Stance Detection in German News Articles

Laura Mascarell[1*] **Tatyana Ruzsics**[1*] **Christian Schneebeli**[1*]
Philippe Schlattner[1*] **Luca Campanella**[1*] **Severin Klingler**[1] **Cristina Kadar**[2]

[1] ETH Zurich, [2] Neue Zürcher Zeitung

{lmascarell,tatiana.ruzsics,christian.schneebeli,philippe.schlattner,
luca.campanella,severin.klingler}@inf.ethz.ch, cristina.kadar@nzz.ch

Abstract

The widespread use of the Internet and the rapid dissemination of information poses the challenge of identifying the veracity of its content. Stance detection, which is the task of predicting the position of a text in regard to a specific target (e.g. claim or debate question), has been used to determine the veracity of information in tasks such as rumor classification and fake news detection (Küçük and Can, 2020). While most of the work and available datasets for stance detection address short texts snippets extracted from textual dialogues, social media platforms, or news headlines with a strong focus on the English language, there is a lack of resources targeting long texts in other languages. Our contribution in this paper is twofold. First, we present a German dataset of debate questions and news articles that is manually annotated for stance and emotion detection. Second, we leverage the dataset to tackle the supervised task of classifying the stance of a news article with regards to a debate question and provide baseline models as a reference for future work on stance detection in German news articles.

1 Introduction

The automatic extraction and analysis of information from text is an important area in natural language processing that tackles a wide range of research problems such as sentiment analysis, veracity and rumor detection, emotion recognition, or fake news detection. With the rapid increase in online content and spread of misinformation, these research topics have gained attention, as it becomes crucial to identify the veracity of the information.

Stance detection is the task of predicting the position of a text (e.g. in favor or against) towards a specific target, which has shown to aid at identifying the veracity of rumors (rumor classification) and fake news (Küçük and Can, 2020). As most of

---

*Equal contribution

the stance detection benchmarks focus on short text snippets, especially tweets, and/or English (Mohammad et al., 2016a,b; Hanselowski et al., 2018), there is a lack of annotated data to address stance detection in longer texts and in other languages.

In this paper, we present the CHeeSE dataset, a collection of manually annotated Swiss (CH) news articles and debate questions in German for Stance and Emotion detection. Specifically, the CHeeSE dataset comprises 91 debate questions that are automatically paired with related news articles, resulting in a total of $3,693$ question and article pairs. A team of native German speakers manually annotated those pairs with the stance of the article towards the debate question, the article emotion, and the emotion of each individual paragraph. Table 1 shows an illustrative example in English of the annotations for a given question and article.

To the best of our knowledge, this is the first available dataset of German news articles annotated for this purpose. While the Fake News Challenge[1] provides document-level stance annotations for English news articles (Hanselowski et al., 2018), other datasets that consider the German language focus on shorter texts (Vamvas and Sennrich, 2020).

Finally, we perform target-specific stance detection as a supervised task using our annotated data. We then build baseline models as a reference in future work on stance detection in German news articles. Since the dataset provides both stance and emotion annotations, future work could assess the use of emotion information to improve stance detection in a multi-task setting (see discussion in Section 5). Moreover, computational journalism applications, such as news recommender systems, could leverage the dataset to mitigate the selective exposure of one-sided articles (Gao et al., 2018).

The CHeeSE dataset and the code to reproduce the baseline models are publicly available.[2]

---

[1] http://www.fakenewschallenge.org
[2] https://github.com/MTC-ETH/CHeeSE

	Example Content	Paragraph Emotion
Question	Should the Corona measures be more strict?	
Article Title	Corona measures insufficient	
Article Text	\<p\>It looks bad. The number of daily cases is increasing since weeks. There is no improvement in sight.\</p\>	Sadness
	\<p\>Again, the parliament failed to respond accordingly.\</p\>	Anger
	\<p\>Yet, our health system is still stable. Let us hope for the best.\</p\>	Trust
Article Emotion	Sadness	
Article Stance	In Favour	

Table 1: Illustrative example in English of the provided annotations for each article and question pairs in the CHeeSE dataset: the article's stance towards the question, the article's emotion and the emotion of each paragraph.

2 Related Work

Most of the available annotated datasets for stance detection focus on short texts such as tweets. The SemEval-2016 shared task hosts a dataset of annotated English tweets and their stance towards the targets 'Atheism', 'Climate Change', 'Feminism', 'Hillary Clinton', and 'Abortion' (Mohammad et al., 2016a). Aside from English, there are available tweets datasets for Catalan and Spanish (Zotova et al., 2020), French (Lai et al., 2020), or Italian (Cignarella et al., 2020). For German, the multilingual dataset X-stance offers a collection of political questions and their corresponding comments written by candidates of elections, partly in German, partly in French and Italian (Vamvas and Sennrich, 2020).

In the journalistic domain, there are two available datasets of English news articles. The Emergent dataset is a collection of claims and their associated news articles (Ferreira and Vlachos, 2016). Similarly, the Fake News Challange detection task (FNC-1) offers a dataset of news articles and their stance towards a headline (Hanselowski et al., 2018). While the Emergent dataset performs the stance detection task on single-sentence headlines towards a specific claim, the FNC-1 is a document-level stance detection task that considers the entire news article. In our dataset, we also address stance detection at document level, but in German text.

Emotion detection datasets are also mostly concerned with short text snippets, such as tweets (Mohammad et al., 2018), news headlines (Strapparava and Mihalcea, 2007), textual dialogues (Chatterjee et al., 2019), or sentences in fairy tales (Volkova et al., 2010). In contrast, our dataset offers article- and paragraph-level annotations for emotion detection in German news articles. Paragraph-level annotations could be leveraged to identify and analyse patterns of emotion sequences (i.e. emotion flow) in news articles. While emotion flow has been explored in books (Maharjan et al., 2018) and movie synopsis (Kar et al., 2018), such analysis has not yet been applied to journalistic content.

To improve stance detection, some research approaches focus on using sentiment information, and more specifically, the polarity of the text (i.e. positive, negative, or neutral). In fact, the annual workshop on Semantic Evaluation (SemEval) organised a stance detection shared task and provided a dataset of tweets also annotated with sentiment information (Mohammad et al., 2016a).[3] However, the conclusions on whether sentiment information aids at improving stance detection are mixed. Some studies report the benefits of using such information (Li and Caragea, 2019; Sun et al., 2019), whereas others claim that the tasks are independent of each other. For example, a text with a positive sentiment could show an opposing stance towards a specific claim (Mohammad et al., 2017; Aldayel and Magdy, 2019; Sobhani et al., 2016).

3 The CHeeSE Dataset

The CHeeSE dataset is a collection of $3,693$ pairs of debate questions and Swiss news articles in German that are manually annotated for stance and emotion detection. Specifically, the dataset comprises a total of $1,970$ different articles paired with up to six questions (~1.87 on average) from a list of 91 questions that we manually compiled to ensure a variety of debate topics (Section 3.1).

The data comes from $\sim$670k news articles provided by the Swiss media companies *Blick* ($\sim$536k articles),[4] *Neue Zürcher Zeitung* ($\sim$125k articles),[5] and *NZZ am Sonntag* ($\sim$9k articles).[6] The articles

---

[3] http://alt.qcri.org/semeval2016/task6/
[4] https://www.blick.ch
[5] https://www.nzz.ch
[6] https://nzzas.nzz.ch/

Negative Example	Issue	Preferred Formulation
Should legislation on environmental protection be relaxed to allow for the building and expansion of wind farms, solar power stations, and hydroelectric plants?	Topic too specific	Should state laws counteract climate change?
Should Saddam Hussein be executed?	Time and topic too specific	Should death penalty be legal?
Should Swiss law be changed to follow EU more restrictive firearms law?	Multiple topics	Should the firearm laws be stricter?
First case of transmission of Sars-CoV-2 identified in the uterus	News, no debate	–

Table 2: Negative examples in English of target questions and their preferred formulation.

Topic	Debate Question	Keyword Examples
Freizeit [Leisure]	Soll die Jagd als Freizeittätigkeit verboten werden? [Should hunting be banned as a leisure activity?]	Jagd, Tiere, Tierartenschutz [hunting, animals, species protection]
Digitalisierung [Digitalisation]	Verstärkt KI Vorurteile und Bias? [Does AI reinforce prejudices and bias?]	Künstliche Intelligenz, Vorurteile, Sexismus [Artificial intelligence, prejudice, sexism]
Gesellschaft [Society]	Sind Abtreibungen moralisch vertretbar? [Are abortions morally acceptable?]	Geburt, Abtreibung, Fötus [birth, abortion, fetus]
Gesellschaft [Society]	Soll die Adoption für alle ermöglicht werden? [Should adoption be possible for all?]	Adoption, Privatrecht, Gleichberechtigung [adoption, private law, equality]

Table 3: Examples of the selected debate questions to build the CHeeSE dataset and their assigned keywords.

were published between 2004 and 2020 and cover a set of 24 topics, such as science, environment, politics, religion, or society.

In this section, we describe the steps involved in the construction of the CHeeSE dataset. First, we compile a list of debate questions (Section 3.1) that we automatically pair with related articles using a two-pass ranking approach (see Section 3.2). Then, a team of native German speakers annotate the collected data with emotion and stance labels (Section 3.3). Finally, we apply a semi-automatic method to balance the stance label distribution of questions without articles in favor or against in the annotated dataset (Section 3.4).

3.1 Question Selection

We generate a list of debate questions that we later pair with news articles for the stance detection task. To do so, we organise a brainstorming session with a team of 12 participants of various backgrounds in the news media industry (e.g. journalists or data scientists). To ensure that the generated questions comprise a wide range of topics as per the IPTC topic taxonomy[7] (e.g. politics, environment, or sport), we divide the session in four iterations and four groups, where each group concerns a different

set of topics. In each iteration, we rotate the group members. As a reference, the participants receive a list of controversial issues from Wikipedia[8] and the ProCon debate platform[9] and a list of negative examples with their preferred formulations (see Table 2). The participants also assign a list of keywords to each debate question as additional data to represent the questions in the stance detection task. For example, a suitable list of keywords for the English question 'should hunting be banned as a leisure activity' is 'hunting', 'animals', and 'species protection'. At the end of the session, the participants generated a total of 125 questions with their corresponding keywords. We further filter the list and eliminate duplicates or similar questions, resulting in a final list of 91 debate questions (see examples in Table 3).

3.2 Question-Article Pairs Collection

To provide a relevant dataset for the stance detection task, we focus on collecting question and article pairs that are related. That is, the article is likely to show a stance (in favor, against, or discussing) with respect to the associated question. To do so, we implement a two-pass ranking approach that re-

---

[7] https://iptc.org/standards/newscodes

[8] https://en.wikipedia.org/wiki/Wikipedia:List_of_controversial_issues
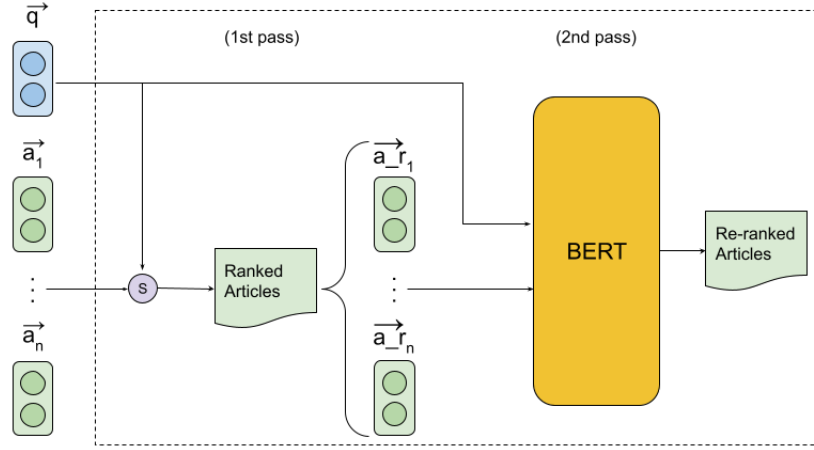
[9] https://www.procon.org

Figure 1: Two-pass ranking approach to collect related question and article pairs. For a given question and a set of news articles, in the first pass, our approach ranks the articles based on the cosine similarity scores of their tf-idf embeddings. In the second pass, it re-ranks the articles using a pre-trained German BERT model, selecting those above the probability threshold 0.75.

sembles the text ranking method used in Nogueira and Cho (2020).[10] Figure 1 illustrates our approach for a given question and the set of news articles.

The method works as follows. For each of the selected questions in section 3.1, we use our two-pass ranking approach to identify its related articles. Specifically, in the first pass, we rank the articles based on the cosine similarity scores of the tf-idf embeddings of the question and each article. Here, we represent a question as the string concatenation of the question itself and the set of its assigned keywords (see Section 3.1) and an article as the concatenation of its title and body.

In the second pass, we score and re-rank the articles according to their probability of having any stance towards the debate question (i.e. the article is in favour, against, or discussing it). As a re-ranker, we use German BERT (Devlin et al., 2019),[11] a general purpose natural language processing model for German text that we fine-tune on our re-ranking task as in Nogueira and Cho (2020).

The fine-tuning data is a set of 275 question and article pairs that we manually annotated as 'related', if the article has any stance with respect to the question, or 'unrelated', otherwise. To obtain those pairs, we first selected 11 questions from different topics through stratified sampling. For each question, we then applied the first pass of our ranking approach to rank the articles based on the

cosine similarity of their tf-idf embeddings, and kept the 25 most similar articles. Note that these pairs are only used for fine-tuning and they are not part of the final CHeeSE dataset. We fine-tune the BERT model on this data for three epochs and a learning rate of 3e-05.

The BERT architecture requires a token sequence as an input. To represent a given question and article as such input sequence, we follow the method described in Devlin et al. (2019). We concatenate a question-article pair with the special token [SEP] and insert the special classification token [CLS] at the beginning of the sequence, which aggregates the features from both question and article. We truncate the input to the maximum length of 512 tokens that BERT supports.

After re-ranking the articles with BERT, we keep the highest-confidence matches. To obtain a threshold, we apply 10-fold cross-validation on the fine-tuning data. The resulting threshold of 0.75 is the average cutoff value of $(tpr + \frac{1}{fpr})/2$, where tpr and fpr are the true and false positive rates, respectively, under the ROC curve.

Our two-pass ranking approach results in a dataset of $2,096$ question and article pairs. Each article is associated with up to two questions.

Dataset extension Since manual annotations are expensive and time-consuming, we extend the collected dataset keeping in mind the annotation efforts. Articles are longer texts than questions, therefore, it is easier to annotate a new question-article pair, where only the question is new to the annota-

---

[10] We refer to Yates et al. (2021) for a comprehensive overview of text ranking approaches.

[11] https://huggingface.co/dbmdz/bert-base-german-uncased

tor. Thus, we extend the dataset by pairing more questions with the collected articles.

To do so, for each collected article after the two-pass ranking approach, we rank all questions using the German BERT classifier as in the second pass. Here, we exclude those questions that were already paired with the article. As a result, we get $1,597$ additional pairs, increasing the number of questions per article from two to five.

In the final dataset, we distinguish between the data collected after the two-pass ranking approach and this additional data, labelling them as 'two-pass' and 'two-pass-ext', respectively.

3.3 Manual Annotation Task

We design a task to manually annotate the collected question and article pairs (Section 3.2). For each pair, the annotation task is twofold: (a) annotate the stance of the article towards the question; (b) annotate the emotions of the news article.

To perform this annotation task, we recruited a team of 23 native German speakers and used a crowd-sourcing labelling tool that was implemented for this purpose (see Appendix A). In particular, the annotators, 15 women and 8 men, are university students or academics between 21 and 49 and a median of 25 years old. To ensure that each crowd-worker understood the annotation task, we closely monitored them during the labelling of their first 50 articles and notified them if they did not follow the annotation instructions as intended.

Stance Annotation To label the stance of an article towards a question, we consider the German counterpart of the English 'in favor', 'against', 'discussion', and 'unrelated', where 'discussion' indicates that the article discusses the topic of the question without taking a position, as in the FNC-1 task (Hanselowski et al., 2018). Table 4 shows an example from the CHeeSE dataset of the different stances with regards to the question 'are gene manipulations morally justifiable?'. In this table, we represent the articles with their snippet for illustrative purposes only. Note that identifying a pair as 'unrelated' is harder in our dataset than in FNC-1. The reason is that the unrelated pairs in the FNC-1 dataset are randomly matched, whereas we collect all pairs using our ranking approach, ensuring that each pair is among the most related according to their tf-idf embeddings similarity and BERT classification scores (Section 3.2).

Formally, given an article a, a question q, and the set of German stance labels *Ja, dafür* ('in favour'), *Diskutierend* ('discussion'), *Nein, dagegen* ('against'), and *Kein Bezug* ('unrelated') $\{s_1, \ldots, s_4\}$, the task is to identify the article's a stance towards the question q,

$$\text{stance}(a,q) \quad \in \{s_1, \ldots s_4\}$$

During the annotation task, three different annotators label the stance of each question-article pair. This annotation process results in 86% of question-article pairs, where at least two annotators agree on the stance. We label the remaining pairs with no annotator agreement as the German *Unklar* ('unclear'). The final inter-rater agreement (IRA) and the Fleiss' κ coefficient of inter-rater reliability (Fleiss, 1971) are 0.52 and 0.33, respectively.

Emotion Annotation The annotation task is designed to identify the emotion that is mostly represented by the text. We consider a set of eight emotions, which are the German counterpart of the basic emotions defined in Plutchik (1980)'s model: *Freude* ('joy'), *Vertrauen* ('trust'), *Angst* ('fear'), *Antizipation* ('anticipation'), *Traurigkeit* ('sadness'), *Ekel* ('disgust'), *Ärger* ('anger'), and *Überraschung* ('surprise'). The 'no emotion' label is defined as the German *Keine*.

This is a challenging annotation task for two main reasons, which negatively impact the inter-rater agreement. First, there are eight different emotion labels to choose from. Second, the text can convey an in-between emotion, also referred as 'dyads' (Plutchik, 2001). For example, 'hope', which is considered a complex emotion, is a combination of 'anticipation' and 'trust'. Indeed, we observe this phenomenon, when manually analysing the most frequent disagreeing pairs of emotions (e.g. anticipation together with trust or anger).

To mitigate this issue, we extend the annotation task as follows. We encourage the annotation of one emotion per paragraph and article and, in case of doubt, the annotators assign the emotion that the text conveys the most as primary emotion and the rest as secondary emotions. Annotators could select multiple secondary emotions. On average, they selected 0.78 and 0.43 secondary emotions at article and paragraph level, respectively.

Formally, we define a primary emotion e_p for a given text t as one of the eight emotions labels $e_p(t) \in \{e_1, \ldots e_8\}$, and the set of secondary emotions e_s as the corresponding subset of labels $e_s(t) \subset \{e_1, \ldots e_8\}$. Given an article a, which is composed of a set of paragraphs $a = \{p_1, \ldots, p_n\}$, the task

	Sind Genmanupulationen moralisch vertretbar?
	[Are gene manipulations morally justifiable?]
Diskutierend	Ein Tabu fällt: Grossbritannien erlaubt es Forschern, die Gene von Embryos zu manipulieren. Die wichtigsten Fragen und Antworten.
[Discussion]	[A taboo has been lifted: The UK allows researchers to manipulate the genes of embryos. The most important questions and answers.]
Ja, dafür	Mäuse, die mit dem Aidserreger infiziert sind, werden virusfrei. Diesen Erfolg haben Forscher mit Medikamenten und der Genschere Crispr/Cas erzielt.
[In favour]	[Mice infected with the AIDS virus become virus-free. Researchers have achieved this success with drugs and the gene scissors Crispr/Cas.]
Nein, dagegen	Ein chinesischer Wissenschaftler verkündet die Geburt zweier genmanipulierter Babys. Das sorgt weltweit für Entsetzen.
[Against]	[A Chinese scientist announces the birth of two genetically engineered babies. This is causing worldwide horror.]
Kein Bezug	Das Personal in Spitälern wird besser geschult, um zu verhindern, dass Patienten Medikamente in falscher Dosierung oder gar nicht erhalten.
[Unrelated]	[Hospital staff will be better trained to prevent patients from receiving the wrong dose of medication or not receiving it at all.]

Table 4: Example from the CHeeSE dataset of the four different stance annotations 'Diskutierend', 'Ja, dafür', 'Nein, dagegen', and 'Kein Bezug' with regards to the same question. The table only contains the article's snippet.

consists of (1) identifying the primary overall article emotion and secondary emotions, if any, or 'no emotion' $\neg e$, otherwise; (2) identifying the primary and secondary emotions of each paragraph, if any, or 'no emotion', otherwise.

emotion$(a) \in \{e_p(a), \{e_p(a)\} \cup e_s(a), \neg e\}$,
emotion$(p_1) \in \{e_p(p_1), \{e_p(p_1)\} \cup e_s(p_1), \neg e\}$,
...
emotion$(p_n) \in \{e_p(p_n), \{e_p(p_n)\} \cup e_s(p_n), \neg e\}$

As in the stance annotation task, three different annotators label each article and paragraph. The final list of confirmed labels are those primary labels with annotator agreement. That is, those primary emotion labels that are also used as primary or secondary emotion by a different annotator. This results in 84% of the articles and 78% of the paragraphs with at least one confirmed label. We compute the IRA according to this definition as follows. For every article, we check each combination of two annotators (i.e. three pairs for three annotators). For each pair of annotators, we count them as agreement (1) if the annotators agree on the primary emotion or the primary emotion of one of the annotators matches one of the secondary emotions of the other. Otherwise, we count them as disagreement (0). We then average over all annotator pairs of all articles. The IRA of the article and paragraph annotations are 0.50 and 0.42, respectively.[12] We then annotate the rest of the articles and paragraphs without agreement as *Unklar* ('unclear').

Overall, there are 76 question and article pairs in the CHeeSE dataset, where both stance and article emotion are annotated as *Unklar*. In contrast, there are 2,588 question-article pairs, where both stance and article emotion have a confirmed label.

3.4 Balancing the Stance Class Distribution

We observe that some of the questions from the annotated data have an imbalanced stance class distribution. In particular, 17 questions have no articles in favor and 40 have no articles against (see examples in Table 5). To balance the distribution, we perform a semi-automatic method to retrieve additional articles that have a positive and negative stance towards those questions.

Our semi-automatic method works as follows. For each of the imbalanced questions, we first search for related debates on discussion platforms where users can argument in favor or against different topics, such as DebateHub.net,[13] using the question itself or the list of defined keywords (Section 3.1). Next, we group the arguments for each stance, resulting in two texts, one with arguments in favor t_{for} and another with arguments against t_{agt}. Assuming that texts sharing the same position have similar embeddings, we use their embeddings to retrieve similar articles from our data (i.e. articles that share the same stance). Specifically, we obtain the embeddings of t_{for} and t_{agt} and, for each embedding, we rank our articles based on their cosine similarity. Since our data and the arguments are in

---

[12]The IRA values include the articles labelled after balancing the stance class distribution (Section 3.4)

[13]https://debatehub.net

Question examples without articles in favor
Soll Werbung für Kinder verboten werden? [Should advertising to children be banned?]
Soll die 32 Stunden Arbeitswoche eingeführt werden? [Should the 32-hour work week be introduced?]
Ist ein Krieg gegen ein anderes Land gerechtfertigt? [Is war against another country justified?]

Questions examples without articles against
Sollte Sterbehilfe legal sein? [Should euthanasia be legal?]
Sollen Plastikverpackungen verboten werden? [Should plastic packaging be banned?]
Sollen Obdachlose staatlich subventioniert werden? [Should homeless people be subsidised by the state?]

Table 5: Examples of questions with no articles in favor or against after the two-pass ranking approach.

different languages (mostly German and English), we use a pre-trained multilingual BERT model to obtain the embeddings.[14] Finally, we manually go through the most similar articles in each ranking and keep those that share the corresponding stance.

Our semi-automatic method reduces the number of questions with no articles against and in favour from 40 and 17 to 21 and 9, respectively, resulting in 116 additional question and article pairs in the final dataset. Table 6 shows the final distribution of the stance classes in the CHeeSE dataset. In addition, the table shows the total number of question and article pairs without annotation agreement on the stance annotation task (i.e. *Unklar* label).

4   Stance Detection Task

In the stance setting, the CHeeSE dataset provides with (a) a natural-language question and its set of related keywords; (b) a news article; (c) the stance of the article towards the question. Similar to the stance detection task defined in the FNC-1, we use the stance classes *Ja, dafür* ('for'), *Nein, dagegen* ('against'), *Diskutierend* ('discussion'), and *Kein Bezug* ('unrelated'). The target-specific stance detection task is then a multiclass classification task that consists of predicting an article's position towards its associated question.

We assess different baselines on the task using the CHeeSE dataset as a reference for future work. The proposed baselines are a bag-of-words linear classifier and a BERT-based model (Devlin et al.,

Label	Two-pass	Two-pass-ext	All
Diskutierend	620	154	774
Ja, dafür	509	193	702
Nein, dagegen	232	54	286
Kein Bezug	403	1,025	1,428
Unklar	332	171	503

Table 6: Stance class distribution in the final dataset ('All' column), consisting of the data collected after the two-pass ranking approach ('Two-pass') and its extension ('Two-pass-ext'). The question-article pairs with the *Unklar* ('unclear') label are those without annotation agreement.

2019), a multi-layer bidirectional Transformer encoder (Vaswani et al., 2017), which has shown to achieve the state-of-the-art performance on a wide range of natural language processing tasks (Radford and Narasimhan, 2018; Wu and Dredze, 2019; Peters et al., 2018). We report their class-wise F_1 and macro-averaged F_1 score, which is not affected by the majority class (Hanselowski et al., 2018).

Since the amount of training data is relatively small compared to other similar datasets, such as X-stance or FNC-1 (see Table 8), we perform five-fold cross-validation on the CHeeSE dataset, keeping the same class distribution in each fold as in the entire dataset. In our experiments, we exclude the instances annotated as *Unklar* ('unclear'), which are those with no annotator agreement.

4.1   Baseline Experiments

We evaluate two baselines: a bag-of-words linear classifier and a BERT-based model, and report their F_1 scores on the CHeeSE dataset.

Bag-of-Words Linear Classifier   Our first baseline is a bag-of-words linear classifier built using fastText[15] as described in Joulin et al. (2017). We train the model for 50 epochs and a learning rate of 1 as recommended in the standard fastText settings.

Fine-tuned German BERT model   As a second baseline, we fine-tune a pre-trained BERT model. Specifically, we first adapt the pre-trained German BERT model bert-base-german-uncased from Hugginface (Wolf et al., 2020)[16] to the stance detection task and extend it with a multiclass classification head. We then fine-tune the model on our data for four epochs and a learning rate of 3e-05.

Model	F_1 micro	F_1 macro	F_1 in favour	F_1 against	F_1 discussion	F_1 unrelated
Bag of words	50.8	41.9	42.8	20.5	39.9	64.3
German BERT	67.6	58.4	59.1	35.8	55.3	83.3

Table 7: Micro-averaged F_1, macro-averaged F_1 and class-wise F_1 scores for each baseline on the stance detection task. Fine-tuning the pre-trained German BERT model on the CHeeSE data outperforms the bag-of-words linear classifier.

Datasets	Lang	Training	Classification	Unrelated samples	Avg. tokens	F_1 macro
X-stance	multiple	∼46k	binary	n/a	61	76.8
FNC-1	English	∼61k	multiclass (4)	randomly selected	432	72.8
CHeeSE	German	∼3k	multiclass (4)	two-pass ranking	843	58.2

Table 8: Overview of different stance detection benchmarks built using pre-trained BERT models on the X-stance (Vamvas and Sennrich, 2020), FNC-1 (Slovikovskaya and Attardi, 2020), and CHeeSE datasets. The benchmarks are not directly comparable as there are differences on the training sizes of the datasets, the classification task (multiclass or binary), or the nature of the unrelated samples. The 'avg. tokens' column refers to the average token length of the stance target (question or headline) and body text pairs.

Results Table 7 shows the macro-averaged and class-wise F_1 scores of the bag-of-words and the fine-tuned German BERT baselines. The results show that the pre-trained German BERT model fine-tuned on the CHeeSE data performs consistently better than the bag-of-words linear classifier. We also observe that the unrelated stance class is the easiest to correctly identify. Note that the unrelated annotations represent about 40% of all stance annotations in our dataset (see Table 6).

BERT-like models have shown to perform well in other stance detection tasks. Table 8 lists different supervised stance detection benchmarks built using BERT models. Note that there are differences on the total of predicted labels in the classification task (binary or multiclass), the nature of the unrelated samples as discussed in Section 3.3, and the training sizes; in fact, the training sizes of the X-stance and FNC-1 are considerably higher than the CHeeSE dataset. Therefore, we provide the scores as a reference rather than for comparative purposes.

5 Discussion and Future Work

The emotion of a text is closely related to sentiment analysis. Zhang et al. (2020), for example, leverage semantic-emotion knowledge in a cross-target stance detection task with positive results. However, there is insufficient research on combining emotion and stance information; perhaps due to the lack of datasets with both annotations. Since the CHeeSE dataset is annotated for both stance

and emotion detection tasks, it allows the research community to further experiment in this direction.

In our experiments, we fine-tune a German pre-trained BERT model on our dataset. BERT-like models are able to deal with input sequences of up to 512 tokens. Since our dataset consists of news articles that are on average 843 and up to 4,130 tokens long, it would be interesting to assess the impact on stance detection when considering longer texts. Therefore, in future work we plan to experiment with model architectures that handle long input sequences, such as the Longformer, which processes sequences up to 4,096 tokens long and even up to 16k with the current GPUs (Beltagy et al., 2020).

The size of the CHeeSE dataset is considerably smaller than other stance datasets as shown in Table 8. Specifically, our dataset contains ∼3k training instances compared to, for example, the ∼61k from the FNC-1. More advanced models than our baselines could benefit from combining multiple datasets to achieve a better performance. For example, in multi-task learning, a model is simultaneously trained on related tasks, such that it learns to better generalise on the original task (Schiller et al., 2021; Hardalov et al., 2021).

We also expect the dataset to be a valuable resource in computational journalism, for example, to generate diverse news article portfolios or to provide a balanced exposure of articles' positions with recommender systems.

6 Conclusion

We presented the CHeeSE dataset, a manually annotated dataset of Swiss news articles. In contrast to most of the available datasets, which focus on short texts and/or the English language, the novelty of our dataset is that it provides document-level annotations on German news articles. Specifically, the dataset consists of a collection of German news articles automatically paired with debate questions. A team of native German speakers labelled each article and question pair with the stance of the article towards the question, the article emotion, and the emotion of each article's paragraph. We then performed target-specific stance detection as a supervised task using the CHeeSE dataset and provide baseline models as a reference in future research on stance detection in German news articles.

Acknowledgements

References

Abeer Aldayel and Walid Magdy. 2019. Assessing sentiment of the expressed stance on social media. In *Proceedings of the 2019 International Social Informatics Conference (SocInfo)*, pages 277–286, Doha, Qatar. Springer.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019. SemEval-2019 task 3: EmoContext contextual emotion detection in text. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 39–48, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

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 2020 Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA)*, Online. CEUR-WS.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1163–1168, San Diego, California. Association for Computational Linguistics.

J.L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Mingkun Gao, Hyo Jin Do, and Wai-Tat Fu. 2018. Burst your bubble! an intelligent system for improving awareness of diverse social opinions. In *Proceedings of the 2018 International Conference on Intelligent User Interfaces*, pages 371–383, Tokyo, Japan. Association for Computing Machinery.

Andreas Hanselowski, Avinesh PVS, 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*, pages 1859–1874, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2021. Cross-domain label-adaptive stance detection. *arXiv preprint arXiv:2104.07467*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Sudipta Kar, Suraj Maharjan, A. Pastor López-Monroy, and Thamar Solorio. 2018. MPST: A corpus of movie plot synopses with tags. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *Association for Computing Machinery*, 53(1).

Mirko Lai, Alessandra Teresa Cignarella, Delia Irazú Hernanández Farías, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2020. Multilingual stance detection in social media political debates. *Computer Speech and Language*, 63:101075.

Yingjie Li and Cornelia Caragea. 2019. Multi-task stance detection with sentiment and stance lexicons. 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), pages 6299–6305, Hong Kong, China. Association for Computational Linguistics.

Suraj Maharjan, Sudipta Kar, Manuel Montes-y-Gómez, Fabio A. González, and Thamar Solorio. 2018. Letting emotions flow: Success prediction by modeling the flow of emotions in books. *arXiv preprint arXiv:1805.0974*.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016a. A dataset for detecting stance in tweets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3945–3952, Portorož, Slovenia. European Language Resources Association (ELRA).

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016b. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. *ACM Transactions on Internet Technology*, 17(3).

Rodrigo Nogueira and Kyunghyun Cho. 2020. Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Robert. Plutchik. 1980. *Emotion, a psychoevolutionary synthesis*. Harper & Row, New York, NY, USA.

Robert Plutchik. 2001. The Nature of Emotions. *American Scientist*, 89(4):344.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Stance detection benchmark: How robust is your stance detection? *KI-Künstliche Intelligenz*, pages 1–13.

Valeriya Slovikovskaya and Giuseppe Attardi. 2020. Transfer learning from transformers to fake news challenge stance detection (FNC-1) task. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1211–1218, Marseille, France. European Language Resources Association.

Parinaz Sobhani, Saif Mohammad, and Svetlana Kiritchenko. 2016. Detecting stance in tweets and analyzing its interaction with sentiment. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 159–169, Berlin, Germany. Association for Computational Linguistics.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, Prague, Czech Republic. Association for Computational Linguistics.

Qingying Sun, Zhongqing Wang, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2019. Stance detection via sentiment information and neural network model. *Frontiers of Computer Science*, 13(1):127.

Jannis Vamvas and Rico Sennrich. 2020. X-Stance: A multilingual multi-target dataset for stance detection. In *Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS)*, Zurich, Switzerland.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, page 6000–6010, Long Beach, California, USA. Curran Associates, Inc.

Ekaterina P. Volkova, Betty Mohler, Detmar Meurers, Dale Gerdemann, and Heinrich H. Bülthoff. 2010. Emotional perception of fairy tales: Achieving agreement in emotion annotation of text. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 98–106, Los Angeles, CA. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. 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), pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Andrew Yates, Rodrigo Nogueira, and Jimmy Lin. 2021. Pretrained transformers for text ranking: BERT and beyond. In *Proceedings of the 2021 ACM International Conference on Web Search and Data Mining (WSDM)*, page 1154–1156, Online. Association for Computing Machinery.

Bowen Zhang, Min Yang, Xutao Li, Yunming Ye, Xiaofei Xu, and Kuai Dai. 2020. Enhancing cross-target stance detection with transferable semantic-emotion knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3188–3197, Online. Association for Computational Linguistics.

Elena Zotova, Rodrigo Agerri, Manuel Nuñez, and German Rigau. 2020. Multilingual stance detection in tweets: The Catalonia independence corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1368–1375, Marseille, France. European Language Resources Association.

A   Labelling Tool

Figure 2 shows an example of the labelling tool that we use to annotate each article and question pairs.

Homopaare können weiterhin nicht adoptieren

BERN - Gleichgeschlechtliche Paare und homosexuelle Einzelpersonen können theoretisch Kinder adoptieren, nicht aber homosexuelle Paare, die ihre Partnerschaft eintragen liessen.

Absatz

Der Nationalrat hat es am Freitag abgelehnt, diese absurde Situation zu ändern.

Mit 97 zu 83 Stimmen und 8 Enthaltungen weigerte sich der Rat, einer Petition des Vereins Familienchancen Folge zu geben. Die Eingabe war im Juni 2010 mit knapp 20000 Unterschriften eingereicht worden und hatte die Unterstützung vor allem der Linken und einiger Freisinniger.

Die Petenten verlangten, dass gleichgeschlechtliche Paare in eingetragener Partnerschaft in Sachen Eltern- und Adoptionsrechte mit heterosexuellen Ehepaaren gleichgestellt werden. Zivilstand und die sexuelle Orientierung der Adoptionswilligen dürften keine Rolle spielen. Auch Kinder von gleichgeschlechtlichen Paaren sollten gleiche Rechte haben wie Kinder von Ehepaaren.

Emotion

Primäre Emotion

Freude	Vertrauen	Antizipation
Angst	Traurigkeit	Ekel
Ärger	Überraschung	Keine

☐ Ich bin mir meiner Antwort nicht sicher

Optional: Sekundäre Emotionen

Freude	Vertrauen	Antizipation
Angst	Traurigkeit	Ekel
Ärger	Überraschung	Keine

☐ Ich bin mir meiner Antwort nicht sicher

Primäre Emotion

Freude	Vertrauen	Antizipation
Angst	Traurigkeit	Ekel
Ärger	Überraschung	Keine

☐ Ich bin mir meiner Antwort nicht sicher

Optional: Sekundäre Emotionen

Freude	Vertrauen	Antizipation
Angst	Traurigkeit	Ekel
Ärger	Überraschung	Keine

☐ Ich bin mir meiner Antwort nicht sicher

Emotion des gesamten Artikels

Welche Emotion wird vom gesamten Text am Stärksten vermittelt?

Primäre Emotion

Freude	Vertrauen	Antizipation
Angst	Traurigkeit	Ekel
Ärger	Überraschung	Keine

☐ Ich bin mir meiner Antwort nicht sicher

Optional: Sekundäre Emotionen

Freude	Vertrauen	Antizipation
Angst	Traurigkeit	Ekel
Ärger	Überraschung	Keine

☐ Ich bin mir meiner Antwort nicht sicher

Standpunkt des Artikels

Bitte wählen Sie die Position aus, die der Artikel bezüglich den folgende(n) Frage(n) einnimmt:

Soll die Adoption für alle ermöglicht werden?

| 👍 Ja, dafür | ✊ Diskutierend | 👎 Nein, dagegen | !? Kein Bezug |

☐ Ich bin mir meiner Antwort nicht sicher

Figure 2: Example of the interface of the labelling tool that we use to annotate the CHeeSE dataset. For each article and question pair, the annotators label the emotion of each article's paragraph, the emotion of the overall article, and the stance of a debate question towards the article.