websites of newspapers. However, we are the first to explore controversy of opinions in the context of query analysis.

In the context of *regional differences between queries* Rogers et al.[12] explore different local Google versions for determining which specific types of rights (e.g. children's rights, patients' rights) are most frequently searched for in different countries. However, they do not analyze the sentiment expressed in queries.

There is a considerable amount of work on *classification of queries* into different taxonomies. Taxonomies can be based on the general user intention such as the Transactional, Navigational and Informational query intents introduced by Broder [Broder 2002] or can be topic-oriented (e.g. "Entertainment" vs. "Sports" vs. "Politics" as well as sub-categories). In [Broder et al. 2007] queries and result documents are used to build language models, and queries are classified into categories of a topic-based taxonomy taken from a leading search engine and consisting of several thousand nodes. In [Kang and Kim 2003] queries are classified as being related to a "homepage finding task" or a "topic relevance task", exploiting information contained in query terms, part-of-speech information in queries, and terms from anchor texts and titles. In addition, there is work on leveraging click graphs and query sessions in order to propagate query category labels [Li et al. 2008], and for category label disambiguation [Cao et al. 2009]. In [Shen et al. 2009] an approach for classifying queries into a product taxonomy is suggested. However, none of these works studies *sentiment* connected to queries.

Query recommendation aims at suggesting additional relevant queries for a given query. Baeza-Yates et. al. [2004], for instance, use a combination of term-based query similarity and query support (obtained through the number of document clicks for queries) to suggest relevant queries. Fonesca et. al. [2003] mine association rules from sets of query sessions in order to identify related queries. A template-based approach for mining recommendation rules involving general entity types such as city, person, or substance is described in [Szpektor et al. 2011]. In [Anagnostopoulos et al. 2010] transition probabilities inferred from subsequent queries in user sessions are leveraged for query recommendation. In contrast to these works, we make use of sentiment-based query relationships in order to recommend queries that are aligned with the sentiment expressed in the original query.

[12] `https://wiki.digitalmethods.net/Dmi/NationalityofIssues`

Many recent works propose techniques for generating, exploiting and presenting *query auto-completions*. Generation of auto-completions can be carried out before or during the submission of a query, and can take into account temporal and contextual factors (e.g., see [Shokouhi and Radinsky 2012; Bar-Yossef and Kraus 2011]). Our work does not address the topic of generating query auto-completions, but rather employs such suggestion services for collecting a large and presumably popular set of opinionated queries that are submitted by real users. Bar-Yossef and Gurevich [2008] propose algorithms to mine search engine suggestion services for other purposes, such as estimating the size and coverage of the suggestion databases and popularity of query terms. A clustering approach for presenting query suggestions is introduced by Jain and Mishne [2010] to increase the user satisfaction. None of these latter works attempt to investigate or exploit the sentiment expressed in queries.

A preliminary (poster) version of this work is published in [Chelaru et al. 2012], and focuses on the sentiment analysis of opinionated queries and provides basic results using manual annotations and lexicon-based techniques that only correspond to a subset of the findings reported in Section 3 of this paper. Our current submission significantly extends this previous work in several ways. First, we provide a more detailed analysis of opinionated queries by investigating the correlation of sentiment in the queries and their top ranked results using manual evaluations and lexicon-based sentiment analysis methods (Sections 3.1 and 3.5, respectively). We also investigate the impact of region on the sentiments expressed in queries (Section 3.6). Second, we develop sentiment detection models using a large set of manually annotated queries (Section 4). Finally, we leverage query sentiment detectors for two novel application scenarios, namely, query recommendation and controversial topic discovery (Sections 5 and 6, respectively).

9. CONCLUSION

We conducted an in-depth analysis to shed some light on different aspects of query sentiment. How frequently are opinions and sentiments expressed in queries? Does query sentiment depend on the regional context? Can query sentiment be an indicator for controversial topics? Can we train models for detecting the sentiment of queries? These are some of the questions we examined in this paper by manually and automatically analyzing a publicly available query log as well as query suggestions gathered from a major search engine. In our classification experiments, we demonstrated that using query text alone is often sufficient for automatically determining the sentiment of queries. This makes a large-scale sentiment-oriented analysis of query logs feasible, and opens many avenues for opinion mining in query logs. In our experiments we showed that automatic sentiment analysis of queries can be applied for discovering controversial topics and for recommending related queries to the user. Directions of our future work involve investigating the benefits of query sentiment detection in other scenarios such as result aggregation, trend analysis, and targeted advertising.

REFERENCES

AHMAD, K. 2011. *Affective Computing and Sentiment Analysis: Emotion, Metaphor and Terminology (Text, Speech and Language Technology)* 1st Edition. Ed. Springer.

AKTOLGA, E. AND ALLAN, J. 2011. Reranking search results for sparse queries. In *CIKM*. 173–182.

ALLAN, J. 2002. *Topic Detection and Tracking: Event-Based Information Organization*. Kluwer Academic Publishers.

ANAGNOSTOPOULOS, A., BECCHETTI, L., CASTILLO, C., AND GIONIS, A. 2010. An optimization framework for query recommendation. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*. WSDM '10. ACM, New York, NY, USA, 161–170.

AWADALLAH, R., RAMANATH, M., AND WEIKUM, G. 2012. Harmony and dissonance: organizing the people's voices on political controversies. In *Proceedings of the fifth ACM international conference on Web search and data mining*. WSDM '12. ACM, New York, NY, USA, 523–532.

BAEZA-YATES, R., HURTADO, C., AND MENDOZA, M. 2004. Query recommendation using query logs in search engines. In *Proceedings of the 2004 International Conference on Current Trends in Database Technology*. Springer-Verlag, Berlin, Heidelberg, 588–596.

BAR-YOSSEF, Z. AND GUREVICH, M. 2008. Mining search engine query logs via suggestion sampling. *Proc. VLDB Endow. 1*, 54–65.

BAR-YOSSEF, Z. AND KRAUS, N. 2011. Context-sensitive query auto-completion. In *Proceedings of the 20th International Conference on World Wide Web*. 107–116.

BERMINGHAM, A. AND SMEATON, A. F. 2010. Classifying sentiment in microblogs: is brevity an advantage? In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. ACM, New York, NY, USA, 1833–1836.

BROCCOLO, D., MARCON, L., NARDINI, F. M., PEREGO, R., AND SILVESTRI, F. 2012. Generating suggestions for queries in the long tail with an inverted index. *Inf. Process. Manage. 48,* 2, 326–339.

BRODER, A. 2002. A taxonomy of web search. *SIGIR Forum 36,* 2, 3–10.

BRODER, A. Z., FONTOURA, M., GABRILOVICH, E., JOSHI, A., JOSIFOVSKI, V., AND ZHANG, T. 2007. Robust classification of rare queries using web knowledge. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, USA, 231–238.

CAO, H., HU, D. H., SHEN, D., JIANG, D., SUN, J.-T., CHEN, E., AND YANG, Q. 2009. Context-aware query classification. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, USA, 3–10.

CHANG, C.-C. AND LIN, C.-J. 2011. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol. 2,* 3, 27:1–27:27.

CHELARU, S., ALTINGOVDE, I. S., AND SIERSDORFER, S. 2012. Analyzing the polarity of opinionated queries. In *Proceedings of the 34th European Conference on IR Research*. Springer-Verlag, Berlin, Heidelberg, 463–467.

DEMARTINI, G. AND SIERSDORFER, S. 2010. Dear search engine: what's your opinion about...?: sentiment analysis for semantic enrichment of web search results. In *Proceedings of the 3rd International Semantic Search Workshop*. ACM, New York, NY, USA, 4:1–4:7.

DENECKE, K. 2009. Are sentiwordnet scores suited for multi-domain sentiment classification? In *Proceedings of the Fourth IEEE International Conference on Digital Information Management*. 33–38.

ESULI, A. AND SEBASTIANI, F. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation*. 417–422.

FELLBAUM, C., Ed. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

FONSECA, B. M., GOLGHER, P. B., DE MOURA, E. S., AND ZIVIANI, N. 2003. Using association rules to discover search engines related queries. In *Proceedings of the First Conference on Latin American Web Congress*. IEEE Computer Society, Washington, DC, USA, 66–71.

GOORHA, S. AND UNGAR, L. 2010. Discovery of significant emerging trends. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, USA, 57–64.

GWET, K. 2010. *Handbook of Inter-Rater Reliability* Second Ed. Advanced Analytics, LLC.

GYLLSTROM, K. AND MOENS, M.-F. 2011. Clash of the typings: finding controversies and children's topics within queries. In *Proceedings of the 33rd European Conference on IR Research*. Springer-Verlag, Berlin, Heidelberg, 80–91.

HALL, M., FRANK, E., HOLMES, G., PFAHRINGER, B., REUTEMANN, P., AND WITTEN, I. H. 2009. The weka data mining software: an update. *SIGKDD Explorations 11,* 1, 10–18.

HATZIVASSILOGLOU, V. AND MCKEOWN, K. 1995. A quantitative evaluation of linguistic tests for the automatic prediction of semantic markedness. In *ACL*. 197–204.

HATZIVASSILOGLOU, V. AND MCKEOWN, K. 1997. Predicting the semantic orientation of adjectives. In *ACL*. 174–181.

JAIN, A. AND MISHNE, G. 2010. Organizing query completions for web search. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. ACM, New York, NY, USA, 1169–1178.

KANG, I.-H. AND KIM, G. 2003. Query type classification for web document retrieval. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, USA, 64–71.

KITTUR, A., SUH, B., PENDLETON, B. A., AND CHI, E. H. 2007. He says, she says: conflict and coordination in wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '07. ACM, New York, NY, USA, 453–462.

KUCUKTUNC, O., CAMBAZOGLU, B. B., WEBER, I., AND FERHATOSMANOGLU, H. 2012. A large-scale sentiment analysis for yahoo! answers. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining*. WSDM '12. ACM, New York, NY, USA, 633–642.

LI, X., WANG, Y.-Y., AND ACERO, A. 2008. Learning query intent from regularized click graphs. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, USA, 339–346.

MANNING, C. D., RAGHAVAN, P., AND SCHTZE, H. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

O'CONNOR, B., BALASUBRAMANYAN, R., ROUTLEDGE, B. R., AND SMITH, N. A. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the Fourth International Conference on Weblogs and Social Media*.

ORIMAYE, S. O., ALHASHMI, S. M., AND SIEW, E.-G. 2011. Frequency of sentential contexts vs. frequency of query terms in opinion retrieval. In *WEBIST*, J. Cordeiro and J. Filipe, Eds. SciTePress, 607–610.

PAK, A. AND PAROUBEK, P. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *LREC*.

PAN, S. J., NI, X., SUN, J.-T., YANG, Q., AND CHEN, Z. 2010. Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th International Conference on World Wide Web*. ACM, New York, NY, USA, 751–760.

PANG, B. AND LEE, L. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr. 2,* 1-2.

PANG, B., LEE, L., AND VAITHYANATHAN, S. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*. Association for Computational Linguistics, Stroudsburg, PA, USA, 79–86.

PASS, G., CHOWDHURY, A., AND TORGESON, C. 2006. A picture of search. In *Proceedings of the 1st International Conference on Scalable Information Systems*. ACM, New York, NY, USA.

PERA, M. S., QUMSIYEH, R., AND NG, Y.-K. 2011. A query-based multi-document sentiment summarizer. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*. ACM, New York, NY, USA, 1071–1076.

PREIS, T., MOAT, H. S., AND STANLEY, H. E. 2013. Quantifying trading behavior in financial markets using google trends. *Scientific Reports 3*.

RADLINSKI, F., SZUMMER, M., AND CRASWELL, N. 2010. Inferring query intent from reformulations and clicks. In *WWW*. 1171–1172.

RIPBERGER, J. T. 2011. Capturing curiosity: Using internet search trends to measure public attentiveness. *Policy Studies Journal 39,* 2, 239–259.

SHEN, D., LI, Y., LI, X., AND ZHOU, D. 2009. Product query classification. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*. CIKM '09. ACM, New York, NY, USA, 741–750.

SHOKOUHI, M. AND RADINSKY, K. 2012. Time-sensitive query auto-completion. In *Proceedings of the 35th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 601–610.

SIERSDORFER, S., CHELARU, S., NEJDL, W., AND SAN PEDRO, J. 2010. How useful are your comments?: analyzing and predicting youtube comments and comment ratings. In *Proceedings of the 19th International Conference on World Wide Web*. ACM, New York, NY, USA, 891–900.

SONG, Y., ZHOU, D., AND HE, L.-W. 2011. Post-ranking query suggestion by diversifying search results. In *Proceedings of the 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, USA, 815–824.

SZPEKTOR, I., GIONIS, A., AND MAAREK, Y. 2011. Improving recommendation for long-tail queries via templates. In *Proceedings of the 20th International Conference on World Wide Web*. ACM, New York, NY, USA, 47–56.

THELWALL, M., BUCKLEY, K., PALTOGLOU, G., CAI, D., AND KAPPAS, A. 2010. Sentiment in short strength detection informal text. *JASIST 61,* 12, 2544–2558.

THOMAS, M., PANG, B., AND LEE, L. 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*. Association for Computational Linguistics, Stroudsburg, PA, USA, 327–335.

TURNEY, P. D. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *ACL*. 417–424.

TURNEY, P. D. AND LITTMAN, M. L. 2002. Unsupervised learning of semantic orientation from a hundred-billion-word corpus. technical report egb-1094. Tech. rep., National Research Council Canada.

VUONG, B.-Q., LIM, E.-P., SUN, A., LE, M.-T., LAUW, H. W., AND CHANG, K. 2008. On ranking controversies in wikipedia: models and evaluation. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*. WSDM '08. ACM, New York, NY, USA, 171–182.

VURAL, A. G., CAMBAZOGLU, B. B., AND SENKUL, P. 2012. Sentiment-focused web crawling. In *CIKM*. 2020–2024.

WEBER, I., GARIMELLA, V. R. K., AND BORRA, E. 2012. Mining web query logs to analyze political issues. In *Proceedings of the 3rd Annual ACM Web Science Conference*. WebSci '12. ACM, New York, NY, USA, 330–334.

WILKINSON, E. 2012. Climate change: Environmental issues vs leadership. Available at `http://www.wateo.org/2012/01/02/climate-change-environmental-issues-vs-leadership-by-elisa-wilkinson/`.

ZARAGOZA, H., CAMBAZOGLU, B. B., AND BAEZA-YATES, R. A. 2010. Web search solved?: all result rankings the same? In *CIKM*. 529–538.