Towards Analyzing the Bias of News Recommender Systems Using Sentiment and Stance Detection

Mehwish Alam*
mehwish.alam@kit.edu
FIZ Karlsruhe, Karlsruhe Institute of
Technology
Germany

Andreea Iana*
andreea@informatik.uni-
mannheim.de
University of Mannheim
Germany

Alexander Grote
alexander.grote@kit.edu
FZI Forschungszentrum Informatik,
Karlsruhe Institute of Technology
Germany

Katharina Ludwig
kaludwig@mail.uni-mannheim.de
University of Mannheim
Germany

Philipp Müller
p.mueller@uni-mannheim.de
University of Mannheim
Germany

Heiko Paulheim
heiko@informatik.uni-mannheim.de
University of Mannheim
Germany

ABSTRACT

News recommender systems are used by online news providers to alleviate information overload and to provide personalized content to users. However, algorithmic news curation has been hypothesized to create filter bubbles and to intensify users' selective exposure, potentially increasing their vulnerability to polarized opinions and fake news. In this paper, we show how information on news items' stance and sentiment can be utilized to analyze and quantify the extent to which recommender systems suffer from biases. To that end, we have annotated a German news corpus on the topic of migration using stance detection and sentiment analysis. In an experimental evaluation with four different recommender systems, our results show a slight tendency of all four models for recommending articles with negative sentiments and stances against the topic of refugees and migration. Moreover, we observed a positive correlation between the sentiment and stance bias of the text-based recommenders and the preexisting user bias, which indicates that these systems amplify users' opinions and decrease the diversity of recommended news. The knowledge-aware model appears to be the least prone to such biases, at the cost of predictive accuracy.

CCS CONCEPTS

• **Information systems → Recommender systems**; **Sentiment analysis**; **Personalization**.

KEYWORDS

news recommendation, filter bubbles, echo chambers, polarization, stance detection, sentiment analysis, German news articles

*First two authors contributed equally to this research.

1 INTRODUCTION

Every day, large volumes of news articles are being published online, leading to an information load that exceeds the consumptive capacities of Internet users. To alleviate this information overload and to provide customized content to users based on their past interests, recommender systems are widely employed by news providers. However, this algorithmic news curation influences users' exposure to diverse content, by selectively filtering out articles that seem irrelevant (in order to maximize user engagement) [32] or inconsistent with the readers' preexisting beliefs and attitudes, as users are more likely to accept information that reinforces their opinions [10]. The power of recommender systems to shape users' perception of the world has over time lead to concerns that individuals are being isolated from diverse perspectives through algorithmically created "filter bubbles" [32] and self-selected "echo chambers" [20, 41], in which they only interact with individuals with similar ideological viewpoints. In the context of (political) news, an over-exposure to less diverse viewpoints may determine an attitude-reinforcing spiral [8, 40] which, in the long run, can lead to opinion polarization, or even radicalization of individuals with extreme political or ideological viewpoints [2, 3, 25].

Diversity represents a paramount characteristic not only for news quality [29], but also for ensuring a balanced and broad variety of information to create a well-informed public in any democratic society [17]. Nonetheless, news diversity is a multi-faceted concept [14] which can refer, simultaneously, to pluralism of sources (*source diversity*) [44], of discussed topics (*content diversity*) [14], and of stances taken on a given topic (*viewpoint diversity*) [1]. Some previous works have focused on constructing datasets for detecting and analyzing the more general notion of media bias in news articles [9, 23, 39], and on establishing a connection between the research on media bias in the social science and computer science fields [15].

From a recommender-system point of view, sentiment analysis and stance detection have been used to measure and control news diversity. In sentiment analysis, systems determine whether a piece of text is positive, negative, or neutral [24, 34]. However,

in stance detection, systems are to determine an article's or author's viewpoints towards a given target issue, which may not even be explicitly mentioned in the text [22, 36]. Stance and sentiment annotations can be used to identify whether news recommender systems intensify selective exposure and increase users' vulnerability to polarized opinions and fake news. At the same time, stance and sentiment labels can constitute cues for controlling the level of polarization in personalized recommender systems, and for generating suggestions with more diverse sentiment orientations and viewpoints.

In this paper, we use sentiment and stance annotations to analyze whether different kinds of recommender systems suffer from an underlying bias towards a certain sentiment or viewpoint, and whether this decreases diversity of recommendations and intensifies users' selective exposure [33]. To this end, we use transfer learning to annotate GeNeG, a dataset of German news articles, with sentiment scores and stance labels. These annotations are then used to examine the results of four recommender systems with regards to the sentiment and stances identified in the suggested articles, as well as those from the user's reading history.

The rest of the paper is structured as follows. Firstly, we discuss related work on stance detection in various languages (Section 2). Secondly, we introduce GeNeG and describe its curation and annotation process (Section 3). Thirdly, in Section 4 we evaluate a set of news recommender systems, and examine their sentiment and stance bias. We conclude with a discussion of the findings and an outlook of the next steps planned for this study in Section 5.

2 RELATED WORK

In the following, various studies are discussed which take into account the creation of datasets for stance detection in different languages since the multilinguality aspect of stance detection is fairly understudied. We then more particularly focus on the multilinguality aspect in stance detection. For an exhaustive survey on techniques for stance detection please refer to [22]. In the current study, we do not discuss knowledge-aware news recommender systems, however, for more details on knowledge-aware news recommender systems please refer to [18].

2.1 Existing Datasets for Stance Detection

STANDER [6] is a dataset for stance detection and fine-grained evidence retrieval task for future research in stance detection, as well as multitask learning. It contains 3,291 annotated articles by experts (professional journalists). This corpus is also aligned with the Twitter dataset WT-WT [6] corpus. The topic of the news articles in STANDER is "mergers of US companies in the healthcare industry". The stance annotations are 'support', 'refute', 'comment', and 'unrelated'. Moreover, the corpus also includes 'evidences' as an annotation which indicates the part of text used to identify the stance of the article. 2 to 4 annotators were employed, and majority voting was used to decide on the final annotation.

In [31], the authors present a dataset of tweets annotated with the tweeter's stance (in favor or against) regarding an already chosen target. The dataset contains six targets of interest commonly debated in the United States, i.e, 'Atheism', 'Climate Change is a Real Concern', 'Feminist Movement', 'Hillary Clinton', 'Legalization of Abortion', and 'Donald Trump'. For each target, 1000 tweets were randomly sampled from the initially collected 1.7 million tweets. The tweets were annotated by 8 annotators. In total, the corpus contains 4,870 annotated tweets. This dataset has been used for the stance detection task in SemEval 2016[1]. This stance dataset, which was subsequently also annotated for sentiment, can be used to better understand the relationship between stance, sentiment, entity relationships, and textual inference.

In the Fake News detection Challenge, FNC-1 [16], the organizers focus on performing stance detection on document level. The documents are to be classified into four classes: 'agree', 'discuss', 'disagree', 'unrelated'. The data contains 2,587 headlines and documents. Out of these documents, 7.4% are labeled 'agree', 2.0% 'disagree', 17.7% 'discuss', and 72.8% 'unrelated', respectively. All the above datasets for stance detection are for English language, and thus, not usable for performing stance detection on GeNeG.

2.2 Multilingual Stance Detection

In [27] the authors create Cheese, a new dataset containing 3,693 pairs of debate questions and the associated Swiss news articles in German, annotated with stances and emotions. The news articles were published between 2004 and 2020, and cover 24 topics, including science, environment, politics, religion, society, etc. The stance annotation include 'in favor', 'against', 'discussion', and 'unrelated', while the emotion annotations are 'joy', 'trust', 'fear', 'anticipation', 'sadness', 'disgust', 'anger', 'surprise', and 'no emotion'. Along with the dataset, the authors provide a supervised classification task targeting the stance of the news article with respect to the question. The classification algorithm uses a contextual language model in German (German BERT).

SardiStance [5] is a stance detection task for tweets in Italian. The shared task was created for two different settings: (i) textual stance detection, exploiting only the information provided by the tweet, and (ii) contextual stance detection, with the addition of information on the tweet itself, such as the number of retweets, the number of favors or the date of posting, contextual information about the author, such as follower count, location, user's biography, and additional knowledge extracted from the user's network of friends, followers, retweets, quotes and replies. The dataset constructed for the task contains 700K tweets written in Italian about "Movimento delle Sardine". The tweets were collected from 11-2019 to 01-2020, and annotated with the labels 'against', 'in favor', 'neutral', and 'irrelevant'. The dataset also contains annotations indicating whether the tweet is 'ironic', 'non-ironic', or 'undefined'.

In [47], the authors present the task of stance detection on Chinese Microblogs. This is further divided into two sub-tasks, where the first is a supervised classification task which detects the stance towards five targets of interest with given labeled data. The second is an optional unsupervised task which requires only unlabeled data.

Previously discussed datasets focus on only one language. As a step further, in x-stance [42] the authors facilitate multilingual, multi-target stance detection, which also helps in performing cross-lingual stance detection. The authors use Multilingual BERT [7] for

---

[1]https://alt.qcri.org/semeval2016/task6/

performing stance classification. The dataset contains question and news article pairs in English, French, German, and Italian.

3 ENRICHING GERMAN NEWS KNOWLEDGE GRAPH WITH SENTIMENTS AND STANCE

This section discusses necessary details about GeNeG, along with the methods adopted for performing sentiment analysis as well as stance detection.

3.1 GeNeG

GeNeG [19] is a news knowledge graph constructed from a dataset of 4,557 German news articles on the topic of refugees and migration, collected from 39 media outlets that cover a wide political spectrum, including far-right and far-left online publications. The corpus consists of news published between 01-01-2019 and 20-10-2020, selected based on keyword stems representative of the topic (e.g., 'flüchtl', 'migrant', 'asyl'), and collected from the news outlets via a Web crawler[2]. The dataset contains rich textual and metadata information, such as content, provenance, publishing dates, authors, or tagged keywords. Furthermore, named entities have been extracted from the articles' content (e.g. persons, locations) and metadata (e.g. publisher, author) and disambiguated using Wikidata [45].

The news knowledge graph built from this dataset represents a heterogeneous network, where the news content and real-world entities are represented as nodes, while different relations between these items constitute the graph's edges. The nodes are divided into *literals*, denoting textual content, dates or polarization scores, and *entities*, indicating identified named entities. In turn, entity nodes can be linked to Wikidata, or non-disambiguated (if they are not found in Wikidata). GeNeG is provided in three flavors: a *base graph* – containing textual information, metadata, and entities extracted from the articles, an *entities graph* – derived from the base version by removing all literal nodes, it contains only entities extracted from the articles and it is enriched with their three-hop Wikidata neighbors, and a *complete graph* – combining the previous two graphs, and incorporating both literals and entities. The *base GeNeG* contains 54,327 nodes and 186,584 edges denoting 16 properties. The *entities GeNeG* consists of 844,935 nodes, and 6,615,972 edges representing 1,263 properties. Lastly, the *complete GeNeG* comprises of 868,159 nodes and 6,656,779 edges denoting 1,271 properties.

3.2 Sentiment Annotations

To annotate each news article with a sentiment score, we used a pre-trained BERT-based [7] sentiment classification model for German language texts [13]. The model uses a softmax function to calculate a probability estimate, which classifies each document as either positive, neutral or negative. To transform these probabilities into a sentiment score in the range of -1 to 1, we take the negative sentiment probability and subtract it from the positive sentiment probability. More formally, this is expressed by $s = p_p - p_n$, where s equals the sentiment score, while p_p and p_n represent the positive and negative sentiment probability, respectively. We ignore the neutrality score, since it is implicitly encoded as $1 - p_p - p_n$. On the

GeNeG corpus we observed a slight skewness towards articles with negative sentiment, as indicated by the mean and median sentiment scores of $s_{mean} = -0.154$, and $s_{median} = -0.019$, respectively.

3.3 Transfer Learning for Stance Detection on German News Articles

In order to perform stance detection on the news articles collected on the topic of refugees and migration, transfer learning was performed. In order to do so, the first step is to select a proper training dataset for articles in German, and the second step is to classify the articles into two classes, namely 'in favor ' or 'against'.

Training Data Selection. There are two datasets which are introduced for German language, i.e., Cheese [27] which specifically focuses on German language, and the German part of x-stance [42]. The Cheese dataset was excluded from consideration for training since the number of instances in each of the classes, i.e., 'in favor', 'against', 'unrelated', 'discussion', were highly unbalanced (i.e., 702, 286, 1428, and 774), leading to unbalanced classes predicted for our corpus. This issue has been discussed in [37] in more detail, where the authors propose a cascade of binary classifiers to consider the hierarchy of classes in stances. More specifically, the authors first classify the text into relevant or irrelevant, and then label the relevant texts as either 'neutral' or 'stance'. The final binary classification is performed using 'in favor' or 'against' class labels for the texts which were previously classified as 'stance'.

In our case, we use the x-stance [42] dataset for training the classifier, which contains a balanced number of articles for both classes, i.e., 'in favor' and 'against'. The training data was extracted from x-stance with the language label 'de', i.e., only the German subset. The statistics of the corpus extracted from x-stance are given in Table 1.

Table 1: Statistics of the the German part of the x-stance dataset.

Dataset	Against	Favor	Total
Training	17130	16720	33850
Validation	1451	1420	2871
Test	5882	6009	11891

Stance Detection in GeNeG. Stance detection takes into account a question and article pair, and classifies the stance of the article as being in favor or against the question at hand. The questions used to build the question and news article pairs for classification are given in Table 2. GermanBert[3] [4] was fine-tuned on the training dataset provided in x-stance. The hyper-parameters for the best performing model on the test set are: *learning rate* = $3e - 5$ and *number of epochs* = 4. Afterwards, transfer learning was performed for classifying the news articles in GeNeG (as test dataset) into one of the two classes, i.e., 'in favor' or 'against', according to one of the predefined questions. In future studies we would like to perform human evaluation on the annotated GeNeG articles. Table 3 show the results of the stance detection for each of the questions, as well as the average score, computed as $\frac{favor - against}{favor + against}$.

---

[2]https://github.com/andreeaiana/german-news

[3]https://huggingface.co/bert-base-german-cased

Table 2: Questions for the question-news article pairs.

German Question	English Translation (for understandability)
(Q1) Befürworten Sie, dass Flüchtlinge nach Deutschland kommen?	Are you in favor of refugees coming to Germany?
(Q2) Befürworten Sie, dass Flüchtlinge in Deutschland leben?	Are you in favor of refugees living in Germany?
(Q3) Befürworten Sie, dass Flüchtlinge in Deutschland arbeiten?	Are you in favor of refugees working in Germany?
(Q4) Sollte Deutschland Flüchtlinge aufnehmen?	Should Germany take in refugees?
(Q5) Sollte Deutschland Flüchtlingen helfen?	Should Germany help refugees?

Table 3: Articles 'in favor' of or 'against' the pairing question.

Question	Articles in Favor	Articles Against	Avg. Score
(Q1)	2165	2392	-0.050
(Q2)	2193	2364	-0.038
(Q3)	2210	2347	-0.030
(Q4)	2120	2437	-0.070
(Q5)	2192	2365	-0.038

Finally, GeNeG was additionally populated with the output of the stance detection with the help of two newly created properties, namely geneg:in_favor and geneg:against. The resulting triples have the form $\langle article, stance, question \rangle$, where $stance \in \{geneg:in\_favor, geneg:against\}$, and $question \in \{Q_n, n = \overline{1,5}\}$. The dataset is available through Zenodo[4] with restricted access.

4 SENTIMENT AND STANCE BIAS OF NEWS RECOMMENDERS

The stance and sentiment of news articles can be used to determine whether news recommenders are biased towards a certain sentiment or stance of the articles, and if this, in turn, correlates with and reinforces users' existing viewpoints on a given topic. In this section, we firstly introduce the four news recommender systems analyzed and describe the user data. Afterwards, we evaluate the performance of the models and investigate their sentiment and stance bias.

4.1 Recommendation models

Collaborative filtering is the most adopted recommendation method in fields such as music or movie recommendation. Nonetheless, content-based approaches are the most widely used in the field of news recommendation mostly due to the fact that users generally do not have long-term profiles on news websites, with reading history being limited to a single session, and feedback being collected almost exclusively implicitly from click logs [18, 21]. We compared four different content-based recommender systems in our analysis. The first three are *textual*, i.e., they are solely based on the articles' texts, whereas the last one is *knowledge-based*, using the graph representation and the information from Wikidata instead of the articles' texts. All text-based models use cosine similarity to determine the similarity between a candidate news article and the user's history of read news.

- The **Term Frequency-Inverse Document Frequency (TF-IDF)** recommender uses a TF-IDF [38] vector representation for each article.
- The **Word2vec** recommender encodes articles using word embeddings learned from a large text corpus [30]. More specifically, we use a pre-trained German word2vec model trained on the Common Crawl and Wikipedia dataset [12] to learn latent representations of the words in an article. The article's vector representation is computed as the average of its words' embeddings.
- The **Transfomer** architecture uses an attention mechanism to incorporate context in text embeddings [43]. We use a pre-trained cross-lingual model for sentence embeddings in English and German [28, 35] to encode the sentences in the news articles. The final representation of the article is obtained by averaging the embedding vectors of all of its sentences.
- **RippleNet** [46] is a knowledge-aware recommender which propagates a user's potential preferences along the edges of the knowledge graph. The model generates *ripple sets* (i.e. sets of multi-hop entity neighbors encoding potential user interests) based on the entities extracted from the user's read articles. These ripple sets are used to explore higher-order preferences, where the strength of the user's preference diminishes proportional to the distance from the original seed in the knowledge graph. The final click probability is predicted using the preference distribution of the user for a candidate news, obtained by superposing multiple ripple sets. For RippleNet, we use the *entities GeNeG* as the knowledge graph for the recommender. In particular, the sentiment and stance information was *not* included in the graph used to compute the recommendations.

4.2 User Data

We collected the user data through an online study aimed at measuring the political polarization effect of recommender systems on users. The study was based on the news dataset described in Section 3.1. However, 732 articles longer than 1,500 words were removed from the experiment in order to limit the reading time required for the participants, and to ensure a strong response to the stimulus. Each of the participants in the study was randomly assigned to one of four recommenders, namely TF-IDF, Word2vec, Transformer, or a random recommendation baseline. Each participant was asked to choose an article from a preview of six articles, and then to read it. The user's choices were included in his or her reading history. This process was repeated four times, resulting in four interactions per participant.

We split the user data into training and test sets using a 80:20 ratio. Since 85% of the user ratings were generated using recommendations of the text-based recommenders, these baselines are prone to overfit. Therefore, we generate a test subset containing only user ratings for randomly suggested articles from the complete test set. The user data statistics can be found in Table 4. For more details on the online study and data collection, please refer to Appendix A.

Table 4: User data statistics.

Dataset	Items	Users
Total	3,825	1,417
Training	3,365	1,414
Complete test	1,633	1,174
Random test	316	177

4.3 Evaluation of Recommender Systems

In the following, we describe the experimental setup and discuss the evaluation results of the proposed recommenders.

4.3.1 Experimental Setup. We evaluated the four recommender models on click-though-rate (CTR) prediction. In this scenario, each recommender is applied on every user-article pair from the test set to predict the user's likelihood of clicking the candidate article. Furthermore, we applied a min-max scaling to the similarity measures outputted by the text-based recommenders as an approximation of probability scores. We use Accuracy (ACC), Area Under the Curve (AUC), and the F1 score to evaluate the performance of the models.

The key parameters settings for the analyzed recommender systems [5] are as follows. For TF-IDF, we use an n-gram range of 1 and 2, and l2-norm regularization. The Word2vec and the Transformer recommenders use embedding vectors of dimension 300, and respectively, 768. In RippleNet, we empirically set the number of hops to $H = 1$, the size of the user's ripple set to 16, the dimension of the item and knowledge graph embeddings to $d = 48$, and the training weight of the knowledge graph embedding to $\lambda_2 = 0.03$. We use the default values for news recommendation from [46] for the other parameters.

4.3.2 Results. The experimental results of the benchmarked recommenders are summarized in Table 5. The text-based recommenders achieve the best scores in terms of AUC and F1 on the complete test set, whereas RippleNet outperforms the Word2vec and Transformer recommenders in terms of accuracy. However, the text-based recommendation models overfit on the complete test data, due to the user ratings provenance. More specifically, the ratings collected from users assigned to a TF-IDF recommender account for more than 55% of the cases (for more in-depth statistics regarding the ratings' provenance, please refer to Appendix A).

Consequently, we investigated the performance of the recommender systems on the random test set, in order to reduce the bias of the text-based models. In this evaluation setting, RippleNet outperforms all other recommenders in terms of accuracy and AUC. Moreover, it obtains a higher F1 score than the TF-IDF model, which

[5]The code and data are available at https://github.com/andreeaiana/geneg_benchmarking.

Table 5: CTR prediction results.

Model	Complete			Random		
	ACC	AUC	F1	ACC	AUC	F1
TF-IDF	**0.732**	**0.873**	0.647	0.487	0.499	0
Word2vec	0.514	0.794	**0.674**	0.499	0.474	0.663
Transformer	0.505	0.779	0.671	0.499	0.515	**0.665**
RippleNet	0.553	0.574	0.523	**0.559**	**0.578**	0.531

is unable to make any correct predictions for the articles in this test set, resulting in a F1 score of 0.

Overall, we conclude that, by using only entities in GeNeG for computing recommendations, RippleNet is able to achieve a decent performance, while being more robust to changes in the structure of the data compared to the purely text-based recommenders.

4.4 Bias Analysis

In addition to the recommenders' predictive accuracy, we analyzed their recommendations to identify whether the recommenders are prone to stance or sentiment bias. In this study, each of the text-based models generated an output containing the top k most similar articles to the ones in the user history, whereas RippleNet recommended the k articles with the highest probability of being clicked by the user. In our subsequent bias analysis, we set the number of recommended articles for all recommenders to $k = 5$, as we observed that the majority of investigated outlets suggest, on average, five other articles related to the one currently read by the user.

In order to quantify the stance of the news articles and to calculate an overall bias score for the users and recommenders, we employ the following transformation function for an article's stance label:

$$stance\_score = \begin{cases} +1 & \text{if } stance\_label = \text{Favor} \\ -1 & \text{if } stance\_label = \text{Against} \end{cases}$$

For each user, we compute the average sentiment bias score as the mean of the sentiment scores of the articles included in his or her reading history. Similarly, the recommender sentiment bias score per user was obtained as the average of all sentiment scores of the recommended articles. Lastly, the mean of all recommender sentiment bias scores over all users constitutes the recommender's average sentiment bias score. The average user and recommender stance bias scores are calculated analogously. The average bias scores for both sentiment and stances fall in the interval $[-1, 1]$, where -1 represents a user's or recommender's tendency for articles with negative sentiments or stances against the given topic, whereas +1 denotes the opposite situation.

4.4.1 Recommender Bias. Firstly, we used the average recommender bias scores to answer the following question: **do the recommender systems have a tendency to recommend articles with a certain sentiment or stance?**

Table 6 shows the results in terms of *sentiment bias*. As it can be seen here, all recommenders are more likely to suggest articles with negative sentiments. However, this skewness towards negative sentiments should be interpreted by taking into account that the news articles in GeNeG have, on average, a negative sentiment

Table 6: Average user and recommender sentiment scores. The statistical significance with a Student's t-test is denoted with * ($p - value < 0.01$) and ** ($p - value < 0.05$). A star in the case of user sentiments denotes statistical significance between the average user and corpus scores. In the case of recommender sentiment, (*/*) denotes statistical significance, firstly with the average user score, and secondly with the average corpus score. A '-' sign shows that no such significance was found.

Test set	Avg. user sentiment score	Avg. recommender sentiment score			
		TF-IDF	Word2vec	Transformer	RippleNet
Complete	-0.171*	-0.162 (-/-)	-0.169 (-/**)	-0.157 (-/-)	-0.148 (*/-)
Random	-0.169	-0.141 (-/-)	-0.170 (-/-)	-0.160 (-/-)	-0.150 (**/-)

Table 7: Average user and recommender stance scores. The statistical significance with a Student's t-test is denoted with * ($p - value < 0.01$) and ** ($p - value < 0.05$). A star in the case of user stance denotes statistical significance between the average user and corpus scores. In the case of recommender stance, (*/*) denotes statistical significance, firstly with the average user score, and secondly with the average corpus score. A '-' sign shows that no such significance was found.

Question	Avg. user stance score	Avg. recommender stance score (complete / random test set)			
		TF-IDF	Word2vec	Transformer	RippleNet
(Q1)	-0.109 / -0.093	-0.140 (-/-) / -0.227 (**/−)	-0.165 (**/-) / -0.219 (**/-)	-0.136 (-/-) / -0.172 (-/-)	-0.082 (-/-) / -0.054 (-/-)
(Q2)	-0.102** / -0.093	-0.132 (-/-) / -0.220 (**/−)	-0.158 (**/-) / -0.207 (**/-)	-0.131 (-/-) / -0.169 (-/-)	-0.074 (-/-) / -0.038 (-/-)
(Q3)	-0.092** / -0.081	-0.127 (-/**) / -0.215 (**/-)	-0.149 (**/**) / -0.205 (**/-)	-0.116 (-/**) / -0.164 (-/-)	-0.062 (-/**) / -0.024 (-/-)
(Q4)	-0.117 / -0.106	-0.167 (**/-) / -0.255 (**/-)	-0.178 (**/-) / -0.268 (*/-)	-0.157 (-/-) / -0.179 (-/-)	-0.095 (-/-) / -0.084 (-/-)
(Q5)	-0.079 / -0.081	-0.130 (**/-) / -0.237 (*/−)	-0.135 (**/-) / -0.199 (**/-)	-0.124 (-/-) / -0.143 (-/-)	-0.060 (-/-) / -0.055 (-/-)

score, as shown in Section 3.2. Nonetheless, we observe that textual recommenders, in 5 out of 6 cases, are more prone towards news with negative sentiments, as indicated by the average recommender sentiment bias, which is larger than the average sentiment score of the articles in the dataset, namely $s_{mean} = -0.154$. In contrast, the knowledge-aware model recommends, in both cases, articles with slightly less negative sentiment than the average sentiment score of the news in the corpus. However, these observations are statistically significant only for the Word2vec recommender on the complete test set, for $p - value < 0.05$, as indicated in Table 6.

Similarly, Table 7 illustrates the average recommender *stance bias* on both test sets. All recommenders show a bias towards news articles that take a stance against the topic of refugees and migration, for all questions used to represent this topic. As shown in Table 3, in GeNeG there are more articles against the topic, than in favor of it, for all of the five questions used in the process of stance detection. While this can partly explain the recommenders' preferences for the more negative articles, the average scores for the articles selected by the recommenders are still more negative than the average scores in the entire corpus depicted above. Again, as for the sentiment bias, the knowledge-based recommender has the weakest negative bias, showing only little deviation from the overall average scores in the dataset. Nonetheless, the difference between the average sentiment scores of the recommenders and the corpus is statistically significant only for question (Q3).

4.4.2 Correlation between the Recommender and User Biases. We then pose the question: **how does the recommenders' sentiment and stance bias correlate with the existing user sentiment and stance bias?** This analysis aims to identify whether the recommenders tend to generate suggestions that agree with or reinforce the user's existing beliefs. Tables 6 and 7 show the average sentiment and stance user biases.

We observed that readers have a slight preference for news articles with negative sentiments (Table 6), which is not fully explained by the overall skewness towards negative sentiment in the dataset. Nevertheless, in seven out of eight cases, the recommender's average sentiment bias is marginally less negative than the user bias, indicating that the recommenders include articles with less negative sentiment in their results list. Here again, one should take into consideration that the articles in our dataset have, on average, a negative sentiment. Even in this context, we notice that users are more prone to reading articles with negative sentiments, with the average user bias on the complete test set being -0.171, and respectively, -0.169 on the random test set, slightly larger than the average sentiment score of the news articles in GeNeG, $s_{mean} = -0.154$. According to the Student's t-test, this is a statistically significant difference. However, there does not seem to be a large difference between the recommender and the users' bias scores, with the former following the same pattern. This could indicate that the recommenders indeed learn from the user's history and preferences and generate suggestions that amplify the readers' existing preferences for a given sentiment. However, we observed a statistically significant difference between the user and the recommender average sentiment scores in the case of RippleNet on the complete test set, which shows that the knowledge-based recommender might be less prone to amplifying existing sentiment preferences.

A similar pattern can be observed in the case of stance bias. On the one hand, the text-based recommenders appear to exacerbate the user's preference towards news against the topic of refugees and migration, as indicated by the larger negative average stance bias scores of the recommenders, on both test sets. Additionally, this behavior of the recommenders appears to be more pronounced on the random test set. The difference between the average stance scores of Word2vec recommendations and the average user stance

scores are significant on both test sets, while in the case of TF-IDF this observation holds true mostly on the random test set. On the other hand, although the knowledge-aware recommender also seems biased towards viewpoints against the given topic, it also appears to suggest, on average, more articles which also have favorable stances towards the topic. This is shown by the average stance bias score of RippleNet, which is lower than the average user stance bias score. We observed no statistically significant differences between the user and recommender stance bias in the case of the Transformer and RippleNet recommenders.

Through a closer look at the correlation between the recommender and the user sentiment and stance biases, we identified five possible cases of bias correlation, as follows:

- (C1) Bias in the same direction: both the user and the recommender have a tendency towards negative (positive) sentiments and/or against (in favor) stances.
- (C2) Bias in opposite directions: the user has a preference for negative (positive) sentiments and/or against (in favor) stances, while the recommender tends to suggest articles with an opposite sentiment and/or stance.
- (C3) Skewed towards the user: the user has a preference for articles with negative (positive) sentiments and/or against stances, while the recommender is balanced with regards to the sentiment and/or stance of the recommended articles.
- (C4) Skewed towards the recommender: the user is balanced with regards to the sentiment and/or stance of the read articles, while the recommender tends to suggest articles with negative (positive) sentiment and/or against (in favor) stance.
- (C5) No bias: neither the user, nor the recommender show any sentiment and/or stance bias.

It should be noted that in the case of sentiment bias, only the first two cases, (C1) and (C2), are possible, given the nature of the sentiment scores, and the fact that there are no perfectly neutral articles in the corpus.

Table 8 shows the number of users that fall in each of the first two cases with regards to *sentiment bias*. The results show that for both test sets, almost always the recommender will suggest articles with the same sentiments as the ones found in the user's history.

Table 8: Recommender-user sentiment correlation counts.

Bias case	Counts (complete / random test set)			
	TF-IDF	Word2vec	Transformer	RippleNet
(C1)	1166 / 177	1166 / 177	1165 / 177	1160 / 177
(C2)	8 / 0	8 / 0	9 / 0	14 / 0

Furthermore, Table 9 shows the number of users for each of the five bias cases and questions in terms of *stance bias*. We observe that all recommenders generate recommendations that most often fall in one of the first two cases, with the same-direction bias being more likely. The third most probable case is the one in which the overall stance bias score of the user is close to 0, denoting a balanced news consumption with articles both in favor and against the topic, whereas the recommender is prone to suggesting articles with one particular viewpoint. The least likely cases are the ones in which the user is biased towards a certain stance, while the recommender

is neutral, followed by the scenario in which both the user and the recommender are neutral with regards to the stance of the news. Overall, the prevalence of the the first bias case is indicative of the fact that recommenders not only learn from user's reading history, but also tend to amplify existing opinions by recommending news with the same stances as the ones previously read by the user.

Table 9: Recommender-user stance correlation counts.

Question	Bias case	Counts (complete / random test set)			
		TF-IDF	Word2vec	Transformer	RippleNet
Q1	(C1)	600 / 81	575 / 68	575 / 88	470 / 83
	(C2)	358 / 63	367 / 70	374 / 52	471 / 61
	(C3)	14 / 1	30 / 7	23 / 5	31/ 1
	(C4)	198 / 31	198 / 31	194 / 29	193/ 31
	(C5)	4 / 1	4 / 1	8 / 3	9 / 1
Q2	(C1)	595 / 79	575 / 69	572 / 86	479 / 82
	(C2)	362 / 65	367 / 69	378 / 54	462 / 62
	(C3)	15 / 1	30 / 7	22 / 5	31 / 1
	(C4)	199 / 31	198 / 31	193 / 29	194 / 31
	(C5)	3 / 1	4 / 1	9 / 3	8 / 1
Q3	(C1)	600 / 78	574 / 68	574 / 86	477 / 81
	(C2)	360 / 66	370 / 70	375 / 54	467 / 63
	(C3)	15 / 1	31 / 7	26 / 5	31 / 1
	(C4)	196 / 31	195 / 31	193 / 30	191 / 31
	(C5)	3 / 1	4 / 1	6 / 2	8 / 1
Q4	(C1)	611 / 82	590 / 76	538 / 86	495 / 87
	(C2)	342 / 54	356 / 61	402 / 48	448 / 50
	(C3)	13 / 2	20 / 1	26 / 4	23 / 1
	(C4)	206 / 38	198 / 39	204 / 37	200 / 39
	(C5)	2 / 1	10 / 0	4 / 2	8 / 0
Q5	(C1)	624 / 89	592 / 77	542 / 88	498 / 84
	(C2)	337 / 51	357 / 62	403 / 49	451 / 57
	(C3)	10 / 2	22 / 3	26 / 5	22 / 1
	(C4)	200 / 34	194 / 35	201 / 34	195 / 35
	(C5)	3 / 1	9 / 0	2 / 1	8 / 0

Lastly, we compute the Pearson correlation coefficient between the recommender and the user average bias scores, to further investigate whether our initial observations are statistically significant. Tables 10 and 11 illustrate the results of recommender-user sentiment, and respectively, stance bias correlation.

Table 10: Recommender-user sentiment bias correlation. Statistical significance is denoted with * ($p-value < 0.01$) and ** ($p-value < 0.05$).

Test set	Pearson Correlation (p-value)			
	TF-IDF	Word2vec	Transformer	RippleNet
Complete	0.325*	0.347*	0.216*	0.045**
Random	0.253*	0.310*	0.129	-0.033

In the case of *sentiment bias*, the results indicate a statistically significant positive correlation between the recommender and the user bias for the TF-IDF and Word2vec models on both test sets. In the case of the Transformer recommender, we notice a statistically significant positive correlation only on the complete test set. This means that if a user prefers only articles with a negative sentiment towards the topic of, for example, refugees living in Germany, these

Table 11: Recommender-user stance bias correlation. Statistical significance is denoted with * ($p-value < 0.01$) and ** ($p-value < 0.05$).

Question	Pearson Correlation (p-value) (complete / random test set)			
	TF-IDF	Word2vec	Transformer	RippleNet
(Q1)	0.248* / 0.078	0.196* / 0.029	0.187* / 0.162**	-0.022 / 0.111
(Q2)	0.241* / 0.078	0.196* / 0.019	0.176* / 0.148**	-0.008 / 0.115
(Q3)	0.252* / 0.072	0.199* / 0.003	0.195* / 0.143	-0.009 / 0.091
(Q4)	0.280* / 0.135	0.249* / 0.002	0.143* / 0.230*	-0.019 / 0.148
(Q5)	0.320* / 0.192**	0.269* / 0.042	0.168* / 0.230*	-0.011 / 0.112

recommenders will continue recommending news with a negative sentiment. In this case, the user's exposure to articles with positive sentiment will be constrained by the recommender, and the user will not have access to news that might not agree with his existing views. In contrast, in the case of RippleNet, we found a statistically significant correlation between the recommender and the user bias for the complete test set, for $p-value < 0.05$, and no statistical significance on the random test. The latter could potentially indicate that the recommender is less prone to amplifying the user's existing preference for a particular sentiment.

As far as *stance bias* is concerned, text-based recommenders are also statistically significant positively correlated with the existing user bias on the complete test set, for all questions, as shown in Table 11. However, this pattern does not hold true in the case of the random test set. In this case, we noticed that only the Transformer recommender shows a positive correlation with the user bias, which is statistically significant for four out of five questions, whereas the Word2vec model has no statistically significant bias correlation with the user history. In comparison, our analysis shows that the stance bias of the TF-IDF model is positively correlated with the user's bias only on question (Q5) for $p-value < 0.05$. Lastly, as in the case of sentiment bias, RippleNet shows no statistically significant correlation with the existing stance bias of users on either of the two tests sets, again demonstrating that the knowledge-based recommender is less prone to amplifying user bias.

5 DISCUSSION

In this paper, we have introduced GeNeG, a German news corpus on the topic of refugees and migration, including polarity annotations with respect to sentiment and stance. We have demonstrated that those annotations can be utilized to quantify the sentiment and stance bias of different recommender algorithms.

Our experiments show that purely text-based recommender systems expose amplification of user attitudes with respect to sentiment and stance, both constituting cues for the creation of filter bubbles in the process of algorithmic news curation, and the decrease of news diversity in terms of sentiments and viewpoints. Similar results have been found by Gao et al. [11], who conducted experiments to understand the effects of stance and credibility labels on online news selection and consumption. The results of their study show that these labels intensify the tendency of the people to look for opinions which are in-line with their own beliefs, and may lead the people being more vulnerable to polarized opinions and fake news.

In contrast, the knowledge-aware recommender appears to be less prone to both types of biases, but also less powerful in terms of accuracy on click-through rate prediction when computing recommendations solely based on entities. On the one hand, text-based recommenders are predisposed to identifying sentiment and stance cues in the text, as their recommendations are based only on textual information. On the other hand, using knowledge bases as sources of side information and taking into account solely named entities identified in the text (regardless of the textual context in which they are mentioned) diminishes such underlying bias, while capturing deeper knowledge-level connections between articles. Therefore, we argue that future research should find a balance between two obviously conflicting goals: performance (as understood by prediction accuracy) and bias (in terms of the sentiments and stances expressed in the recommended news articles). The results of this paper indicate that knowledge-based approaches are a good candidate when it comes to develop less biased recommenders. However, future work should investigate also the behavior of collaborative-filtering approaches in terms of sentiment and stance bias. Moreover, the analysis could be extended to other datasets in order to examine if the same patterns can be observed for a larger variety of topics.

As a future perspective, we would like to enable the recommender systems to take into account the sentiments, as well as the stance of an article. The news recommender system should be able to recommend topics which allow the user to consider differing opinions as compared to their own beliefs, hence reducing the effect of filter bubbles. Since in this paper we have shown how bias can be quantified, it would also be possible to introduce a combined score as a new optimization goal for algorithmic recommenders, which trades off accuracy and bias in a single score (just like, e.g., F1 score trades off recall and precision). Optimizing news recommender systems towards such a target function could be one possible approach towards developing less biased news recommender systems.

REFERENCES

[1] Christian Baden and Nina Springer. 2017. Conceptualizing viewpoint diversity in news discourse. *Journalism* 18, 2 (2017), 176–194.
[2] Eytan Bakshy, Solomon Messing, and Lada A Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348, 6239 (2015), 1130–1132.
[3] Andrei Boutyline and Robb Willer. 2017. The social structure of political echo chambers: Variation in ideological homophily in online networks. *Political*

psychology 38, 3 (2017), 551–569.

[4] Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's Next Language Model. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 6788–6796. https://doi.org/10.18653/v1/2020.coling-main.598

[5] Alessandra Teresa Cignarella, Mirko Lai, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2020. SardiStance @ EVALITA2020: Overview of the Task on Stance Detection in Italian Tweets. In *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online event, December 17th, 2020 (CEUR Workshop Proceedings, Vol. 2765)*. CEUR-WS.org. http://ceur-ws.org/Vol-2765/paper159.pdf

[6] Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020. Will-They-Won't-They: A Very Large Dataset for Stance Detection on Twitter. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 1715–1724. https://doi.org/10.18653/v1/2020.acl-main.157

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/n19-1423

[8] Wolfgang Donsbach and Cornelia Mothes. 2013. The dissonant self: Contributions from dissonance theory to a new agenda for studying political communication. *Annals of the International Communication Association* 36, 1 (2013), 3–44.

[9] Michael Färber, Victoria Burkard, Adam Jatowt, and Sora Lim. 2020. A multi-dimensional dataset based on crowdsourcing for analyzing and detecting news bias. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 3007–3014.

[10] Jonathan L Freedman and David O Sears. 1965. Selective exposure. In *Advances in experimental social psychology*. Vol. 2. Elsevier, 57–97.

[11] Mingkun Gao, Ziang Xiao, Karrie Karahalios, and Wai-Tat Fu. 2018. To Label or Not to Label: The Effect of Stance and Credibility Labels on Readers' Selection and Perception of News Articles. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 55:1–55:16. https://doi.org/10.1145/3274324

[12] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

[13] Oliver Guhr, Anne-Kathrin Schumann, Frank Bahrmann, and Hans Joachim Böhme. 2020. Training a Broad-Coverage German Sentiment Classification Model for Dialog Systems. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 1627–1632.

[14] Mario Haim, Andreas Graefe, and Hans-Bernd Brosius. 2018. Burst of the filter bubble? Effects of personalization on the diversity of Google News. *Digital journalism* 6, 3 (2018), 330–343.

[15] Felix Hamborg, Karsten Donnay, and Bela Gipp. 2019. Automated identification of media bias in news articles: an interdisciplinary literature review. *International Journal on Digital Libraries* 20, 4 (2019), 391–415.

[16] Andreas Hanselowski, Avinesh P. V. S., Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M. Meyer, and Iryna Gurevych. 2018. A Retrospective Analysis of the Fake News Challenge Stance-Detection Task. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*. Association for Computational Linguistics, 1859–1874. https://aclanthology.org/C18-1158/

[17] Natali Helberger. 2019. On the democratic role of news recommenders. *Digital Journalism* 7, 8 (2019), 993–1012.

[18] Andreea Iana, Mehwish Alam, and Heiko Paulheim. 2022. A Survey On Knowledge-Aware News Recommender Systems. *Semantic Web Journal* (2022).

[19] Andreea Iana, Alexander Grote, Kaharina Ludwig, Mehwish Alam, Philipp Müller, Christof Weinhardt, Harald Sack, and Heiko Paulheim. 2022. GeNeG: German News Knowledge Graph. https://doi.org/10.5281/zenodo.5913171

[20] Kathleen Hall Jamieson and Joseph N Cappella. 2008. *Echo chamber: Rush Limbaugh and the conservative media establishment*. Oxford University Press.

[21] Mozhgan Karimi, Dietmar Jannach, and Michael Jugovac. 2018. News recommender systems–Survey and roads ahead. *Information Processing & Management* 54, 6 (2018), 1203–1227.

[22] Dilek Küçük and Fazli Can. 2020. Stance Detection: A Survey. *ACM Computing Surveys (CSUR)* 53, 1 (2020), 12:1–12:37. https://doi.org/10.1145/3369026

[23] Sora Lim, Adam Jatowt, Michael Färber, and Masatoshi Yoshikawa. 2020. Annotating and analyzing biased sentences in news articles using crowdsourcing. In *Proceedings of the 12th Language Resources and Evaluation Conference*. 1478–1484.

[24] Bing Liu et al. 2010. Sentiment analysis and subjectivity. *Handbook of natural language processing* 2, 2010 (2010), 627–666.

[25] Ping Liu, Karthik Shivaram, Aron Culotta, Matthew A Shapiro, and Mustafa Bilgic. 2021. The Interaction between Political Typology and Filter Bubbles in News Recommendation Algorithms. In *Proceedings of the Web Conference 2021*. 3791–3801.

[26] Katharina Ludwig, Alexander Grote, Andreea Iana, Mehwish Alam, Heiko Paulheim, Harald Sack, Christof Weinhardt, and Philipp Müller. 2022. Recommended Polarization? The (limited) effects of content- and sentiment-based news recommendation on affective, ideological, and perceived polarization. *72nd Annual ICA Conference, One World, One Network?* (2022).

[27] Laura Mascarell, Tatyana Ruzsics, Christian Schneebeli, Philippe Schlattner, Luca Campanella, Severin Klingler, and Cristina Kadar. 2021. Stance Detection in German News Articles. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*. Association for Computational Linguistics, Dominican Republic, 66–77. https://doi.org/10.18653/v1/2021.fever-1.8

[28] Philip May. 2020. *Cross English & German RoBERTa for Sentence Embeddings*. https://huggingface.co/T-Systems-onsite/cross-en-de-roberta-sentence-transformer

[29] Denis McQuail. 1992. *Media performance: Mass communication and the public interest*. Vol. 144. Sage London.

[30] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[31] Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiao-Dan Zhu, and Colin Cherry. 2016. A Dataset for Detecting Stance in Tweets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA).

[32] Eli Pariser. 2011. *The filter bubble: What the Internet is hiding from you*. Penguin UK.

[33] Evaggelia Pitoura, Kostas Stefanidis, and Georgia Koutrika. 2021. Fairness in Rankings and Recommendations: An Overview. *The VLDB Journal* (2021), 1–28.

[34] Kumar Ravi and Vadlamani Ravi. 2015. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-based systems* 89 (2015), 14–46.

[35] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 3982–3992.

[36] Myrthe Reuver, Antske Fokkens, Suzan Verberne, H Toivonen, and M Boggia. 2021. No NLP task should be an island: multi-disciplinary for diversity in news recommender systems. *Proceedings of the EACL Hackashop on news media content analysis and automated report generation* (2021), 45–55.

[37] Arjun Roy, Pavlos Fafalios, Asif Ekbal, Xiaofei Zhu, and Stefan Dietze. 2022. Exploiting stance hierarchies for cost-sensitive stance detection of Web documents. *Journal of Intelligent Information Systems* 58, 1 (2022), 1–19.

[38] Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*. Vol. 39. Cambridge University Press Cambridge.

[39] Timo Spinde, Felix Hamborg, and Bela Gipp. 2020. An integrated approach to detect media bias in german news articles. In *Proceedings of the ACM/IEEE joint conference on digital libraries in 2020*. 505–506.

[40] Natalie Jomini Stroud. 2017. Selective exposure theories. In *The Oxford handbook of political communication*.

[41] Cass R Sunstein. 2001. *Echo chambers: Bush v. Gore, impeachment, and beyond*. Princeton University Press Princeton, NJ.

[42] Jannis Vamvas and Rico Sennrich. 2020. X-Stance: A Multilingual Multi-Target Dataset for Stance Detection. *CoRR* abs/2003.08385 (2020). arXiv:2003.08385 https://arxiv.org/abs/2003.08385

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[44] Paul S Voakes, Jack Kapfer, and David Shano-yeon Chern. 1996. Diversity in the news: A conceptual and methodological framework. *Journalism & Mass Communication Quarterly* 73, 3 (1996), 582–593.

[45] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM* 57, 10 (2014), 78–85.

[46] Hongwei Wang, Fuzheng Zhang, Jialin Wang, Miao Zhao, Wenjie Li, Xing Xie, and Minyi Guo. 2018. Ripplenet: Propagating user preferences on the knowledge graph for recommender systems. In *Proceedings of the 27th ACM international conference on information and knowledge management*. 417–426.

[47] Ruifeng Xu, Yu Zhou, Dongyin Wu, Lin Gui, Jiachen Du, and Yun Xue. 2016. Overview of NLPCC Shared Task 4: Stance Detection in Chinese Microblogs. In *Natural Language Understanding and Intelligent Applications - 5th CCF Conference on Natural Language Processing and Chinese Computing, NLPCC 2016, and 24th International Conference on Computer Processing of Oriental Languages, ICCPOL 2016, Kunming, China, December 2-6, 2016, Proceedings (Lecture Notes in Computer Science, Vol. 10102)*. Springer, 907–916. https://doi.org/10.1007/978-3-319-50496-4_85

A USER STUDY AND DATA COLLECTION

The experimental setup is extensively described in [26]. The participants were recruited through the online-access panel of Respondi AG, and selected using a quote procedure to match German Internet users aged 18 to 74. The mean average age was 46.56, with a standard deviation of 15.68, 47.7% of the participants were females, and 38.5% of the respondents had an *Abitur* (German university entrance certificate) or higher qualification. Only participants with desktop-based devices were included given the display of the experimental conditions. Moreover, some participants were removed from the final user dataset if these did not meet the age range of the quota sample, spent less than 120 seconds or more than 2.5 hours with the recommender system, or provided very consistent answering patterns across all constructs, resulting in a final sample of 1,417 participants out of 1,801 initial ones.

The study consisted in a two-factorial experiment with an incomplete design. The first factor consisted in using one of three types of text-based news recommender systems (i.e. TF-IDF, Word2vec, Transformer) or a random recommendation. The second factor concerned the inclusion of sentiment in the recommender. Only the TF-IDF recommender was enriched with sentiment scores, and thus, ratings gathered based on it are over-represented in the data. More specifically, in the test data, there are 786 unique users with TF-IDF generated ratings, 211 with Word2vec-generated ratings, 209 with Transformer-generated ratings, and 211 with randomly generated ratings.

Each participant was firstly asked to complete a questionnaire including questions regarding polarization, political interests, attitudes towards refugees, and demographics. Afterwards, the subjects were asked to select the news that they found the most interesting from a list of six news previews, consisting of their titles and first lines of text. The participants would then read the news, and the process would be repeated four times. At the end of the experiment, the participants would be asked again to complete a follow-up questionnaire, in order to examine how the recommended news influenced their political polarization measured by the initial questionnaire.