Analysis and Prediction of Multilingual Controversy on Reddit

Philipp Koncar
Graz University of Technology
Graz, Austria
philipp.koncar@tugraz.at

Simon Walk
Detego GmbH
Graz, Austria
s.walk@detego.com

Denis Helic
Graz University of Technology
Graz, Austria
dhelic@tugraz.at

ABSTRACT

Social media users express their opinions about arbitrary subjects, including controversial matters such as the 2020 U.S. presidential election or climate change. Controversial topics typically attract user attention, which often lead to fruitful, but sometimes also heated discussions potentially segregating the community. Understanding features that are predictive of controversy in social media can improve moderation of communities and therefore the public discourse. In this paper, we analyze and predict controversy on the multilingual social platform Reddit. In particular, we compare a large set of textual and user activity features in controversial and non-controversial comments posted in six different languages. Using these features we perform a prediction task and study their predictive strengths for controversy. Our results indicate that, regardless of the language, controversial comments are harder to read, more negative and users follow up faster and more frequently to such comments. Moreover, with our prediction experiment (ROC AUC = 0.79) we find that across all languages user activity is the most predictive of controversy on Reddit. Our results contribute to an improved understanding of controversy in social media and can serve as a foundation for tools and models to automatically detect controversial content posted on such platforms.

CCS CONCEPTS

• **Information systems** → **Social networks**; • **Applied computing** → *Sociology*.

KEYWORDS

controversy, Reddit, analysis, prediction

ACM Reference Format:
Philipp Koncar, Simon Walk, and Denis Helic. 2021. Analysis and Prediction of Multilingual Controversy on Reddit. In *13th ACM Web Science Conference 2021 (WebSci '21), June 21–25, 2021, Virtual Event, United Kingdom.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3447535.3462481

1 INTRODUCTION

Discussions in social media frequently evolve around controversial topics, such as gun laws or abortion, that separate users into agreeing and disagreeing communities [25]. While the discussion of controversial topics may lead to new insights [34], break down user stereotypes [29], raise quality of collaborative efforts due to a higher diversity in debates [30], or increase (anonymous) user attention [37], it is also a catalyst for disputes among opposing communities, eventually resulting in destructive discussions [32].

Due to the sheer amount of discussions taking place on social media platforms, researchers and practitioners already recognized the automatic identification of controversy as an indispensable tool for monitoring of such discussions. In practice, this allows moderators and mediators to intervene timely and resolve conflicts or to advise users to include references backing up their claims in debates on social media. Hence, the prediction of controversy has been studied extensively in existing research [7, 17, 19, 23, 24, 31, 37, 38]. However, most of these studies focused on English content and platforms such as Twitter [15, 16, 19, 26] or Wikipedia [7, 8, 28, 38].

Research question. In this work, we study controversy on the multilingual social news aggregation website Reddit, which our community has not yet widely analyzed with respect to controversy. In particular, we ask how commonly studied discussion features, such as *word usage* [17, 23, 31], *writing style* [17, 19, 31], *sentiment* [7, 23, 38], as well as *user involvement* [24, 37] are predictive of controversy on Reddit and whether their predictive properties carry over to languages other than English.

Approach. We define controversy as a discussion topic that separates users into agreeing and disagreeing groups, which captures a general controversy definition (e.g., from Wiktionary[1]) as a debate of contrary and opposing views. More precisely, on Reddit users can post all types of digital content (e.g., text, videos, pictures and links to other websites) that other users can comment on as well as up-vote or down-vote to indicate agreement or disagreement, respectively. Reddit automatically labels controversial submissions or comments based on these up- and down-votes. Here, we operationalize these controversy labels and analyze over 123 million English, German, French, Italian, Portuguese and Spanish comments posted on 50 different discussion boards (Subreddits).

We base our analysis on well-established features of social media users and postings previously studied in settings different to ours [7, 17, 19, 23, 24, 31, 37, 38]. In particular, we study word usage in comments and perform subgroup discovery to find words that have been distinctively used in controversial and non-controversial comments. We then compute various textual features (e.g., readability, POS tags) to study differences in writing styles and sentiment as well as structural and temporal features (e.g., number of replies, time to first reply) to investigate differences in user involvement. For comparison, we use statistical hypothesis testing to identify significant differences in feature distributions between controversial and non-controversial comments. Next, to answer our research question regarding the predictive strengths of individual features for controversy on Reddit, we perform a range of prediction experiments. Finally, we repeat the same analysis for six languages

[1] https://en.wiktionary.org/wiki/controversy

WebSci '21, June 21–25, 2021, Virtual Event, United Kingdom

ACM ISBN 978-1-4503-8330-1/21/06.
https://doi.org/10.1145/3447535.3462481

to learn more about cultural and linguistic differences regarding controversy in online social platforms. Note that all data and code used for our work is publicly available[2].

Findings & Contributions. Overall, our results indicate that, except for word usage, our features reflect controversy similarly across languages and cultural differences. For example, we find that controversial comments are harder to read and contain significantly higher fractions of negative sentiment as compared to non-controversial comments. We confirm results from previous research [24, 37] and find increased user participation in discussions of controversial topics also on Reddit. Our prediction experiments reveal that user involvement features are most predictive of controversy on Reddit, regardless of language.

Our findings contribute to a better understanding of controversy in social media by uncovering its universal properties across languages and by identifying features highly predictive of controversy. As such, our work can inform the development of novel and existing models to advance the automatic detection of controversy, hence improving the civil discourse by supporting moderators to timely intervene in user separating debates and disputes.

2 RELATED WORK

Existing research studied controversy, for example, in the context of weblogs [1, 24], news articles [6, 23, 31, 37], Twitter [15, 16, 19, 26], search engines [9, 21, 36] or Wikipedia [7, 8, 28, 38]. We now briefly recap some of the studies that are most relevant for our work.

Word Usage. Hessel and Lee [17] predicted controversy of comments in six Subreddits by capturing textual content through TFIDF and word2vec models. Mejova et al. [23] used crowdsourcing to manually identify controversial and non-controversial words, which they used to label news articles of 15 major U.S. news outlets. Siersdorfer et al. [31] defined controversy of comments from YouTube and Yahoo! News based on comment ratings (i.e., up- and down-votes) and investigated words occurring in them. Similar to that, we analyze words that are distinctively used in both controversial and non-controversial comments but extend our analysis to six different languages and 50 different Subreddits in total.

Writing Style. Hessel and Lee [17] and Siersdorfer et al. [31] used basic textual features, such as the number of words or sentences as well as readability, to predict and analyze controversy in comments. Jang and Allan [19] further incorporated POS tags to summarize stances in controversial Twitter discussions. We take up and extend such features to study their influence on controversy on Reddit.

Sentiment. Dori and Allan [7] introduced a method to detect the controversy of arbitrary web pages. Authors used sentiment analysis as a baseline, which bares a high recall but performs worse than their proposed method. Mejova et al. [23] found that sentiment and emotions are tempered in controversial news articles. Zielinski et al. [38] detected controversy of Wikipedia articles by considering the sentiment of their respective talk pages. We also investigate and compare sentiment of comments for all six languages.

User Involvement. Ziegele et al. [37] first conducted qualitative interviews with users who comment on news articles and then performed a quantitative analysis of user comments taken from *Spiegel.de* and *Bild.de* (both German news outlets) and their respective Facebook pages. Authors reported that controversy attracts much attention and provokes increased user engagement. Mishne et al. [24] support these findings and uncovered similar behavior for comments posted in various weblogs. In our work, we also study the impact of controversial comments on user involvement on Reddit. Further, we go beyond the listed works and combine user involvement with word usage, writing style and sentiment characteristics to automatically detect controversy in such comments.

3 DATASET AND DESCRIPTIVE ANALYSIS

Dataset. On Reddit, users can express opinions about submissions (i.e., new threads) and comments (i.e., replies to existing threads) from other users by commenting as well as by up- or down-voting them. Each contribution has a score, which aggregates up- (counting as +1) and down-votes (counting as −1). Reddit automatically labels contributions as controversial if the number of up- and down-votes is both high (indicating that many users read the comment) and close to each other, resulting in a score close or equal to 0. The detailed procedure of how Reddit labels controversial contributions is not publicly available. However, the general idea behind this largely reflects the definition of controversy (e.g., from Wiktionary), specifying it as a debate or discussion that involves contrary and opposing views. Additionally, controversial contributions are visible to all users and can be filtered at the top of each comment listing. Thus, Reddit users can specifically search for (or ignore) controversial comments and are aware of comments controversiality.

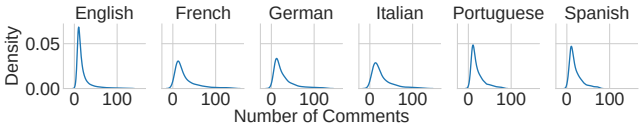
Reddit is structured into Subreddits, each addressing a different topic, such as politics or sports. While the majority of Subreddits addresses English speakers, there are also Subreddits in which users communicate in other languages. Usually, these non-English Subreddits are dedicated to residents of a given country and cover multiple topics at once (e.g., in the German Subreddit r/Austria, Austrians discuss news and sports but also post memes). We analyze 50 different Subreddits, which either address English, French, German, Italian, Portuguese or Spanish speaking users, are thematically different and have varying numbers of users and comments to cover a wide spectrum of possible Subreddits. We use a publicly available dataset[3] including Reddit's controversy labels and extract all comments posted in these Subreddits during the year 2019.

In preprocessing, we remove all hyperlinks, HTML tags, comments posted in submissions with less than ten comments as well as comments containing less than one word (empty and deleted comments). Further, we automatically detect the language of all comments through the Compact Language Detector 2[4] and remove all comments for which the detected language deviates from the expected Subreddit language. As such, we obtain a total of 123, 026, 308 comments for our analysis. In Table 1, we provide detailed statistics of our dataset, including the Subreddits we chose for our analysis. Note that, due to the smaller number of non-English users, the

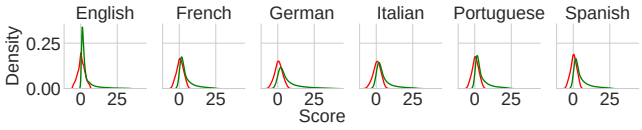
[2] https://github.com/philkon/reddit-controversy

[3] https://files.pushshift.io/reddit

[4] https://github.com/CLD2Owners/cld2; We keep all comments for which the Compact Language Detector 2 only detected one language with at least 90% confidence.



(a) KDE of Number of Comments by Submission



(b) KDE of Scores (green = non-controversial; red = controversial)

Figure 1: Characteristics of Our Dataset. The figure illustrates selected key characteristics of our dataset, including kernel density estimations (KDE) of submission length as well as comment scores, respectively for each language contained in our dataset. We observe that the majority of submissions receive only minimal attention, whereas a small number of submissions receive more comments (cf. Figure 1a; distribution truncated at 120 comments which is still above the 95th percentile). As expected based on the definition of controversy by Reddit, we report higher probabilities for scores around zero for controversial comments (red color) compared to non-controversial comments (green color) for all six languages (cf. Figure 1b; distributions truncated at -5 and 30 which still is below the 5th and above the 95th percentile, respectively).

majority of comments is in English. We selected English Subreddits based on the variety of topics (e.g., news, sports, politics) and Subreddits in other languages based on their activity levels[5].

Preliminary Descriptive Analysis. In Figure 1, we depict selected characteristics of our dataset. Regarding the length of Submissions (i.e., the number of comments posted in a discussion thread), we observe that the majority of submissions on Reddit receive smaller numbers of comments (mean submission length over all Subreddits = 54.42; median submission length over all Subreddits = 14.0), which is independent from language (cf. Figure 1a). Only a small number of submissions receive higher numbers of comments for all six languages, suggesting that not all contributions can attract substantial amounts of attention.

As expected, we observe that controversial comments have scores around zero (cf. Figure 1b; mean score over all Subreddits = 0.60; median score over all Subreddits = 0.00) independent from language. Non-controversial comments across all analyzed languages have a median score of 2.00 and a mean score of 18.72, indicating a rather positive attitude of users on Reddit.

[5]Specifically, we manually checked available Subreddits in respective (non-English) languages and selected those with frequent user activity and at least 1, 000 members.

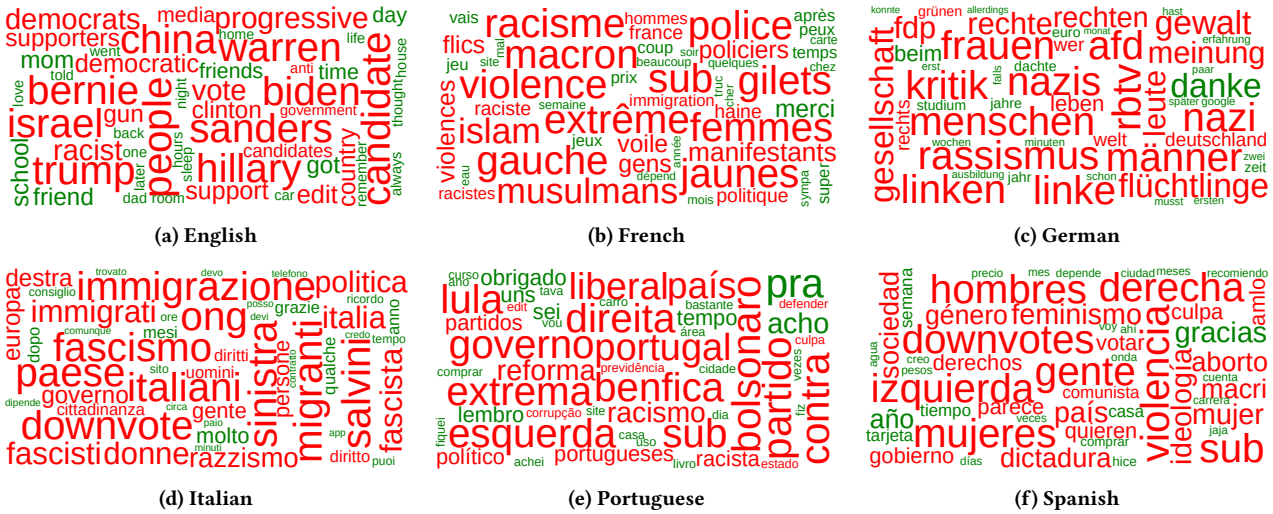
In Table 1, we list the ratios of controversial comments for each of our 50 Subreddits. In general, we observe rather small ratios (ranging between 16.22% and 0.07%) of controversial comments, indicating that most discussions on Reddit do not lead to conflicts or disputes among its users, which again supports our assumption of a general positive mood on the platform.

For English, most controversy can be found in the Subreddit *r/worldnews* (7.84%). We argue that the world affairs discussed in this Subreddit across multiple continents and countries invite a plethora of different views, which increases the probability for conflicts among users. On the contrary, *r/AskReddit*, a Subreddit in which users can pose arbitrary questions to fellow users, has the smallest ratio (0.99%) of controversial comments among English Subreddits. This indicates an open-minded and welcoming community, willing to answer a wide range of user questions.

For other languages, Subreddits that are dedicated to languages, countries or cities are most controversial (e.g., *r/france*, *r/es*). Two exceptions are *r/rocketbeans* for German Subreddits and *r/PrimeriaLiga* for Portuguese Subreddits which are most controversial for respective languages. The former addresses viewers of the German live streaming channel *Rocket Beans TV*, which deals with topics related

Table 1: Dataset Statistics. The table lists an overview of our dataset, including the number of unique users, submissions, comments, controversial comments as well as the Subreddits (numbers in brackets show the ratio of controversial comments in respective Subreddits) we extracted comments from, respectively for each language.

Language	# Users	# Submissions	# Comments	# Controversial	Subreddits
English	5, 225, 561	2, 029, 507	115, 186, 784	3, 176, 622 (2.76%)	r/AskReddit (0.99%), r/facepalm (3.46%), r/funny (3.78%), r/me_irl (2.26%), r/nfl (3.47%), r/philosophy (5.18%), r/politics (4.20%), r/sports (4.50%), r/StarWars (5.42%), r/technology (5.98%), r/todayilearned (4.27%), r/worldnews (7.84%)
French	34, 092	33, 752	1, 431, 249	87, 638 (6.12%)	r/france (6.27%), r/FranceLibre (2.47%), r/jeuxvideo (0.22%), r/montreal (5.71%), r/Quebec (5.47%), r/rance (1.47%)
German	67, 546	49, 707	1, 867, 328	94, 256 (5.05%)	r/Austria (4.87%), r/Dachschaden (4.44%), r/de (5.34%), r/de_IAmA (1.61%), r/Finanzen (2.52%), r/FragReddit (1.58%), r/ich_iel (1.37%), r/rocketbeans (16.22%), r/wasletztepreis (1.67%)
Italian	16, 621	13, 145	669, 072	23, 008 (3.44%)	r/Italia (1.89%), r/italy (3.58%), r/ItalyInformatica (0.65%), r/Libri (0.07%), r/litigi (1.44%)
Portuguese	43, 995	65, 488	1, 765, 419	74, 820 (4.24%)	r/brasil (3.62%), r/BrasildoB (2.90%), r/brasilivre (3.99%), r/circojeca (0.38%), r/portugal (5.75%), r/PORTUGALCARALHO (2.45%), r/PrimeiraLiga (7.10%)
Spanish	57, 713	69, 140	2, 106, 456	76, 394 (3.63%)	r/argentina (3.07%), r/chile (4.13%), r/Colombia (3.60%), r/es (7.94%), r/espanol (0.19%), r/mexico (3.93%), r/podemos (1.83%), r/spain (10.48%), r/uruguay (5.80%), r/vzla (1.65%), r/yo_elvr (0.68%)



(a) English (b) French (c) German

(d) Italian (e) Portuguese (f) Spanish

Figure 2: Word Usage Results. The figure depicts the top 25 words according to χ^2 (represented through word sizes) respectively for controversial (red color) and non-controversial (green color) comments and each of our six languages. We observe differences in discussed topics not only between controversial and non-controversial topics but also between languages. For example, English (cf. Figure 2a) controversial comments focus more on politics and foreign countries, whereas French (cf. Figure 2b), German (cf. Figure 2c) and Italian (cf. Figure 2d) controversial comments address racism and immigration.

to the gaming industry or other issues focusing on a younger audience, potentially involving a more reckless user behavior and, thus, resulting in more controversial comments. In the latter, users specifically discuss the *Primeira Liga*, the highest division of the Portuguese football league system. Here, we argue that the rivalry among fans is clearly reflected in this Subreddit's controversial comments.

To infer the influence of user activity on resulting numbers of controversial comments, we compute Pearson correlation coefficients between the total number of comments and the number of controversial comments in submissions, respectively for each Subreddit. We find significantly positive correlations for all Subreddits (exceptions are *r/Libri*, *r/espanol* and *r/jeuxvideo* for which the correlation is non-significant), but with varying strengths. For example, $\rho = 0.89$ for *r/worldnews* and $\rho = 0.84$ for *r/chile*, but $\rho = 0.27$ for *r/ItalyInformatica* and $\rho = 0.31$ for *r/rance*. This suggests that the ratio of controversial comments is not only impacted by activity levels (e.g., the more attention the more controversial comments), but also depends on topics discussed and their individual communities.

4 EMPIRICAL RESULTS

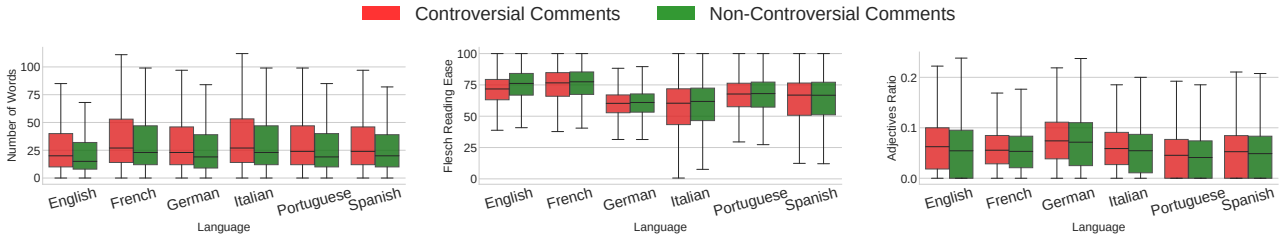
We first investigate word usage and then analyze a set of 23 writing style, sentiment and user involvement features for which we report median and mean differences between distributions of controversial and non-controversial comments and use statistical hypothesis testing to assess whether these differences are significant.

4.1 Word Usage

Motivated by existing research [17, 23, 31], which found that certain words (e.g., "abuse", "killing" or "race") are more related to controversy than others, we analyze word usage differences between controversial and non-controversial comments posted on Reddit. For that, we perform subgroup discovery and adopt the method from Hofland and Johansson [18], which is based on contingency tables and chi-squared (χ^2) tests to assess which words have significantly different distributions in two text corpora. More precisely, for each language we look at the sets of the top 500 words (after removing stop words[6]) included in controversial and non-controversial comments and build the union of those two sets. The union of the top words contains 581 words for English, 565 words for French, 568 words for German, 586 words for Italian, 560 words for Portuguese and 580 words for Spanish. Next, for each language and each word from the respective union we build a 2×2 contingency table, which keeps the count of a given word, as well as the total count of all other words in both controversial and non-controversial comments. The null hypothesis of the χ^2 test (which we perform with Yates Correction [35]) states that the occurrence of a given word is independent of the controversy of the comment. Hence, words for which we can reject this null hypothesis are used distinctively in either controversial or non-controversial comments.

Results. In Figure 2, we depict the top 25 significant words with regard to their χ^2 values and to their relative frequencies in controversial and non-controversial comments (to decide where their usage is significantly higher; all p-values < 0.0005), respectively for each of the six languages. We observe that, independent from language, top controversial words reflect topics that have been discussed in broader public during 2019, whereas non-controversial words cannot be related to topics that easily. These findings are similar to those from previous studies [17, 23, 31]. For example, English top controversial words include political persons, such as

[6]Link to stop words lists: https://github.com/Alir3z4/python-stop-words



(a) Number of Words **(b) Flesch Reading Ease** **(c) Ratio of Adjectives**

Figure 3: Writing Style Results. The figure depicts box plots for selected writing style features comparing non-controversial (green color) and controversial (red color) comments for all six languages. Horizontal black lines indicate the median and the first and third quartile. Whiskers indicate minimum and maximum values still within 1.5 interquartile ranges. Note that we do not depict outliers for better representation of data and that these characteristics apply to remaining box plots in this paper as well. We report that controversial comments have more words (a), are harder to read (b) and contain more adjectives (c) compared to non-controversial comments in any of the six languages.

Bernie Sanders and *Donald Trump*, but also the foreign countries *China* and *Israel*. Top controversial words of other languages relate to racism, fascism, immigration, religion or gender equality.

We find that top words largely reflect local issues of respective countries. For example, English top words, to a large extent, address U.S. politics as well as foreign countries in the focus thereof. Note that *r/politics* and *r/worldnews* have the highest absolute numbers of English controversial comments (845, 089 and 682, 402 respectively), explaining the majority of top words being related to politics. In the case of French, controversial comments address *Islam* and *Moslems*, which may be caused by the lengthy series of terrorist attacks. For German top controversial words, we observe that an increasing rightward shift of society is reason for disputes. Similarly, Italians discuss rightwing politics as well as migration, which could be due to the ongoing refugee crisis. Portuguese top words focus on politics in Portugal as well Brazil, but also include *Benfica*, which stems from the Portuguese football club *Sport Lisboa e Benfica*, suggesting strong rivalry among sport fans in Portugal. Spanish controversial top words address gender equality as well as abortion, which has very restrictive laws in Latin America [14].

4.2 Writing Style

We now set our focus on how writing style of comments on Reddit reflects controversy. For that, we follow existing research [17, 19, 31] and consider differences in text length, readability and POS tags between controversial and non-controversial comments.

Statistical Hypothesis Tests. To assert whether the difference between the two types of comments is significant, we use statistical hypothesis testing. In particular, our null hypothesis assumes equal distributions for controversial and non-controversial comments. Thus, we first perform the Brown-Forsythe test (at a significance level $\alpha = 0.05$) to assess the equality of variances between distributions. Based on these results, we then use the median test in cases of unequal variances, and in case of equal variances we use the Mann-Whitney-Wilcoxon test, which has a higher statistical power (cf. [10]). We select these tests as they make no assumptions about the underlying distributions (manual inspection of kernel density estimation plots revealed many different shapes). To counteract the problem of multiple comparisons, we perform the Bonferroni correction [4] reducing the commonly used significance level $\alpha = 0.05$ for the entire set of n comparisons to $\frac{\alpha}{n}$. In our work, we test 23 features (including features of subsequent sections), hence, $n = 23$ and $\alpha \approx 2.17 \times 10^{-3}$. Note that we provide a detailed overview of hypothesis test results as well as differences in distribution medians and means for all features and languages in Table 2.

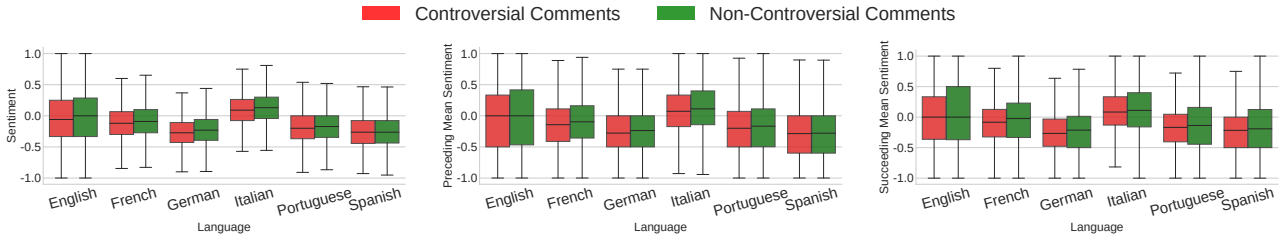
Text Length. For each comment we extract the *number of characters*, the *number of syllables*, *the number of words* and the *number of sentences*. Further, we investigate the *characters to sentences ratio* and the *words to sentences ratio*.

Readability. We determine the readability of comments by computing the *Flesch Reading Ease* [12]. Since the original formula is intended for English only, we use its respective derivatives for French [20], German [3], Italian [13], Portuguese [2] and Spanish [11] comments. These measurements are comparable and represent the reading difficulty of a text by a score ranging between 0 and 100, where higher values indicate easier to read texts and lower values indicate harder to read texts. Note that in this case we limit our dataset to comments with a minimum of 100 words as the readability formulas might return inaccurate values otherwise. Hence, for readability, we analyze 5, 363, 089 English, 122, 204 French, 120, 578 German, 54, 832 Italian, 121, 549 Portuguese as well as 136, 772 Spanish comments.

POS Tags. Using Spacy's POS tagger[7], we extract the *ratio of nouns*, the *ratio of verbs*, the *ratio of adjectives*, the *ratio of adverbs* and the *ratio of pronouns* of each comment.

Results. We illustrate distributions of selected writing style features for controversial and non-controversial comments as well as each language in Figure 3. Starting with comment text length, we find that controversial comments are significantly longer across all languages (cf. Figure 3a). The number of characters, syllables and sentences positively correlate with the number of words (all Pearson $\rho > 0.754$ with p-values < 0.0005), further strengthening this observation. Similarly to comment length, controversial comments of all languages have significantly higher characters to sentences and words to sentences ratios.

[7] https://spacy.io (version used: 2.2.3)



(a) Sentiment **(b) Preceding Mean Sentiment** **(c) Succeeding Mean Sentiment**

Figure 4: Sentiment Results. The figure depicts box plots for our sentiment features comparing non-controversial (green color) and controversial (red color) comments for all six languages. We find that, independent form language, sentiment is significantly more negative in discussion evolving around controversial comments than in those around non-controversial ones.

In Figure 3b, we illustrate distributions of the Flesch Reading Ease for controversial and non-controversial comments, respectively for each language. Overall, we find that comments on Reddit are between fairly difficult and fairly easy to read. According to median differences, controversial comments are harder to read as compared to non-controversial comments for all languages. The median difference is significant for English, French as well as German (all p-values < our corrected α), but non-significant for Italian (p-value = 0.007), Portuguese (p-value = 0.096) and Spanish (p-value = 0.168). These results in combination with text length features suggest a more complex content for controversial comments.

Finally, we report significantly higher ratios of nouns (except for English with a p-value of 0.004) and adjectives (cf. Figure 3c) in controversial comments across all languages. Regarding the ratio of verbs, we observe significantly lower ratios for English, Portuguese and Spanish controversial comments as well as significantly higher ratios for German and Italian controversial comments. For French comments, there is no significant difference (p-value = 0.208) in the ratio of verbs. The ratio of adverbs is significantly higher for English, German and Spanish controversial comments, whereas it is not significantly higher for French (p-value = 0.016), Italian (p-value = 0.346) and Portuguese (p-value = 0.211) comments. The ratio of pronouns is significantly lower in English and French controversial comments, while it is significantly higher in German and Portuguese controversial comments. We observe no significant differences in the ratio of pronouns for Italian (p-value = 0.004) and Spanish (p-value = 0.012) comments. Overall, our POS tag features imply that controversial comments are written more impersonally than non-controversial comments.

4.3 Sentiment

According to previous research [7, 23, 38], sentiment is predictive of controversy. To compute the sentiment of comments posted on Reddit, we rely on existing sentiment dictionaries that have already been created[8] and evaluated for our six languages in existing work [5]. Using the respective sentiment dictionaries for our languages, we compute the sentiment s of each comment with $s = (W_p - W_n)/(W_p + W_n)$, where W_p is the number of positive words in a comment and W_n is the number of negative words in a comment. Hence, s ranges between -1 and $+1$, where values close to -1 are considered as negative, values close to $+1$ as positive, and values close to zero as neutral sentiment.

Besides computing the sentiment for each comment, we also investigate the *preceding mean sentiment* (i.e., the mean sentiment over preceding comments) and the *succeeding mean sentiment* (i.e., the mean sentiment over succeeding comments).

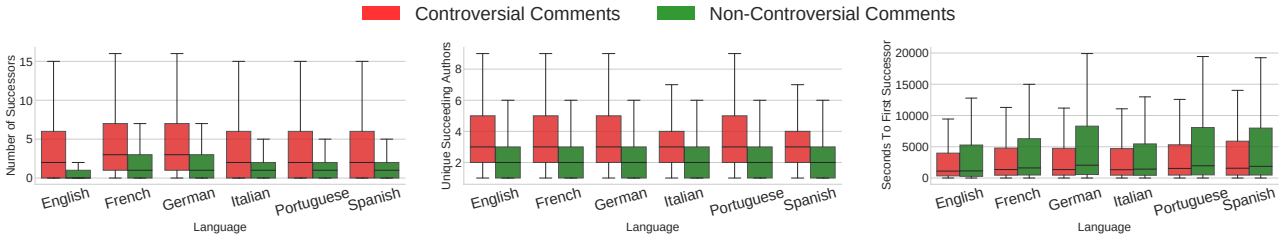
Similar to writing style features (cf. Section 4.2), we only consider comments with at least 100 words for the analysis of sentiment to prevent inaccurate values. Further, we use the same hypothesis testing approach to assess the significance of differences between controversial and non-controversial comments.

Results. Figure 4 depicts the distributions for controversial and non-controversial comments and each language. In general, we observe that the median sentiment of comments is rather negative for all languages, except for Italian, where the median sentiment is slightly positive. However, controversial comments have a significantly more negative sentiment compared to non-controversial comments for all languages, except Spanish (p-value = 0.327). The preceding mean sentiment (cf. Figure 4b) suggests that comments preceding controversial ones are also significantly more negative than those preceding non-controversial ones. This difference is not significant for Spanish comments (p-value = 0.188). The succeeding mean sentiment (cf. Figure 4c) is significantly more negative for controversial comments in all languages, indicating a general negative attitude for discussion threads with controversial comments.

4.4 User Involvement

Previous studies [24, 37] have shown that controversial topics attract a lot of attention in online discussions. Hence, we investigate whether users of our Subreddits exhibit a similar behavior and compute eight features capturing structural and temporal aspects of discussion threads. We inspect the *number of predecessors* (i.e., the number of preceding comments), the *number of successors* (i.e., the number of succeeding comments), the *number of unique preceding users* as well as the *number of unique succeeding users*. Analogously, we analyze the *time from predecessor*, the *mean time between predecessors* (i.e., the mean time between all preceding comment), the

[8] Authors extracted most frequent words of Wikipedia articles and created a knowledge graph to combine similar words of different languages through Wiktionary, machine translation (via Google translate), transliteration links and WordNet. Starting with sentiment of English vertices based on an existing dictionary, authors propagated sentiment to vertices of other languages and created dictionaries for 136 languages.



(a) Number of Successors **(b) Unique Succeeding Users** **(c) Time to First Successor**

Figure 5: User Involvement Results. The figure depicts box plots for selected user involvement features comparing non-controversial (green color) and controversial (red color) comments for all six languages. We observe that controversial comments attract more attention (a and b) more quickly (c) compared to non-controversial comments, regardless of language.

time to the first successor and the *mean time between successors* (i.e., the mean time between all succeeding comments), providing insights into how fast comments attract attention. Note that we measure time in seconds for all temporal features.

We use the same hypothesis testing approach described in Section 4.2 to infer whether differences between controversial and non-controversial comments are significant.

Results. We depict distributions for selected user involvement features in Figure 5. The number of predecessors is significantly lower for controversial comments in any language (except English for which it is significantly higher), indicating that such comments are posted closer to the original submission of the discussion. Similarly, the number of unique preceding users is significantly lower for controversial comments in all languages except German (p-value = 0.025), Spanish (p-value = 0.149) and English (here the number is significantly higher). Controversial comments receive more attention as the number of successors and the number unique succeeding users is significantly higher for them across all languages.

The time from predecessor suggests that controversial comments in any language are posted significantly faster to preceding comments than non-controversial ones. Similar to that, the mean time between predecessors is significantly lower for controversial comments across languages. The time to the first successor (cf. Figure 5c) indicates that controversial comments are attracting attention significantly faster. Further, the mean time between successors is significantly lower for controversial comments, except for French (p-value = 0.128) and Italian (p-value = 0.023).

Table 2: Distribution Differences and Hypothesis Test Results. The table lists the median and mean (in brackets) differences between controversial and non-controversial distributions for each of our 23 features and each of the six languages. Depending on the differences in variances between non-controversial and controversial comment distributions, we either use the median test or the Mann-Whitney-Wilcoxon test (indicated with colored cells). Underlined values indicate significance (p-values of hypothesis tests < our Bonferroni corrected $\alpha \approx 2.17 \times 10^{-3}$). We can reject our null hypothesis for 122 out of 138 tests.

Feature	English	French	German	Italian	Portuguese	Spanish
Number of Characters	29.0 (35.764)	25.0 (20.529)	29.0 (28.024)	26.0 (28.142)	26.0 (28.53)	24.0 (18.076)
Number of Syllables	7.0 (8.712)	5.0 (4.946)	8.0 (7.579)	8.0 (9.494)	8.0 (9.376)	8.0 (5.873)
Number of Words	5.0 (5.431)	4.0 (3.094)	4.0 (3.932)	4.0 (4.266)	5.0 (4.638)	4.0 (3.022)
Number of Sentences	0.0 (0.306)	0.0 (0.097)	0.0 (0.199)	0.0 (0.131)	0.0 (0.269)	0.0 (0.071)
Characters to Sentences Ratio	7.455 (7.699)	6.0 (8.304)	7.0 (6.994)	7.667 (10.337)	5.75 (5.746)	9.0 (10.37)
Words to Sentences Ratio	1.0 (1.068)	1.0 (1.223)	0.75 (0.936)	1.0 (1.504)	0.929 (0.864)	1.571 (1.578)
Flesch Reading Ease	−4.41 (−4.116)	−0.862 (−1.476)	−0.675 (−0.743)	−1.409 (−1.696)	−0.329 (0.085)	0.112 (−0.309)
Ratio of Nouns	0.0 (−0.005)	0.003 (0.002)	0.002 (0.0)	0.004 (0.003)	0.0 (0.002)	0.003 (0.003)
Ratio of Verbs	0.0 (−0.001)	0.001 (0.0)	0.002 (0.001)	0.002 (0.002)	−0.002 (−0.002)	−0.004 (−0.003)
Ratio of Adjectives	0.008 (0.005)	0.002 (0.002)	0.003 (0.003)	0.004 (0.003)	0.004 (0.003)	0.004 (0.003)
Ratio of Adverbs	0.002 (−0.0)	0.0 (0.0)	−0.001 (−0.001)	−0.0 (−0.001)	−0.001 (−0.001)	0.002 (0.0)
Ratio of Pronouns	−0.006 (−0.006)	−0.004 (−0.003)	0.002 (0.001)	0.001 (0.001)	0.001 (−0.001)	0.0 (−0.001)
Sentiment	−0.059 (−0.031)	−0.029 (−0.033)	−0.042 (−0.043)	−0.04 (−0.04)	−0.028 (−0.023)	0.0 (−0.002)
Preceding Mean Sentiment	0.0 (−0.03)	−0.046 (−0.039)	−0.041 (−0.041)	−0.038 (−0.041)	−0.033 (−0.026)	−0.006 (−0.002)
Succeeding Mean Sentiment	0.0 (−0.078)	−0.06 (−0.042)	−0.054 (−0.059)	−0.028 (−0.027)	−0.032 (−0.038)	−0.025 (−0.04)
Number of Predecessors	1.0 (0.398)	0.0 (−0.473)	0.0 (−0.274)	0.0 (−0.288)	0.0 (−0.447)	0.0 (−0.205)
Number of Successors	2.0 (3.413)	2.0 (2.978)	2.0 (3.018)	1.0 (2.507)	1.0 (2.912)	1.0 (2.429)
Unique Preceding Users	1.0 (0.485)	−1.0 (−0.194)	0.0 (−0.086)	0.0 (−0.106)	0.0 (−0.068)	0.0 (−0.048)
Unique Succeeding Users	1.0 (0.328)	1.0 (0.972)	1.0 (0.86)	1.0 (0.845)	1.0 (1.06)	1.0 (0.835)
Time From Predecessor	−2768.0 (−17558.083)	−1677.0 (−10135.157)	−2602.0 (−14428.74)	−887.0 (−10041.005)	−2116.5 (−13544.689)	−2084.0 (−14370.263)
Time To First Successor	−30.0 (−7102.537)	−260.0 (−3316.991)	−697.0 (−6166.192)	−97.5 (−2917.029)	−426.0 (−5341.833)	−303.0 (−6640.838)
Mean Time Bet. Predecessors	−454.625 (−3028.215)	−2010.875 (−4830.73)	−1465.917 (−4552.319)	−4522.5 (−8133.433)	−1430.9 (−6668.522)	−2483.5 (−8757.049)
Mean Time Bet. Successors	333.486 (−6697.816)	−37.4 (−3719.733)	−911.786 (−6424.904)	107.635 (−3577.459)	−323.429 (−5278.054)	−168.264 (−5918.054)

Figure 6: Feature Importances. The figure illustrates mean feature importances over the ten-fold cross-validation. Our model achieves a mean ROC AUC of 0.79. Note that, for visualization purposes, we only show features with an importance of at least 0.01. The error bars indicate 95% bootstrap confidence intervals. We observe that involvement features (pink color) are most predictive, while other features, such as writing style (orange color), contribute less to the prediction of controversy on Reddit.

4.5 Summary of Empirical Results

For word usage, we find that controversial comments often include words related to topics frequently addressed in the public discourse of respective language, such as the refugee crisis in Italy or abortion in Latin America. On the other hand, non-controversial comments in all languages include more moderate words or words that cannot clearly be related to a topic, such as "time" or "friend".

For writing style, sentiment and user involvement features, we find significant differences for 122 out of 138 cases according to our hypothesis tests. We list median and mean differences for each feature and language in Table 2. Our results for writing style and sentiment differences suggest that controversial comments are significantly longer, harder to read, more impersonal and more negative than non-controversial comments. Our user involvement analysis reveals that controversial comments attract more user attention than non-controversial comments and that users are quicker to follow up on a controversial comment.

Overall, our results on writing style, sentiment and user involvement are similar across languages, indicating that discussions on controversial issues on Reddit follow common modalities shared between languages. However our word usage analysis suggests a different relevance of topics across languages and countries.

5 PREDICTING CONTROVERSY

5.1 Experimental Setup

Based on our empirical results, we now conduct a prediction task and investigate the predictive power of various features for controversy on Reddit. Note that the aim of this prediction task is not to achieve best performance, but rather to investigate what features are most predictive of controversy. As such, our proposed method needs to be easily interpretable and may not reach the performance of more sophisticated approaches, such as neural networks.

Features. We use all features from the previous word usage (cf. Section 4.1), writing style (cf. Section 4.2), sentiment (cf. Section 4.3) and user involvement (cf. Section 4.4) analyses. In the case of word usage features, we count the number of the top 25 controversial and non-controversial words in comments, respectively for each language, and report these features as the *number of controversial words* and the *number of non-controversial words*. Additionally, we include the language of comments as well as the Subreddit they had been posted in. Note that we use one-hot-encoding to transform categorical features (language and Subreddit) and that we apply robust scaling (due to the presence of outliers) to numerical features.

Prediction Samples. We remove all comments that have less than 100 words in order to exclude unreliable textual features. This leaves us with a total of 5, 919, 024 comments (218, 053 controversial and 5, 700, 971 non-controversial). To address the unbalance between classes, we perform random undersampling and finally obtain 436, 106 comments (376, 302 English; 16, 796 French; 14, 092 German; 4, 540 Italian; 12, 204 Portuguese; 12, 172 Spanish).

Model and Evaluation. We employ *Gradient Boosted Decision Trees* (GBDTs) as implemented in scikit-learn[9]. We tune hyper parameters of the GBDTs through a grid search[10] and evaluate our model by using ten-fold cross-validation for which we report mean ROC AUC values over the ten cross-validation folds.

5.2 Prediction Results

We report a mean ROC AUC of 0.79, indicating that we can achieve moderate prediction performance and improve a random baseline of 0.50 by 0.29. In Figure 6, we depict the importance of selected features (we exclude all features with importance smaller than 0.01) to assess which of the previously analyzed features are most predictive of controversy. Here, we find involvement features to carry most predictive strengths. In particular, the number of unique succeeding users (0.32), the number of predecessors (0.18), the mean time between predecessors (0.09), the time from predecessor (0.06), the Subreddit *r/AskReddit* (0.06) as well as the number of controversial words (0.05) are most predictive. Other features have importance values equal or smaller than 0.03.

These results suggest that the more users participate in a discussion, the likelier it is to include a controversial comment. Other aspects of comments, such as the words it contains or the sentiment it conveys, are less important when predicting controversy on Reddit. Further, we see that languages have no or only minimal influence on prediction performance, suggesting that controversy on Reddit

[9] https://scikit-learn.org/0.23 (version used: 0.23.2)
[10] We include parameters for the best model in our GitHub repository.

behaves similarly across languages. Interestingly, only three Subreddits are at least to some extent predictive of controversy. These include *r/AskReddit*, *r/politics* and *r/worldnews*. Combining these findings with ratios of controversial comments per Subreddit (cf. Table 1), we argue that it is either particularly unlikely (*r/AskReddit*) or particularly likely (*r/politics*, *worldnews*) for a comment to be controversial in these Subreddits.

6 DISCUSSION

We now connect our results to our research question and compare our findings to those from existing controversy studies [7, 17, 19, 23, 24, 31, 37, 38] conducted in other contexts.

Word Usage. We found that controversial comments often include terms that are frequently discussed in the public discourse, such as "abortion", "society" or "politics". Contrary, non-controversial comments include terms, such as "friend" or "mother", that cannot be related to a specific topic that easily. In particular, we find that our extracted top words coincide with controversial and non-controversial words manually extracted by crowdworkers from news articles provided by NewsCred [23]. We observe that our extracted controversial and non-controversial top words are respectively related to words frequently found in accepted and not accepted comments posted on YouTube or Yahoo! News [31]. This suggests that controversy on Reddit behaves similarly to controversy in other contexts. Further, we found that language communities on Reddit not only discuss topics related to their respective countries, but also global issues, implying a possible domain transfer across cultures. This is different when studying Subreddits in a micro-perspective, where text features, such as TFIDF or word embeddings, are very community specific [17].

Writing Style. We observed that controversial comments are longer, harder to read and more impersonal compared to non-controversial comments. We argue that this is due to an overall more complex writing style used in controversial comments to persuade or deceive opinions of other users. Similar findings were obtained by Tan et al. [33], where authors analyzed *r/ChangeMyView*, a Subreddit specifically dedicated to the understanding of contrasting views.

Sentiment. We saw that controversial comments convey a more negative sentiment than non-controversial comments. Further, comments preceding and succeeding a controversial comment also have more negative sentiment, suggesting that discussion threads with controversial comments have a general negative mood. Our findings are again similar to existing controversy studies in the context of edit wars on Wikipedia [38] or online news articles [23].

In future work, we are interested in further analyzing the particular negative sentiment associated with controversial comments. For example, we would also expect controversial comments in which their authors write positive about a controversial topic (e.g., a user is for abortion and not against it) and, thus, convey a positive sentiment. Also, the investigation of a potential bias introduced by sentiment dictionaries, in which terms related to controversies are generally labelled as negative, can be promising for future work.

User Involvement. Similarly to the existing studies [24, 37], we found that users on Reddit feel particularly attracted to controversial comments. In our data, we observed correlations between the temporal and structural user involvement features. For example, there are weak negative correlations between the time to first successor and the number of successors (Spearman $\rho = -0.164$; p-value < 0.0005), suggesting that the faster a controversial comment gets its first reply, the more total replies it receives. However, this result may be somewhat obfuscated due to Reddit's way of displaying comments (i.e., the easy reachability of lower discussion tree levels), where deeper levels of a discussion tree do not attract as much attention as the lower levels due to the position bias in users perception of Web pages [22]. Further, additional clicks are required to expand discussion branches, requiring additional effort from users to see comments further along the discussion trees.

Predicting Controversy. User involvement features discriminate the most between controversial and non-controversial comments. Especially the number of preceding comments and the number of unique succeeding authors indicate the influence of increased exposure (through commenting right from the beginning of a discussion) on comment controversy. This also indicates that comments posted at a later time of a discussion might get lost in the crowd. Overall, we suggest to consider word usage, writing style, sentiment and Subreddits next to user involvement features when predicting controversy, as they add useful information to controversy prediction, similar to what has been shown in existing studies [7, 17, 38]. To assess the performance gain of these features in our context, we redo our prediction experiment (cf. Section 5.1), but this time only include user involvement, Subreddit as well as language features. We achieve a ROC AUC of 0.77 and, thus, observe a gain of 0.02 when including word usage, writing style and sentiment features.

Multilingual Controversy. The insignificant importance of languages in our prediction model further confirms that controversy on Reddit is language agnostic. However, we are aware that this might be due the unbalanced distribution of languages in our prediction samples (the majority of comments is in English). Thus, to control for the number of comments per language we perform an additional prediction experiment. In particular, we randomly draw 2,270 (this is the number of Italian controversial samples and also the minimum across languages) controversial comments for each language for which we then repeat the experiment as described in Section 5.1. With this prediction experiment we confirm our previous findings as we again report that language is not important (all importance values < 0.01) and user involvement features are once more most predictive. Overall, this indicates that existing controversy detection models can be transferred to other languages, as long as they incorporate features similar to our user involvement features and are not based on words specific to a certain language.

Limitations. The scope of our work is limited to 50 Subreddits with comments posted during the year 2019. Thus, we observe a relatively short time period and also only a small fraction of all Subreddits hosted by Reddit. However, we argue that our results are representative for Reddit, as our analysis included some of the largest and most active Subreddits in their respective languages. Further, our work investigates controversy as defined by Reddit, which is different from other definitions based, for example, on Twitter [15, 16, 26, 27]. We leave a comparison of different controversy definitions to future work. Also, we want to note that we analyzed correlations and not causality. For example, a comment receiving much attention may not necessarily be controversial as it

can also be popular due to other reasons. We leave the study of the causality to future work. Finally, we did not consider sarcasm and its influence on sentiments expressed in comments, a separate and non-trivial problem, which is out of scope for this paper.

7 CONCLUSION

In this paper we analyzed word usage, writing style, sentiment and user involvement in the context of controversial and non-controversial comments posted in six languages on Reddit. We performed subgroup discovery and computed a set of 23 features for comments posted in 50 different Subreddits and used statistical hypothesis testing to infer differences between controversial and non-controversial comments. Most notably, we found that users are more engaged and that they write more complex and negative comments when debating controversial issues. We observed this behavior for all languages, suggesting that controversy on Reddit is universal across languages, except for the fact that languages reflect local issues of respective countries. Further, we demonstrated that our analyzed and described features are predictive of controversy on Reddit, with user involvement features having the most predictive strengths independent from language. Our approach enables moderators to timely intervene if required or to simply automatically flag submissions that are in danger of derailing conversations.

For future work, we want to further extend our analysis to other datasets. We also plan to characterize users involved in controversial discussions. Moreover, we intend to experiment with early prediction of controversial comments using machine learning.

ACKNOWLEDGMENTS

Parts of this work were funded by the go!digital Next Generation program of the Austrian Academy of Sciences. Supported by the Graz University of Technology Open Access Publishing Fund.

REFERENCES

[1] Lada A Adamic and Natalie Glance. 2005. The political blogosphere and the 2004 US election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*. ACM, 36–43.
[2] Patricia Estefania Ayala Aguirre, Melina Martins Coelho, Daniela Rios, Maria Aparecida Andrade Moreira Machado, Agnes Fátima Pereira Cruvinel, and Thiago Cruvinel. 2017. Evaluating the dental caries-related information on Brazilian websites: qualitative study. *Journal of medical Internet research* 19, 12 (2017).
[3] T Amstad. 1978. Wie verständlich sind unsere Zeitungen?(unveröffentlichte Dissertation). *Zürich: Universität* (1978).
[4] Carlo E Bonferroni. 1935. Il calcolo delle assicurazioni su gruppi di teste. *Studi in onore del professore salvatore ortu carboni* (1935), 13–60.
[5] Yanqing Chen and Steven Skiena. 2014. Building sentiment lexicons for all major languages. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. ACL, Baltimore, Maryland, 383–389.
[6] Yoonjung Choi, Yuchul Jung, and Sung-Hyon Myaeng. 2010. Identifying controversial issues and their sub-topics in news articles. *Intelligence and Security Informatics* (2010), 140–153.
[7] Shiri Dori-Hacohen and James Allan. 2013. Detecting controversy on the web. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*. ACM, 1845–1848.
[8] Shiri Dori-Hacohen and James Allan. 2015. Automated controversy detection on the web. In *European Conference on Information Retrieval*. Springer, 423–434.
[9] Shiri Dori-Hacohen, Elad Yom-Tov, and James Allan. 2015. Navigating Controversy as a Complex Search Task.. In *SCST@ ECIR*.
[10] Nick Feltovich. 2003. Nonparametric tests of differences in medians: comparison of the Wilcoxon–Mann–Whitney and robust rank-order tests. *Experimental Economics* 6, 3 (2003), 273–297.
[11] José Fernández Huerta. 1959. Medidas sencillas de lecturabilidad. *Consigna* 214 (1959), 29–32.
[12] Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology* 32, 3 (1948), 221.
[13] Valerio Franchina and Roberto Vacca. 1986. Adaptation of Flesh readability index on a bilingual text written by the same author both in Italian and English languages. *Linguaggi* 3 (1986), 47–49.
[14] Beatriz Galli. 2020. Challenges and opportunities for access to legal and safe abortion in Latin America based on the scenarios in Brazil, Argentina, and Uruguay. *Cadernos de Saúde Pública* 36 (2020), e00168419.
[15] Anatoliy Gruzd and Jeffrey Roy. 2014. Investigating political polarization on Twitter: A Canadian perspective. *Policy & Internet* 6, 1 (2014), 28–45.
[16] Pedro Henrique Calais Guerra, Wagner Meira Jr, Claire Cardie, and Robert Kleinberg. 2013. A Measure of Polarization on Social Media Networks Based on Community Boundaries.. In *ICWSM*.
[17] Jack Hessel and Lillian Lee. 2019. Something's Brewing! Early Prediction of Controversy-causing Posts from Discussion Features. *arXiv preprint arXiv:1904.07372* (2019).
[18] Knut Hofland and Stig Johansson. 1982. *Word frequencies in british and american english*. Norwegian computing centre for the Humanities.
[19] Myungha Jang and James Allan. 2018. Explaining controversy on social media via stance summarization. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 1221–1224.
[20] Liliane Kandel and Abraham Moles. 1958. Application de l'indice de Flesch à la langue française. *Cahiers Etudes de Radio-Télévision* 19, 1958 (1958), 253–274.
[21] Danai Koutra, Paul N Bennett, and Eric Horvitz. 2015. Events and controversies: Influences of a shocking news event on information seeking. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 614–624.
[22] Daniel Lamprecht, Kristina Lerman, Denis Helic, and Markus Strohmaier. 2017. How the structure of wikipedia articles influences user navigation. *New Review of Hypermedia and Multimedia* 23, 1 (2017), 29–50.
[23] Yelena Mejova, Amy X Zhang, Nicholas Diakopoulos, and Carlos Castillo. 2014. Controversy and sentiment in online news. *arXiv preprint arXiv:1409.8152* (2014).
[24] Gilad Mishne, Natalie Glance, et al. 2006. Leave a reply: An analysis of weblog comments. In *Third annual workshop on the Weblogging ecosystem*. Edinburgh, Scotland.
[25] Amita Misra and Marilyn A Walker. [n.d.]. Topic Independent Identification of Agreement and Disagreement in Social Media Dialogue. *Evolution* 460, 460 ([n. d.]), 920.
[26] AJ Morales, Javier Borondo, Juan Carlos Losada, and Rosa M Benito. 2015. Measuring political polarization: Twitter shows the two sides of Venezuela. *Chaos: An Interdisciplinary Journal of Nonlinear Science* 25, 3 (2015), 033114.
[27] Marco Pennacchiotti and Ana-Maria Popescu. 2010. Detecting controversies in Twitter: a first study. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*. ACL, 31–32.
[28] Hoda Sepehri Rad and Denilson Barbosa. 2012. Identifying controversial articles in Wikipedia: A comparative study. In *Proceedings of the eighth annual international symposium on wikis and open collaboration*. ACM, 7.
[29] Marlene Schommer-Aikins and Rosetta Hutter. 2002. Epistemological beliefs and thinking about everyday controversial issues. *The journal of Psychology* 136, 1 (2002), 5–20.
[30] Feng Shi, Misha Teplitskiy, Eamon Duede, and James A Evans. 2019. The wisdom of polarized crowds. *Nature Human Behaviour* (2019), 1.
[31] Stefan Siersdorfer, Sergiu Chelaru, Jose San Pedro, Ismail Sengor Altingovde, and Wolfgang Nejdl. 2014. Analyzing and mining comments and comment ratings on the social web. *ACM Transactions on the Web (TWEB)* 8, 3 (2014), 1–39.
[32] Róbert Sumi, Taha Yasseri, et al. 2011. Edit wars in Wikipedia. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*. IEEE, 724–727.
[33] Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web*. 613–624.
[34] VG Vydiswaran, ChengXiang Zhai, Dan Roth, and Peter Pirolli. 2012. BiasTrust: Teaching biased users about controversial topics. In *Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM, 1905–1909.
[35] Frank Yates. 1934. Contingency tables involving small numbers and the χ 2 test. *Supplement to the Journal of the Royal Statistical Society* 1, 2 (1934), 217–235.
[36] Elad Yom-Tov, Susan Dumais, and Qi Guo. 2014. Promoting civil discourse through search engine diversity. *Social Science Computer Review* 32, 2 (2014), 145–154.
[37] Marc Ziegele, Timo Breiner, and Oliver Quiring. 2014. What creates interactivity in online news discussions? An exploratory analysis of discussion factors in user comments on news items. *Journal of Communication* 64, 6 (2014), 1111–1138.
[38] Kazimierz Zielinski, Radoslaw Nielek, Adam Wierzbicki, and Adam Jatowt. 2018. Computing controversy: Formal model and algorithms for detecting controversy on Wikipedia and in search queries. *Information Processing & Management* 54, 1 (2018), 14–36.