Classification of German Newspaper Comments

Christian Godde and Konstantina Lazaridou and Ralf Krestel

Hasso-Plattner-Institut, Potsdam, Germany
Christian.Godde@student.hpi.uni-potsdam.de,
konstantina.lazaridou@hpi.de,
ralf.krestel@hpi.de

Abstract. Online news has gradually become an inherent part of many people's every day life, with the media enabling a social and interactive consumption of news as well. Readers openly express their perspectives and emotions for a current event by commenting news articles. They also form online communities and interact with each other by replying to other users' comments. Due to their active and significant role in the diffusion of information, automatically gaining insights of these comments' content is an interesting task. We are especially interested in finding systematic differences among the user comments from different newspapers. To this end, we propose the following classification task: Given a news comment thread of a particular article, identify the newspaper it comes from. Our corpus consists of six well-known German newspapers and their comments. We propose two experimental settings using SVM classifiers build on comment- and article-based features. We achieve precision of up to 90% for individual newspapers.

Keywords: media analysis, news comment analysis, comment classification

1 Introduction

Many online news sites offer their readers the possiblity to comment on news articles either directly below the article in a forum-style way, or via Twitter or Facebook. While the latter is more suitable for sharing news, the former is more appropriate for discussion of the articles' contents. These online comments are huge reservoirs of user generated content with readers expressing opinions on various news-related topics. These range from comments on the article's style, specific arguments of the article, to general opinions about greater questions.

Not only does the discussion in these sections often reflect the readers' opinions about the article itself, but also about the overall topic and beyond, with readers referring to each other or introducing new arguments. Figure 1 shows excerpts of an article together with a comment and a reply to this comment. In general, the discussions are not limited to the specific article's topic and often introduce new arguments and opinions. Sentiments are expressed as well, towards either the content of the article or statements of other users. The content of one individual comment is not easily machine-understandable. It needs to be

evaluated in the context of the surrounding thread and associated article. Nevertheless, we argue that discussion style and topics may differ between various news providers, depending on their respective audience and possibly bias in the article's coverage. For example, German newspapers and the majority of their readers are traditionally associated with a certain political alignment. If this is true, the political leaning should be reflected in the comment sections of the respective news sites as well. Even if the bias in the articles themselves is minimal, the reaction of the readers to the covered event may be much more diverse, which in return could be used to infer arguments for the political alignment of the news sites.

(a) Excerpt of article

(b) Excerpt of comment

(c) Excerpt of reply

Fig. 1: Example of an article, comment, and reply from "Zeit"

In this paper we analyze the user comments on six major German news sites regarding their differences in discussion focus, language and sentiment. Based on the assumption that user comments on various news sites differ in these characteristics, we propose a classifier to predict the source of specific comments, that is, the news site on which the comments have been posted. To analyze this, a prediction method is developed and evaluated, which, given a set of user comments, predicts the originating news site.

2 Related Work

User comments can be found in different online platforms and communities. Social media plattforms, such as Twitter, Facebook, and Youtube, are the most popular environment for users to generate personal content, share pieces of news, build social relations etc. Recent research focuses on analyzing comments' content on these platforms, as well as analyzing the commenters. An extensive analysis [13] of comments in social media communities investigates comments'

sentiment, rating and popularity in Youtube videos and Yahoo! News posts. Momeni and Sageder [9] perform a comparative analysis of comments in Flickr and Youtube. The authors point out different textual, semantic, and topical features of the comments, which are later used to predict the comment's usefulness. Towards identifying the characteristics of influential users, Martin et al. [8] introduce an emotion lexicon-based technique that predicts the helpfulness of reviews posted on Trip advisor and Yelp.

In addition to social media, related research focuses on news media as well. Here, understanding and potentially predicting the user characteristics and preferences is the main goal. The problem of user profiling in media is tackled in [1], where the authors introduce the notion of *comment-worthy* news articles. They predict the comments' interestingness in blogs and news sites using an adapted topic model aiming at personalized recommendation of news articles to users. Similarly, Shmueli et al. [12] address the problem of ranking news comments according to the reader's personal interests in Yahoo! News using a factor model. Instead of analyzing existing comments, Cao et al. [2] extract relevant microblog posts to news articles and use them to automatically generate user comments for these news articles.

Moreover, since users shape the general public's opinion with their comments by often supplementing the news stories with new facts and expertise, approaches that automatically evaluate the comments' quality have received high interest in the literature. To this end, tools distinguishing the (in)appropriate and (ir)relevant comments could assist media to improve the news quality they offer. Related work includes the analysis of the quality of comments [4], and the measurement of the comment sentiment in order to conclude about the media's political leaning [10]. Additionally, the problem of comment relevance is addressed by [9], [3] and [5], with the latter assessing the degree of pertinence of comments by comparing their tf-idf vectors to the articles' in News York Times. Detecting the comments that shift the main article topic and change the article's focus at Digg.com is tackled by Wang et al. [15], while Zhang and Setty [16] identify sets of topic-wise diverse user comments in Reddit news articles.

Finally, multiple interesting prediction tasks emerge from news comments analysis. Among others, the volume of news comments is predicted with a random forest classifier by Tsagkias et al. [14] using a variety of comment and article metadata, as well as textual and semantic features derived from the comments. Rizos et al. predict news stories popularity based on users' comments and the properties of the social graph they form [11]. Since users abuse the commenting mechanism frequently by stating offensive or hate comments, Kant et al. [7] compare an SVM classifier to a pattern mining approach in order to detect spam comments in Yahoo! News articles.

In contrast to the above works, we analyze comments to investigate differences in readership and bias among different German newspapers. Automatically gaining insights in the huge amount of user-generated content in media will help us discover people's opinion over several issues. More specifically, the way readers perceive reality regularly depends on the different writing styles of different

news outlets and their respective journalists. For instance, it would be interesting to discover that users tend to leave more informative or insightful comments, when a newspaper is being brief and doesn't discuss thoroughly certain topics. Alternatively, a user may post funny or hate comments, when an article criticizes openly a person or an event.

Furthermore, the ability to identify a comment's origin is a step towards detecting correlations between the news providers and the news consumers. We share the intuition of [10] regarding media bias detection in news articles, that is, users tend to leave negative comments to articles that oppose their perspective and positive otherwise. Additionally, as introduced in [6], readers often choose to be informed by the sources that share their beliefs. Namely, one is more likely to perceive bias the further the slant of the news is from their own political position.

3 Predicting comments' original source

Our motivation stems from the idea that readers from different newspapers might use unique language and present different commenting patterns. There are indeed differences among users in news media in general: some users tend to be objective and include new facts to the articles, others leave subjective messages (e.g. supporting a party, an opinion), others may attack the journalist or comment writers with hate comments, etc. We are interested in whether the above styles are indicative of the comments' source or not. Hence, we aim at identifying the comment features that distinguish the users of different news outlets. This will allow us to classify comment threads belonging to certain newspapers. To this end, all the direct comments and comment replies in a given article are considered as a single document in our prediction task. That is, one document is the complete news comment thread of a given news article. We then use an SVM classifier to classify each instance to its respective newspaper. The feature selection and the parameter setting are described below.

3.1 Datasets

We analyzed six popular German newspapers, namely *Bild*, *Focus*, *Welt*, *Spiegel*, *Zeit* and *Faz*. The dataset characteristics are shown in Table 1. Our crawled data span from March 2016 until June 2016. The fifth column depicts the average comment length for each source after removing stop words. It appears that *Spiegel* and *Faz* readers tend to leave longer comments than users from other sources. Additionally, we also observed that *Bild* commenters could be characterized as more active in comparison to the rest of the outlets, as the average number of comments per article in *Bild* is higher than in the rest of the newspapers.

Although the number of articles does not vary significantly among the newspapers, we can observe that *Welt* is the outlet with the most comments and commented articles in total. In our experiments, after considering all articles having at least 1, 5 or 10 comments in separate configurations, we concluded

that the threshold (H) of 5 yields the best precision results and thus we only report on results using this threshold. The last column in Table 1 represents the number of articles with at least 5 comments for each source.

Table 1: Dataset Characteristics

Source	**Articles**	**Articles with ≥ 1 Comments**	**Comments**	**Avgerage Comment Length**	**Articles with ≥ 5 Comments**
Bild	1,358	316	11,332	21.6	186
Focus	1,764	965	2,651	58	80
Welt	1,852	1782	31,125	31.7	830
Spiegel	1,654	664	5,771	61.8	188
Zeit	1,045	1032	8,553	46.1	642
Faz	1,656	458	1,329	71.3	61

3.2 Features

This subsection describes the comment-based and article-based features that we use for the SVM classifier.

Number of comments and average comment length. The number of direct comments and comment replies are summed up representing the first dimension of the feature vector. In addition, the average comment length is calculated for each article after filtering out the terms that appear in our stop word list. As shown in Table 1, there are significant differences among the outlets regarding the volume of comments and their length. Hence, our intuition is that the above-mentioned features will constitute an important indicator for the respective news source.

Direct comment/reply ratio and distinct authors. The next two features refer to the users, regarding their activity and commenting behavior. The ratio between the direct comments and the nested ones is a numerical indicator of how interactive the commenters are and whether discussions are initiated by them or not. For instance, as illustrated in Figure 2, *Zeit* and *Bild* appear to have a higher number of user discussions than the other sources.

Moreover, the distinct number of authors per article is interesting as well, as it informs us about the comment availability and potential diversity. Articles with multiple commenters should contain a variety of opinions and statements, in comparison to stories that don't attract high user interest. Figure 3 presents the news articles that are covered by certain numbers of commenters. That is, e.g. around 90% of *Bild* and *Faz* news articles would be covered, if the top-30 commenters were considered. It should be also noted that for this plot we only

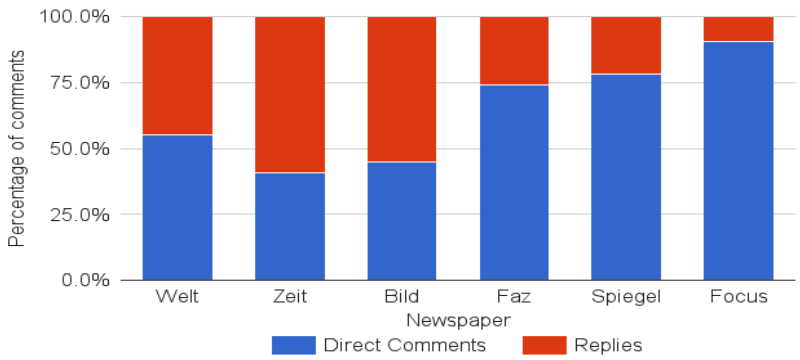

Fig. 2: Direct comments and nested replies for all news sources

use articles with H equal to 5. Our findings are in line with the work of Park et al. [10], where 50 commenters appear to cover around 80% of the overall dataset (when also considering solely articles with more than 5 comments).

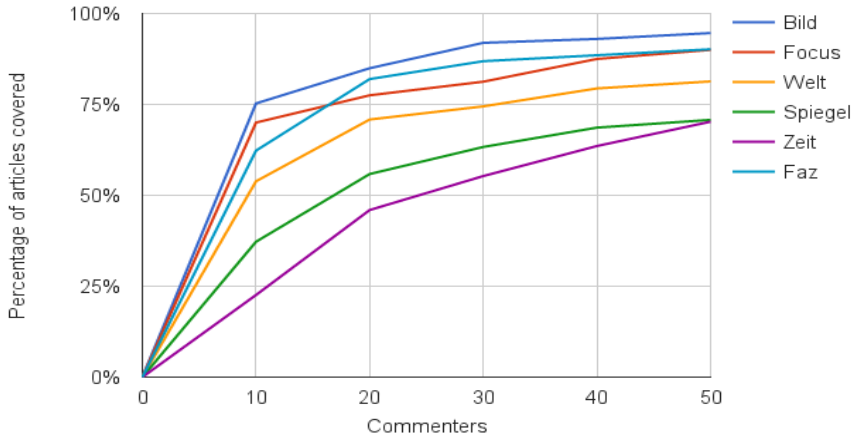

Fig. 3: Percentage of articles covered by k commenters, where $k = 10$, 20, 30, 40 and 50.

Comment terms. The current feature targets the comments' content and possibly opinionated language. We argue that the choice of language in the comments is the most representative feature of the users' perspective. Some comments aim at pointing out neglected facts from the articles and others might criticize the article's position or a politician's behavior, etc. Figure 2 illustrates

an example of a comment in *Zeit* and one of its replies, where the two users express two different sides of the same story. It is notable that we only consider the terms' *tf-idf* scores that are not stopwords, since only these provide semantic and meaningful information about the users' interests.

Newspaper uniqueness metric. Apart from user features, newspapers' characteristics play a key-role to our prediction task as well. Towards discovering representative and specific language used by different newspapers, we measure the similarity between comments and news articles of all sources, in terms of their common words. We compare the comments' terms with the articles' terms from all sources and measure their *overlap coefficient*. That is, for each comment thread to be classified, we compute the *overlap* (or also known as Szymkiewicz-Simpson) *coefficient* between its terms and the overall vocabulary from the articles of each newspaper, which results in six separate numeric counts as individual features. Our intuition is that this metric indicates whether the journalists and the readers from a given newspaper mention the same words.

Since commenters are often subjective and emotional, the current feature might also extract words that are not expected to be found in news media. This word set is a possible bias indicator, considering that news articles are expected to publish objective and well-rounded news pieces, so that readers are adequately informed.

4 Topic analysis

To ensure that all articles/comments are comparable across media outlets, we analyze the topics discussed in each news outlet.

As a first step towards understanding the discussions in our data, we are interested in detecting the topics mentioned in the newspapers' articles during our given time frame. For this purpose we use the latent Dirichlet allocation (LDA) implementation in Mallet, a Machine Learning Java Toolkit. We experiment with different values for the number of topics, namely 10, 20 and 40, but report only our findings for 20 topics, since the results are rather stable with varying topic numbers.

As shown in Table 2, the most discussed topics (15, 0) among all newspapers are focused on local affairs, with topic_0 touching upon financial issues. The least mentioned topics (9, 11, 1, 10, 17, 16) concentrate more on foreign politics, especially U.S. politics, which is an emerging topic as the general elections are approaching in the U.S.

In addition, Figure 4 presents the topic distributions across all newspapers. The x-axis represents the topics and the y-axis the volume of the discussion. One could infer that there are no extreme differences in the topic distributions among the outlets, that is, the same events/issues are covered by all newspapers. However, one notable exception are the comments in *Welt*, where the U.S. election topics (9,16,17) are clearly over represented.

Table 2: Top Terms for Each Topic (Ordered by Descending Popularity)

Topic Id	Frequent Terms
15	leben, politik, land, frage, deutschland, sagen, steht, kinder, sogar
0	prozent, deutschland, regierung, deutschen, zahl, land, praesident, frankreich, millionen
19	polizei, polizisten, frauen, demonstranten, koelner, maenner, verletzt, silvesternacht, koeln
7	euro, milliarden, deutschland, schaeuble, griechenland, geld, spd, gesetz, integration
3	spd, cdu, merkel, prozent, gabriel, afd, csu, seehofer, partei
5	russland, putin, usa, russischen, russische, praesident, obama, ukraine, nato
12	syrien, getoetet, stadt, waffenruhe, syrischen, terrormiliz, staat, aleppo, syrische
14	hofer, oesterreich, prozent, stimmen, fpoe, partei, wahl, parlament, van
4	afd, partei, deutschland, petry, islam, gruenen, cdu, kretschmann, npd
8	tuerkei, erdogan, boehmermann, tuerkischen, merkel, tuerkische, ankara, tayyip, recep
18	nordkorea, kim, journalisten, regierung, gericht, duendar, verurteilt, urteil, land
2	bruessel, anschlaegen, paris, anschlaege, flughafen, bruesseler, polizei, abdeslam, terroristen
13	panama, rousseff, papers, bundeswehr, zeitung, briefkastenfirmen, leyen, praesidentin, temer
6	cameron, khan, buergermeister, honecker, duterte, grossbritannien, london, johnson, britischen
16	trump, clinton, donald, sanders, republikaner, demokraten, hillary, cruz, vorwahlen
17	trump, clinton, sanders, donald, obama, hillary, prozent, cruz, trumps
10	the, waehler, and, twitter, primaries, staat, you, com, pic
1	trump, trumps, kasich, cruz, republikaner, senator, new, york, partei
11	fluechtlinge, tuerkei, griechenland, deutschland, grenze, fluechtlingskrise, migranten, fluechtlingen, europa
9	trump, sanders, clinton, cruz, rubio, donald, prozent, hillary, ted

Future work would be to incorporate this topical information in the classification task and discover whether it can improve our results, i.e. the users' commenting behavior differs for different combinations of topics and newspapers.

5 Classification results

The main goal of our work is to identify the newspaper that a certain comment thread comes from. Due to the small length of a single comment and the absence of rich content, we classify all the comments for a given article at once, instead

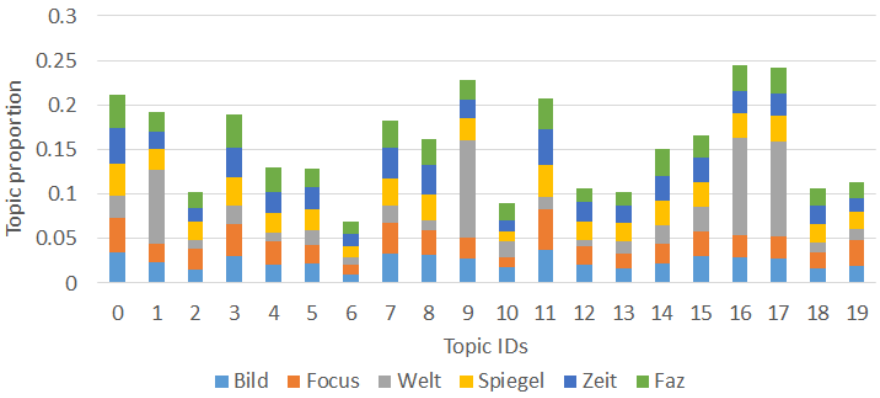

Fig. 4: Topic distributions for all sources using 20 topics

of considering them separately. For this purpose, we use the implementation of SVM classifier in Weka with the default parameter settings.

Regarding the training phase, we initially perform one-versus-one classification, training m=k*(k-1)/2 classifiers (one for each pair of newspapers) and output the majority vote among all classifiers for each input instance. Namely, we train the model with 40 documents per source and tested it on 20 documents per source — all randomly selected from our original dataset. Our second experiment is a one-versus-all classifier that is trained and tested on articles from all outlets, but it performs binary classification for a single given source. In particular, six different classifiers are built (one for each outlet) using 40 articles from the target source and 40 random articles from the remaining sources. The test set consists of 20 articles from the target news outlet and 20 arbitrary ones from the other outlets.

The above numbers of articles are set after examining the last column of Table 1. The maximum possible numbers are considered, in order to obtain a sufficient and equal amount of comments per source that will result to balanced training and test sets. Our future work includes obtaining more articles and subsequently more comments, fairly distributed to all six outlets, to achieve a higher comment quantity and diversity.

One-versus-one classification. The results of our first experiment are depicted in Table 3 and Table 4a. We can observe that the classifier performs best for *Bild* and yields inadequate results for *Focus* and *Zeit* with low recall or precision values respectively. The confusion matrix illustrated in Table 3 reveals that there is at least one comment from each source that is incorrectly classified as originating by *Zeit*. Considering that *Zeit* is the top-2 news outlet regarding the published number of articles with more than 5 comments, one might argue that highly popular and centrist newspapers, such as *Zeit*, contain a variety of

Table 3: One-Versus-One Classification Confusion Matrix

classified as →	**a**	**b**	**c**	**d**	**e**	**f**
a = Bild	19	0	0	0	1	0
b = Focus	0	8	0	0	5	7
c = Welt	2	0	12	0	5	1
d = Spiegel	1	2	2	11	3	1
e = Zeit	1	1	2	1	13	2
f = Faz	0	1	0	4	2	13

comments and commenter behaviors. This makes such news sources a good candidate for an unseen comment, as they could contain a wide range of different commenting styles.

Additionally, *Bild* articles are largely classified successfully. According to Table 1, *Bild* is also one the sources with the most overall comments, whereas the average comment length is relatively very low. Observing Table 1 and Table 3 concurrently, one can distinguish that when taking into account the most right-wing sources, namely *Bild*, *Welt* and *Focus*, the lower the average comment length is the higher our precision result becomes. Since short user comments can often be sharp or pithy, this is an interesting observation for readers of right-wing newspapers.

The average achieved precision is 65% and average recall 63%. Although the average performance score is a promising start, there is significant room for improvement, which we will further discuss in the following paragraph.

Table 4: Classification Results

(a) One-Versus-One

Newspaper	**Precision**	**Recall**
Bild	0.82	0.95
Focus	0.66	0.40
Welt	0.75	0.60
Spiegel	0.68	0.55
Zeit	0.44	0.65
Faz	0.54	0.65
Aaverage	0.65	0.63

(b) One-Versus-All

Newspaper	**Precision**	**Recall**
Bild	0.85	0.80
Focus	0.83	0.80
Welt	0.73	0.72
Spiegel	0.74	0.70
Zeit	0.80	0.75
Faz	0.90	0.90
Average	0.80	0.77

One-versus-all classification. Our next experiment is a one-versus-all classification. As previously mentioned, we build six different classifiers considering 40 articles from the target source and 40 random articles from the rest for the training set. The results are shown in Figure 4b. Surprisingly, although for the *Faz* articles the previous classifier achieved the worst results regarding precision, the

current classifier performs best for this particular outlet. The overall results vary from 73% (*Welt*) to 90% (*Faz*) precision. Moreover, recall is significantly higher, ranging from 70% (*Spiegel*) to 90% (*Faz*). This leads to an average precision of 80% and an average recall of 77%.

6 Conclusion and Future work

In this paper, we address the problem of automatically identifying the original newspaper that a comment thread of a given article belongs to. We analyze six well-known German newspapers, namely *Bild*, *Focus*, *Welt*, *Spiegel*, *Zeit* and *Faz*. To this end, we use an SVM classifier with different comment- and article-based features. For instance, the comment terms' *tf-idf* values and the number of available comments for an article are considered. The best results are accomplished by six one-versus-one classifiers (one for each newspaper pair), where our precision scores range from 70% to 90%. In order to reduce the variance of our results between the two classifiers and also among all news sources, we will perform the experimental evaluation on multiple random training and test sets.

Towards improving our current work, we would like to experiment with different feature combinations and evaluate their impact to our classifier. Apart from our version of measuring the unique characteristics of the newspapers, one could also attempt to take into account the levels of subjectivity in the news text, as an indication of the writing style. Moreover, the polarity (positive, negative, neutral) of each comment is a valuable information as well. It might hold that users in certain newspapers express their emotions more than in others or that users from specific outlets tend to express more their disapproval and criticism to certain issues than in other sources.

Instead of using all comment terms, we also experimented with using only the named entities found in the comments. The results were slightly worse than the reported ones, therefore we will continue to use all terms of the comments, as presented in this work. Although the named entities along with different feature combinations might work in the future, it is interesting to note that not only named entities are crucial for this problem, but verbs, adjectives and adverbs as well. Named entities mainly depict a text's topic, whereas adjectives and adverbs represent the author's perspective and discussion style. Finally, as previously mentioned, a direction we would like to follow is the incorporation of topical information in our classifier, which could potentially lead us to identify the original source of a comment thread more reliably.

References

1. Bansal, T., Das, M., Bhattacharyya, C.: Content Driven User Profiling for Comment-Worthy Recommendations of News and Blog Articles. RecSys pp. 195–202 (2015)
2. Cao, X., Chen, K., Long, R., Zheng, G., Yu, Y.: News comments generation via mining microblogs. In: WWW. pp. 471–472 (2012)

3. Das, M.K., Bansal, T., Bhattacharyya, C.: Going beyond Corr-LDA for detecting specific comments on news & blogs. In: WSDM. pp. 483–492 (2014)
4. Diakopoulos, N., Naaman, M.: Towards quality discourse in online news comments. In: CSCW. pp. 133–142 (2011)
5. Diakopoulos, N.A.: The Editor's Eye. In: CSCW. pp. 1153–1157 (2015)
6. Groseclose, T., Milyo, J.: A measure of media bias. In: The Quarterly Journal of Economics. pp. 1191–1237 (2005)
7. Kant, R., Sengamedu, S.H., Kumar, K.S.: Comment spam detection by sequence mining. In: WSDM. pp. 183–192 (2012)
8. Martin, L., Sintsova, V., Pu, P.: Are influential writers more objective? In: WWW. pp. 799–804 (2014)
9. Momeni, E., Sageder, G.: An empirical analysis of characteristics of useful comments in social media. In: WebSci. pp. 258–261 (2013)
10. Park, S., Ko, M., Kim, J., Liu, Y., Song, J.: The Politics of Comments: Predicting Political Orientation of News Stories with Commenters' Sentiment Patterns. CSCW (2011)
11. Rizos, G., Papadopoulos, S., Kompatsiaris, Y.: Predicting News Popularity by Mining Online Discussions. WWW pp. 737–742 (2016)
12. Shmueli, E., Kagian, A., Koren, Y., Lempel, R.: Care to comment? Recommendations for commenting on news stories. In: WWW. p. 429 (2012)
13. Siersdorfer, S., Chelaru, S., Pedro, J.S., Altingovde, I.S., Nejdl, W.: Analyzing and Mining Comments and Comment Ratings on the Social Web. TWeb 8, 1–39 (2014)
14. Tsagkias, M., Weerkamp, W., de Rijke, M.: Predicting the volume of comments on online news stories. In: CIKM. pp. 1765–1768 (2009)
15. Wang, J., Yu, C.T., Yu, P.S., Liu, B., Meng, W.: Diversionary comments under political blog posts. In: CIKM. pp. 1789–1793 (2012)
16. Zhang, H., Setty, V.: Finding diverse needles in a haystack of comments. In: WebSci. pp. 286–290 (2016)