NewsMTSC: A Dataset for (Multi-)Target-dependent Sentiment Classification in Political News Articles

Felix Hamborg[1,2] **Karsten Donnay**[3,2]

[1]Dept. of Computer and Information Science, University of Konstanz, Germany
[2]Heidelberg Academy of Sciences and Humanities, Germany
[3]Dept. of Political Science, University of Zurich, Switzerland
felix.hamborg@uni-konstanz.de donnay@ipz.uzh.ch

Abstract

Previous research on target-dependent sentiment classification (TSC) has mostly focused on reviews, social media, and other domains where authors tend to express sentiment explicitly. In this paper, we investigate TSC in news articles, a much less researched TSC domain despite the importance of news as an essential information source in individual and societal decision making. We introduce NewsMTSC, a high-quality dataset for TSC on news articles with key differences compared to established TSC datasets, including, for example, different means to express sentiment, longer texts, and a second test-set to measure the influence of multi-target sentences. We also propose a model that uses a BiGRU to interact with multiple embeddings, e.g., from a language model and external knowledge sources. The proposed model improves the performance of the prior state-of-the-art from $F1_m = 81.7$ to 83.1 (real-world sentiment distribution) and from $F1_m = 81.2$ to 82.5 (multi-target sentences).

1 Introduction

Previous research on target-dependent sentiment classification (TSC, also called aspect-term sentiment classification) and related tasks, e.g., aspect-based sentiment classification (ABSC, or aspect-category sentiment classification) and stance detection, has focused mostly on domains in which authors tend to express their opinions explicitly, such as reviews and social media (Pontiki et al., 2015; Nakov et al., 2016; Rosenthal et al., 2017; Zhang et al., 2020; Hosseinia et al., 2020; AlDayel and Magdy, 2020; Liu, 2012; Zhang et al., 2018).

We investigate TSC in news articles – a much less researched domain despite its critical relevance, especially in times of "fake news," echo chambers, and news ownership centralization (Hamborg et al., 2019). How persons are portrayed in news on political topics is very relevant, e.g., for individual and societal opinion formation (Bernhardt et al., 2008).

We define our problem statement as follows: we seek to detect polar judgments towards target persons (Steinberger et al., 2017). Following the TSC literature, we include only in-text, specifically in-sentence, means to express sentiment. In news texts such means are, e.g., word choice and generally framing[1] (Kahneman and Tversky, 1984; Entman, 2007), e.g., "freedom fighters" vs. "terrorists," or describing actions performed by the target, and indirect sentiment through quoting another person (Steinberger et al., 2017). Other means may also alter the perception of persons and topics in the news, but are not in the scope of the task (Balahur et al., 2010), e.g., because they are not on sentence-level. For example, story selection, source selection, article's placement and size (Hamborg et al., 2019), and epistemological bias (Recasens et al., 2013).

The main contributions of this paper are: (1) We introduce *NewsMTSC*, a large, manually annotated dataset for TSC in political news articles. We analyze the quality and characteristics of the dataset using an on-site, expert annotation. Because of its fundamentally different characteristics compared to previous TSC datasets, e.g., as to sentiment expressions and text lengths, NewsMTSC represents a challenging novel dataset for the TSC task. (2) We propose a neural model that improves TSC performance compared to prior state-of-the-art models. Additionally, our model yields competitive performance on established TSC datasets. (3) We perform an extensive evaluation and ablation study of the proposed model. Among others, we investigate the recently claimed "degeneration" of TSC to sequence-level classification (Jiang et al., 2019)

[1]Political frames determine what a causal agent does with which benefit and cost (Entman, 2007). They are defined on a higher level than semantic frames (Fillmore and Baker, 2001).

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 1663–1675
April 19 - 23, 2021.

finding a performance drop in all models when comparing single- and multi-target sentences.

In a previous short-paper, we explored the characteristics of how sentiment is expressed in news articles by creating and analyzing a small-scale TSC dataset (Hamborg et al., 2021). The paper at hand addresses our former exploratory work's critical findings, including essential improvements to the dataset. Key differences and improvements are as follows. We significantly increase the dataset's size and the number of annotators per example and address its class imbalance. Further, we devise annotation instructions specifically created to capture a broad spectrum of sentiment expressions specific to news articles. In contrast, the early dataset misses the more implicit sentiment expressions commonly used by news authors (Hamborg et al., 2021; Steinberger et al., 2017). Also, we comprehensively test various consolidation strategies and conduct an expert annotation to validate the dataset.

We provide the dataset and code to reproduce our experiments at:
https://github.com/fhamborg/NewsMTSC

2 Related Work

Analogously to other NLP tasks, the TSC task has recently seen a significant performance leap due to the rise of language models (Devlin et al., 2019). Pre-BERT approaches yield up to $F1_m = 63.3$ on the SemEval 2014 Twitter set (Kiritchenko et al., 2014). They employ traditional machine learning combining hand-crafted sentiment dictionaries, such as SentiWordNet (Baccianella et al., 2010), and other linguistic features (Biber and Finegan, 1989). On the same dataset, vanilla BERT (also called BERT-SPC) yields 73.6 (Devlin et al., 2019; Zeng et al., 2019). Specialized downstream architectures improve performance further, e.g., LCF-BERT yields 75.8 (Zeng et al., 2019).

The vast majority of recently proposed TSC approaches employ BERT and focus on devising specialized down-stream architectures (Sun et al., 2019a; Zeng et al., 2019; Song et al., 2019). More recently, to improve performance further, additional measures have been proposed. For example, domain adaption of BERT, i.e., domain-specific language model finetuning prior to the TSC finetuning (Rietzler et al., 2019; Du et al., 2020); use of external knowledge, such as sentiment or emotion dictionaries (Hosseinia et al., 2020; Zhang et al., 2020), rule-based sentiment systems (Hosseinia et al., 2020), and knowledge graphs (Ghosal et al., 2020); use of all mentions of a target and/or related targets in a document (Chen et al., 2020); and explicit encoding of syntactic information (Phan and Ogunbona, 2020; Yin et al., 2020).

To train and evaluate recent TSC approaches, three datasets are commonly used: Twitter (Nakov et al., 2013, 2016; Rosenthal et al., 2017), Laptop and Restaurant (Pontiki et al., 2014, 2015). These and other TSC datasets (Pang and Lee, 2005) suffer from at least one of the following shortcomings. First, implicitly or indirectly expressed sentiment is rare in them. In their domains, e.g., social media and reviews, typically authors explicitly express their sentiment regarding a target (Zhang et al., 2018). Second, they largely neglect that a text may contain coreferential mentions of the target or mentions of different concepts (with potentially different polarities), respectively (Jiang et al., 2019).

Texts in news articles differ from reviews and social media in that news authors typically do not express sentiment toward a target explicitly (exceptions include opinion pieces and columns). Instead, journalists implicitly or indirectly express sentiment (Section 1) because language in news is typically expected to be neutral and journalists to be objective (Balahur et al., 2010; Godbole et al., 2007; Hamborg et al., 2019).

Our problem statement (Section 1) is largely identical to prior news TSC literature (Steinberger et al., 2017; Balahur et al., 2010) with key differences: we do not generally discard the "author-" and "reader-level." Doing so would neglect large parts of sentiment expressions. Thus, it would degrade real-world performance of the resulting dataset and models trained on it. For example, word choice (listed as "author-level" and discarded from their problem statement) is in our view an in-text means that may in fact strongly influence how readers perceive a target, e.g., "freedom fighters" or "terrorists." While we do not exclude their "reader-level," we do seek to exclude polarizing or contentious cases, where no uniform answer can be found in a set of randomly selected readers (Sections 3.3 and 3.4). As a consequence, we generally do not distinguish between the three levels of sentiment ("author," "reader," and "text") in this paper.

Previous news TSC approaches mostly employ sentiment dictionaries, e.g., created manually (Balahur et al., 2010; Steinberger et al., 2017) or extended semi-automatically (Godbole et al., 2007),

but yield poor or even "useless" (Steinberger et al., 2017) performances. To our knowledge, there exist two datasets for evaluation of news TSC methods (Steinberger et al., 2017), which – perhaps due to its small size ($N = 1274$) – has not been used or tested in recent TSC literature. Recently, Hamborg et al. (2021) proposed a dataset ($N = 3002$) used to explore target-dependent sentiment in news articles. The dataset suffers from various shortcomings, particularly its small size, class imbalance, and lacking the more ambiguous and implicit types of sentiment expressions described above. Another dataset contains quotes extracted from news articles, since quotes more likely contain explicit sentiment ($N = 1592$) (Balahur et al., 2010).

3 NewsMTSC: Dataset Creation

In creating the dataset, we rely on best practices reported in literature on the creation of datasets for NLP (Pustejovsky and Stubbs, 2012), especially for the TSC task (Rosenthal et al., 2017). Compared to previous TSC datasets though, the nature of sentiment in news articles requires key changes, especially in the annotation instructions and consolidation of answers (Steinberger et al., 2017).

3.1 Data sources

We use two datasets as sources: POLUSA (Gebhard and Hamborg, 2020) and Bias Flipper 2018 (BF18) (Chen et al., 2018). Both satisfy five criteria that are important to our problem. First, they contain news articles reporting on political topics. Second, they approximately match the online media landscape as perceived by an average US news consumer.[2] Third, they have a high diversity in topics due to the number of articles contained and time frames covered (POLUSA: 0.9M articles published between Jan. 2017 and Aug. 2019, BF18: 6447 articles associated to 2781 events). Fourth, they feature high diversity in writing styles because they contain articles from across the political spectrum, including left- and right-wing outlets. Fifth, we find that they contain only few minor content errors albeit being created through scraping or crawling.

3.2 Creation of examples

To create a batch of examples for annotation, we devise a three tasks process: first, we extract example candidates from randomly selected articles. Second, we discard non-optimal candidates. Only for the train set, third, we filter candidates to address class imbalance. We repeatedly execute these tasks so that each batch yields 500 examples for annotation, contributed equally by both sources.

First, we randomly select articles from the two sources. Since both are at least very approximately uniformly distributed over time (Gebhard and Hamborg, 2020; Chen et al., 2018), randomly drawing articles will yield sufficiently high diversity in both writings styles and reported topics (Section 3.1). To extract from an article examples that contain meaningful target mentions, we employ coreference resolution (CR).[3] We iterate all resulting coreference clusters of the given article and create a single example for each mention and its enclosing sentence.

Extraction of mentions of named entities (NEs) is the commonly employed method to create examples in previous TSC datasets (Rosenthal et al., 2017; Nakov et al., 2016, 2013; Steinberger et al., 2017). We do not use it since we find it would miss $\gtrapprox 30\%$ mentions of relevant target candidates, e.g., pronominal or near-identity mentions.

Second, we perform a two level filtering to improve quality and "substance" of candidates. On coreference cluster level, we discard a cluster c in a document d if $|M_c| \leq 0.2|S_d|$, where $|...|$ is the number of mentions of a cluster (M_c) and sentences in a document (S_d). Also, we discard non-persons clusters, i.e., if $\exists m \in M_c : t(m) \notin \{-, P\}$, where $t(m)$ yields the NE type[4] of m, and $-$ and P represent the unknown and person type, respectively. On example level, we discard short and similar examples e, i.e., if $|s_e| < 50$ or if $\exists \hat{e} : sim(s_e, s_{\hat{e}}) > 0.6 \wedge m_e = m_{\hat{e}} \wedge t_e = t_{\hat{e}}$ where s_e, m_e, and t_e are the sentence of e, its mention, and the target's cluster, respectively, and $sim(...)$ the cosine similarity. Lastly, if a cluster has multiple mentions in a sentence, we try to select the most meaningful example. In short, we prefer the cluster's representative mention[5] over nominal mentions, and those over all other instances.

Third, for only the train set, we filter candidates to address class imbalance. Specifically, we discard examples e that are likely the majority class ($p(\text{neutral}|s_e) > 0.95$) as determined by a simple binary classifier (Sanh et al., 2019). Whenever annotated and consolidated examples are added to the

[2] POLUSA by design (Gebhard and Hamborg, 2020) and BF18 was crawled from a news aggregator on a daily basis.

[3] We employ spaCy 2.1 and neuralcoref 4.0.

[4] Determined by spaCy.

[5] Determined by neuralcoref.

train set of NewsMTSC, we retrain the classifier on them and all previous examples in the train set.

3.3 Annotation

Instructions used in popular TSC datasets plainly ask annotators to rate the sentiment of a text toward a target (Rosenthal et al., 2017; Pontiki et al., 2015). For news texts, we find that doing so yields two issues (Balahur et al., 2010): low inter-annotator reliability (IAR) and low suitability. Low suitability refers to examples where annotators' answers can be consolidated but the resulting majority answer is incorrect as to the task. For example, instructions from prior TSC datasets often yield low suitability for polarizing targets, independently of the sentence they are mentioned in. Figure 2 (Appendix) depicts our final annotation instructions.

In an interactive process with multiple test annotations (six on-site and eight on Amazon Mechanical Turk, MTurk), we test various measures to address the two issues. We find that asking annotators to think from the perspective of the sentence's author strongly facilitates that annotators overcome their personal attitude. Further, we find that we can effectively draw annotators' attention not only at the event and other "facts" described in sentence (the "what") but also at word choice ("how" it is described) by exemplarily mentioning both factors and abstracting these factors as the author's holistic "attitude." [6] We further improve IAR and suitability, e.g., by explicitly instructing annotators to rate sentiment only regarding the target but not other aspects, such as the reported event.

While most TSC dataset creation procedures use 3- or 5-point Likert scales (Nakov et al., 2013, 2016; Rosenthal et al., 2017; Pontiki et al., 2014, 2015; Balahur et al., 2010; Steinberger et al., 2017), we use a 7-point scale to encourage rating also only slightly positive or negative examples as such.

Technically, we closely follow previous literature on TSC datasets (Pontiki et al., 2015; Rosenthal et al., 2017). We conduct the annotation of our examples on MTurk. Each example is shown to five randomly selected crowdworkers. To participate in our annotation, crowdworkers must have the "Master" qualification, i.e., have a record of successfully completed, high quality work on MTurk. To ensure quality, we implement a set of objective measures and tests (Kim et al., 2012). While we pay all crowdworkers always (USD 0.07 per assignment), we discard all of a crowdworker's answers if at least one of the following conditions is met. A crowdworker (a) was not shown any test question or answered at least one incorrectly[7], (b) provided answers to invisible fields in the HTML form (0.3% of crowdworkers did so, supposedly bots), or (c) the average duration of time spent on the assignments was extremely low ($< 4s$).

The IAR is sufficiently high ($\kappa_C = 0.74$) when considering only examples in NewsMTSC. The expected mixed quality of crowdsourced work becomes apparent when considering all examples, including those that could not be consolidated and answers of those crowdworkers who did not pass our quality checks ($\kappa_C = 0.50$).

3.4 Consolidation

We consolidate the answers of each example to a majority answer by employing a restrictive strategy. Specifically, we consolidate the set of five answers A to the single-label 3-class polarity $p \in \{\text{pos.}, \text{neu.}, \text{neg.}\}$ if $\exists C \subseteq A : |C| \geq 4 \wedge \forall c \in C : s(c) = p$, where $s(c)$ yields the 3-class polarity of an individual 7-class answer c, i.e., neutral $\Rightarrow$ neutral, any positive (from slightly to strongly) $\Rightarrow$ positive, and respectively for negative. If there is no such consolidation set C, A cannot be consolidated and the example is discarded. Consolidating to 3-class polarity allows for direct comparison to established TSC dataset.

While the strategy is restrictive (only 50.6% of all examples are consolidated this way), we find it yields the highest quality. We quantify the dataset's quality by comparing the dataset to an expert annotation (Section 3.6) and by training and testing models on dataset variants with different consolidations. Compared to consolidations employed for previous TSC datasets, quality is improved significantly on our examples, e.g., our strategy yields $F1_m = 86.4$ when comparing to experts' annotations and models trained on the resulting set yield up to $F1_m = 83.1$ whereas the two-step majority strategy employed for the Twitter 2016 set (Nakov et al., 2016) yields 50.6 and 53.4 respectively.

3.5 Splits and multi-target examples

NewsMTSC consists of three sets as depicted in Table 1. For the *train* set, we employ class balanc-

[6] To think from the author's perspective is not to be confused with the "author-level" defined by Balahur et al. (2010).

[7] Prior to submitting a batch of examples to MTurk, we add 6% test examples with unambiguous sentiment, e.g., "Mr. Smith is a bad guy."

Set	Total	Pos.	Neu.	Neg.	MT-a	MT-d	+Corefs	Pos.	Neu.	Neg.
Train	8739	2395	3028	3316	972	341	11880	3434	3744	4702
Test-mt	1476	246	748	482	721	294	1883	333	910	640
Test-rw	1146	361	587	624	73	30	1572	361	587	624

Table 1: Statistics of NewsMTSC. Columns (f.l.t.r.): name; count of targets with any, positive, neutral, and negative sentiment, respectively; count of examples with multiple targets of any and different polarity, respectively; count of targets and their coreferential mentions with any, pos., neu. and neg. sentiment, respectively.

ing prior to annotation (Section 3.2). To minimize dataset shift, which might yield a skewed sentiment distribution in the dataset compared to the real-world (Quionero-Candela et al., 2009), we do not use class balancing for either of the two test sets. Sentences can have multiple targets (MT) with potentially different polarities. We call this *MT property*. To investigate the effect on TSC performance of considering or neglecting the MT property (Jiang et al., 2019), we devise a test set named *test-mt*, which consists only of examples that have at least two semantically different targets, i.e., each belonging to a separate coreference cluster (Section 3.2). Since the additional filtering required for *test-mt* naturally yields dataset shift, we create a second test set named *test-rw*, which omits the MT filtering and is thus designed to be as close as possible to the real-world distribution of sentiment. We seek to provide a sentiment score for each person in each sentence in *train* and *test-rw* but mentions may be missing, e.g., because of erroneous coreference resolution or crowdworkers' answers could not be consolidated.

3.6 Quality and characteristics

We conduct an expert annotation of a random subset of 360 examples used during the creation of NewsMTSC with five international graduate students (studying Political or Communication Science at the University of Zurich, Switzerland, 3 female, 2 male, aged between 23 and 29). Key differences compared to the MTurk annotation are: first, extensive training until high IAR is reached (considering all examples: $\kappa_C = 0.72$, only consolidated: $\kappa_C = 0.93$). We conduct five iterations, each consisting of individual annotations by the students, quantitative and qualitative review, adaption of instructions, and individual and group discussions. Second, comprehensive instructions (4 pages). Third, no time pressure, since the students are paid per hour (crowdworkers per assignment).

When comparing the expert annotation with our dataset, we find that NewsMTSC is of high quality ($F1_m = 86.4$). The quality of unfiltered answers from MTurk is, as expected, much lower (50.1).

What is contained in NewsMTSC? In a random set of 50 consolidated examples from MTurk, we find that most frequent, non-mutually exclusive means to express a polar statement (62% of the 50) are usage of quotes (in total, direct, and indirect 42%, 28%, and 14%, respectively), target being subject to action (24%), evaluative expression by the author or an opinion holder mentioned outside of the sentence (18%), target performing an action (16%), and loaded language or connotated terms (14%). Direct quotes often contain evaluative expressions or connotated terms, indirect quotes less. Neutral examples (38% of the 50) contain mostly objective storytelling about neutral events (16%) or variants of "[target] said that ..." (8%).

What is not contained in NewsMTSC? We qualitatively review all examples where individual answers could not be consolidated to identify potential causes why annotators do not agree. The predominant reason is technical, i.e., the restrictiveness of the consolidation (MTurk compared to experts: $26\% \approx 30\%$). Other examples lack apparent causes ($24\% \gg 8\%$). Further potential causes are (not mutually exclusive): ambiguous sentence ($16\% \approx 18\%$), sentence contains positive and negative parts ($8\% \approx 6\%$), opinion holder is target ($6\% \approx 8\%$), e.g., "[...] Bauman asked supporters to 'push back' against what he called a targeted campaign to spread false rumors about him online."

What are qualitative differences in the annotations by crowdworkers and experts? We review all 63 cases (18%) where answers from MTurk could be consolidated but differ to experts' answers. The major reason for disagreement is the restrictiveness of the consolidation (53 cases have no consolidation among the experts). In 10 cases the consolidated answers differ. We find that in few examples (2-3%) crowdsourced annotations are superficial and fail to interpret the full sentence correctly.

Texts in NewsMTSC are much longer than in

prior TSC datasets (mean over all examples): 152 characters compared to 100, 96, and 90 in Twitter, Restaurant, and Laptops, respectively.

4 Methodology

The goal of TSC is to find a target's polarity $y \in \{\text{pos.}, \text{neu.}, \text{neg.}\}$ in a sentence. Our model consists of four key components (Figure 1): a pre-trained language model (LM), a representation of external knowledge sources (EKS), a target mention mask, and a bidirectional GRU (BiGRU) (Cho et al., 2014). We adapt our model from Hosseinia et al. (2020) and change the design as follows: we employ a target mask (which they did not) and use multiple EKS simultaneously (instead of one). Further, we use a different set of EKS (Section 5) and do not exclude the LM's parameters from fine-tuning.

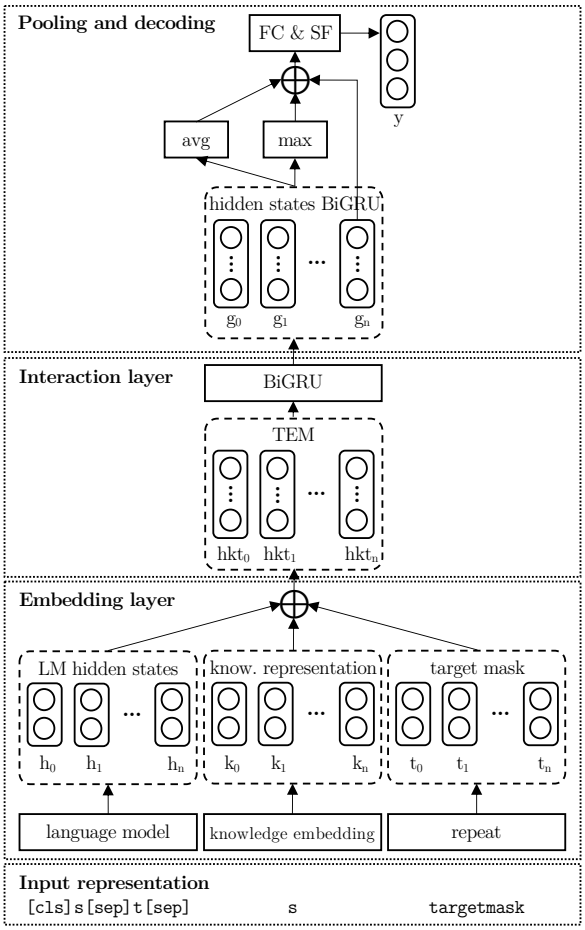

Figure 1: Overall architecture of the proposed model.

4.1 Input representation

We construct three model inputs. The first is a text input T constructed as suggested by Devlin et al. (2019) for question-answering (QA) tasks. Specifically, we concatenate the sentence and target mention and tokenize the two segments using the LM's tokenizer and vocabulary, e.g., WordPiece for BERT (Wu et al., 2016).[8] This step results in a text input sequence $T = [\text{CLS}, s_0, s_1, ..., s_p, \text{SEP}, t_0, t_1, ..., t_q, \text{SEP}] \in \mathbb{N}^n$ consisting of n word pieces, where n is the manually defined maximum sequence length.

The second input is a feature representation of the sentence, which we create using one or more EKS, such as dictionaries (Hosseinia et al., 2020; Zhang et al., 2020). Given an EKS with d dimensions, we construct an EKS representation $E \in \mathbb{R}^{n \times d}$ of S, where each vector $e_{i \in \{0,1,...,p\}}$ is a feature representation of the word piece i in the sentence. To facilitate learning associations between the token-based EKS representation and the WordPiece-based sequence T, we create E so that it contains k repeated vectors for each token where k is the token's number of word pieces. Thereby, we also consider special characters, such as CLS. If multiple EKS with a total number of dimensions $\hat{d} = \sum d$ are used, their representations of the sentence are stacked resulting in $E \in \mathbb{R}^{n \times \hat{d}}$.

The third input is a target mask $M \in \mathbb{R}^n$, i.e., for each word piece i in the sentence that belongs to the target, $m_i = 1$, else 0 (Gao et al., 2019).

4.2 Embedding layer

We feed T into the LM to yield a contextualized word embedding of shape $\mathbb{R}^{n \times h}$, where h is the number of hidden states in the language model. We feed E into a randomly initialized matrix $W_E \in \mathbb{R}^{\hat{d} \times h}$ to yield an EKS embedding. We repeat M to be of shape $\mathbb{R}^{n \times h}$. By creating all embeddings in the same shape, we facilitate a balanced influence of each input to the model's downstream components. We stack all embeddings to form a matrix $TEM \in \mathbb{R}^{n \times 3h}$.

4.3 Interaction layer

We allow the three embeddings to interact using a single-layer BiGRU (Hosseinia et al., 2020), which yields hidden states $H \in \mathbb{R}^{n \times 6h} = \text{BiGRU}(TEM)$. RNNs, such as LSTMs and GRUs, are commonly used to learn a higher-level representation of a word embedding, especially in state-of-the-art TSC prior to BERT-based models but also recently (Liu et al., 2015; Li et al., 2019; Hosseinia et al., 2020; Zhang et al., 2020). We

[8] For readability, we showcase inputs as used for BERT.

choose an BiGRU over an LSTM because of the smaller number of parameters in BiGRUs, which may in some cases result in better performance (Chung et al., 2014; Jiang et al., 2019; Hosseinia et al., 2020; Gruber and Jockisch, 2020).

4.4 Pooling and decoding

We employ three common pooling techniques to turn the interacted, sequenced representation H into a single vector (Hosseinia et al., 2020). We calculate element-wise (1) mean and (2) maximum over all hidden states H and retrieve the (3) last hidden state h_{n-1}. Then, we stack the three vectors to P, feed P into a fully connected layer FC so that $z = FC(P)$ and calculate $y = \sigma(z)$.

5 Experiments

5.1 Experimental data

In addition to NewsMTSC, we use the three established TSC sets: Twitter, Laptop, and Restaurant.

5.2 Evaluation metrics

We use metrics established in the TSC literature: macro F1 on all ($F1_m$) and only the positive and negative classes ($F1_{pn}$), accuracy (a), and average recall (r_a). If not otherwise noted performances are reported for our primary metric, $F1_m$.

5.3 Baselines

We compare our model with TSC methods that yield state-of-the-art results on at least one of the established datasets: SPC-BERT (Devlin et al., 2019): input is identical to our text input. FC and softmax is calculated on CLS token. TD-BERT (Gao et al., 2019): masks hidden states depending on whether they belong to the target mention. LCF-BERT (Zeng et al., 2019): similar to TD but additionally weights hidden states depending on their token-based distance to the target mention. We use the improved implementation (Yang, 2020) and enable the dual-LM option, which yields slightly better performance than using only one LM instance (Zeng et al., 2019). We also planned to test LCFS-BERT (Phan and Ogunbona, 2020) but due to technical issues we were not able to reproduce the authors' results and thus exclude LCFS from our experiments.

5.4 Implementation details

To find for each model the best parameter configuration, we perform an exhaustive grid search. Any number we report is the mean of five experiments that we run per configuration. We randomly split each test set into a dev-set (30%) and the actual test-set (70%). We test the base version of three LMs: BERT, RoBERTa, and XLNET. For all methods, we test parameters suggested by their respective authors.[9] We test all 15 combinations of the following 4 EKS: (1) *SENT* (Hu and Liu, 2004): a sentiment dictionary (number of non-mutually exclusive dimensions: 2, domain: customer reviews). (2) *LIWC* (Tausczik and Pennebaker, 2010): a psychometric dictionary (73, multiple). (3) *MPQA* (Wilson et al., 2005): a subjectivity dictionary (3, multiple). (4) *NRC* (Mohammad and Turney, 2010): dictionary of sentiment and emotion (10, multiple).

5.5 Overall performance

Table 2 reports the performances of the models using different LMs and evaluated on both test sets. We find that the best performance is achieved by our model ($F1_m = 83.1$ on *test-rw* compared to 81.8 by prior state-of-the-art). For all models, performances are (strongly) improved when using RoBERTa, which is pre-trained on news texts, or XLNET, likely because of its large pre-training corpus. Because of limited space, XLNET is not reported in Table 2, but results are generally similar to RoBERTa except for the TD model, where XLNET degrades performance by 5-9pp. Looking at BERT, we find no significant improvement of GRU-TSC over prior state-of-the-art. Even if we domain-adapt BERT (Rietzler et al., 2019) for 3 epochs on a random sample of 10M English sentences (Gebhard and Hamborg, 2020), BERT's performance ($F1_m = 81.8$) is lower than RoBERTa. We notice a performance drop for all models when comparing *test-rw* and *test-mt*. It seems that RoBERTa is better able to resolve in-sentence relations between multiple targets (performance degeneration of only up to -0.6pp) than BERT (-2.9pp). We suggest to use RoBERTa for TSC on news, since fine-tuning it is faster than fine-tuning XLNET, and RoBERTa achieves similar or better performance than other LMs.

While GRU-TSC yields competitive results on

[9] Epochs $\in \{2, 3, 4\}$; batch size $\in \{8, 16\}$ (due to constrained resources not 32); learning rate $\in \{2e{-}5, 3e{-}5, 5e{-}5\}$; label smoothing regularization (LSR) (Szegedy et al., 2016): $\epsilon \in \{0, .2\}$; dropout rate: .1; $\mathcal{L}_2$ regularization: $\lambda = 1e{-}5$; SRD for LCF $\in \{3, 4, 5\}$. We use Adam optimization (Kingma and Ba, 2014), Xavier uniform initialization (Glorot and Bengio, 2010), and cross-entropy loss.

Model		Test-rw				Test-mt			
		$F1_m$	a	$F1_{pn}$	r_a	$F1_m$	a	$F1_{pn}$	r_a
BERT	SPC	**80.1**	80.7	79.5	79.8	73.7	76.1	71.1	76.0
	TD	79.4	79.9	78.9	80.0	75.6	79.1	72.0	75.8
	LCF	79.7	80.9	78.9	79.2	**77.7**	80.5	74.6	79.1
	GRU	**80.2**	81.1	79.7	80.0	77.3	80.0	74.1	77.9
RoBERTa	SPC	81.1	82.7	80.5	80.6	79.4	81.6	77.0	79.9
	TD	81.7	82.5	81.3	81.4	78.4	81.1	75.3	78.2
	LCF	81.4	82.5	80.8	81.1	81.2	83.8	78.6	81.7
	GRU	**83.1**	83.8	82.9	83.3	**82.5**	84.6	80.2	81.0

Table 2: Experimental results on the two test sets.

	Laptop		Restaurant		Twitter	
	$F1_m$	a	$F1_m$	a	$F1_m$	a
SPC	77.4	80.3	78.8	86.0	73.6	75.3
TD	74.4	78.9	78.4	85.1	74.3	77.7
LCF	**79.6**	82.4	**81.7**	87.1	**75.8**	77.3
GRU	79.0	82.1	80.7	86.0	74.6	76.0

Table 3: Results on previous TSC datasets.

Name	$F1_m$	a
no EKS	78.2	81.0
zeros	78.4	81.1
SENT	**80.7**	83.0
LIWC	**80.8**	83.1
MPQA	78.8	80.8
NRC	80.0	82.0
best combination	**81.0**	83.3

Table 4: Results of exemplary EKS combinations.

previous TSC datasets (Table 3), LCF is the top performing model.[10] When comparing the performances across all four datasets, the importance of the consolidation becomes apparent, e.g., performance is lowest on Twitter, which employs a simplistic consolidation (Section 3.4). The performance differences of individual models when contrasting their use on prior datasets and NewsMTSC highlight the need LCF performs consistently best on prior datasets but worse than GRU-TSC on NewsMTSC. One reason might be that LCF's weighting approach relies on a static distance parameter, which seems to degrade performance when used on longer texts as in NewsMTSC (Section 3.6). When increasing LCF's window width SRD, we notice a slight improvement of 1pp (SRD=5) but degradation for larger SRD.

5.6 Ablation study

We perform an ablation study to test the impact of four key factors: target mask, EKS, coreferential mentions, and fine-tuning the LM's parameters. We test all LMs and if not noted otherwise report results for RoBERTa since it generally performs best (Section 5.5). We report results for *test-mt* (performance influence is similar on either test set, with performances generally being $\approx$ 3-5pp higher on *test-rw*). Overall, we find that our changes to the initial design (Hosseinia et al., 2020) contribute to an improvement of approximately 1.9pp. The most influential changes are the selected EKS and in part use of coreferential mentions. Using the target mask input channel without coreferences and LM fine-tuning yield insignificant improvements of up to 0.3 each. We do not test the VADER-based sentence classification proposed by Hosseinia et al. (2020) since we expect no improvement by using it for various reasons. For example, VADER uses a dictionary created for a domain other than news and classifies the sentence's overall sentiment and thus is target-independent.

Table 4 details the results of exemplary EKS, showing that the best combination (SENT, MPQA, and NRC) yields an improvement of 2.6pp compared to not using an EKS (zeros). The single best EKS (LIWC or SENT) each yield an improvement of 2.4pp. The two EKS "no EKS" and "zeros" represent a model lacking the EKS input channel and an EKS that only yields 0's, respectively.

The use of coreferences has a mixed influence on performance (Table 5). While using coreferences has no or even negative effect in our model for large LMs (RoBERTa and XLNET), it can be beneficial

[10] For previous models, Table 3 lists results reported by their authors. In our experiments, we find 0.4-1.8pp lower performance compared to the reported results.

Name	BERT	RoBERTa
none	73.1	**78.1**
target mask	73.3	**78.2**
add coref. to mask	**75.6**	**78.1**
add coref. as example	73.0	73.4

Table 5: Influence of target mask and coreferences.

for smaller LMs (BERT) or batch sizes (8). When using the mode "ignore," "add coref. to mask," and "add coref. as example" we ignore coreferences, add them to the target mask, and create an additional example for each, respectively. Mode "none" represents a model that lacks the target mask input channel.

6 Error Analysis

To understand the limitations of GRU-TSC, we carry out a manual error analysis by investigating a random sample of 50 incorrectly predicted examples for each of the test sets. For *test-rw*, we find the following potential causes (not mutually exclusive): edge cases with very weak, indirect, or in part subjective sentiment (22%) or where both the predicted and true sentiment can actually be considered correct (10%); sentiment of given target confused with different target (14%). Further, sentence's sentiment is unclear due to missing context (10%) and consolidated answer in NewsMTSC is wrong (10%). In 16% we find no apparent reason. For *test-mt*, potential causes occur approximately similarly often except that targets are confused more often (20%).

7 Future Work

We identify three main areas for future work. The first area is related to the dataset. Instead of consolidating multiple annotators' answers during the dataset creation, we propose to test to integrate the label selection into the model (Raykar et al., 2010). Integrating the label selection into the machine learning part could improve the classification performance. It could also allow us to include more sentences in the dataset, especially the edge cases that our restrictive consolidation currently discards.

To improve the model design, we propose to design the model specifically for sentences with multiple targets, for example, by classifying multiple targets in a sentence simultaneously. While we early tested various such designs, we did not report them in the paper due to their comparably poor performances. Further work in this direction should perhaps also focus on devising specialized loss functions that set multiple targets and their polarity into relation. Lastly, one can improve various technical details of GRU-TSC, e.g., by testing other interaction layers, such as LSTMs, or using layer-specific learning rates in the overall model, which can increase performance (Sun et al., 2019b).

8 Conclusion

We present NewsMTSC, a dataset for target-dependent sentiment classification (TSC) on news articles consisting of 11.3k manually annotated examples. Compared to prior TSC datasets, it is different in key factors, such as its texts are on average 50% longer, sentiment is expressed explicitly only rarely, and there is a separate test set for multi-target sentences. As a consequence, state-of-the-art TSC models yield non-optimal performances. We propose GRU-TSC, which uses a bidirectional GRU on top of a language model (LM) and other embeddings, instead of masking or weighting mechanisms as employed by prior state-of-the-art. We find that GRU-TSC achieves superior performances on NewsMTSC and is competitive on prior TSC datasets. RoBERTa yields best results compared to using BERT, because RoBERTa is pre-trained on news and we find it can better resolve in-sentence relations of multiple targets.

Acknowledgments

The work described in this paper is partially funded by the WIN program of the Heidelberg Academy of Sciences and Humanities, financed by the Ministry of Science, Research and the Arts of the State of Baden-Wurttemberg, Germany. We are grateful to the crowdworkers who participated in our online annotation and the UZH students who participated in the expert annotation: F. Jedelhauser, Y. Kipfer, J. Roberts, L. Sopa, and F. Wallin. We thank the anonymous reviewers for their valuable comments that helped to improve this paper.

References

Abeer AlDayel and Walid Magdy. 2020. Stance Detection on Social Media: State of the Art and Trends. *Preprint*.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion

Mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, volume 10, pages 2200–2204, Valletta, Malta. European Language Resources Association (ELRA).

Alexandra Balahur, Ralf Steinberger, Mijail Kabadjov, Vanni Zavarella, Erik Van Der Goot, Matina Halkia, Bruno Pouliquen, and Jenya Belyaeva. 2010. Sentiment analysis in the news. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Dan Bernhardt, Stefan Krasa, and Mattias Polborn. 2008. Political polarization and the electoral effects of media bias. *Journal of Public Economics*, 92(5):1092–1104.

Douglas Biber and Edward Finegan. 1989. Styles of stance in English: Lexical and grammatical marking of evidentiality and affect. *Text - Interdisciplinary Journal for the Study of Discourse*, 9(1).

Wei-Fan Chen, Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. 2018. Learning to Flip the Bias of News Headlines. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 79–88, Stroudsburg, PA, USA. Association for Computational Linguistics.

Xiao Chen, Changlong Sun, Jingjing Wang, Shoushan Li, Luo Si, Min Zhang, and Guodong Zhou. 2020. Aspect Sentiment Classification with Document-level Sentiment Preference Modeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3667–3677, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Stroudsburg, PA, USA. Association for Computational Linguistics.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling.

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*, pages 4171–4186, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chunning Du, Haifeng Sun, Jingyu Wang, Qi Qi, and Jianxin Liao. 2020. Adversarial and Domain-Aware BERT for Cross-Domain Sentiment Analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4019–4028, Stroudsburg, PA, USA. Association for Computational Linguistics.

Robert M. Entman. 2007. Framing Bias: Media in the Distribution of Power. *Journal of Communication*, 57(1):163–173.

Charles J. Fillmore and Collin F. Baker. 2001. Frame semantics for text understanding. In *Proceedings of NAACL WordNet and Other Lexical Resources Workshop*, pages 1–6, Pittsburgh, US.

Zhengjie Gao, Ao Feng, Xinyu Song, and Xi Wu. 2019. Target-Dependent Sentiment Classification With BERT. *IEEE Access*, 7:154290–154299.

Lukas Gebhard and Felix Hamborg. 2020. The POLUSA Dataset: 0.9M Political News Articles Balanced by Time and Outlet Popularity. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, pages 467–468, New York, NY, USA. ACM.

Deepanway Ghosal, Devamanyu Hazarika, Abhinaba Roy, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2020. KinGDOM: Knowledge-Guided DOMain Adaptation for Sentiment Analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3198–3210, Stroudsburg, PA, USA. Association for Computational Linguistics.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Journal of Machine Learning Research*, pages 249–256.

Namrata Godbole, Manja Srinivasaiah, and Steven Skiena. 2007. Large-Scale Sentiment Analysis for News and Blogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, volume 7, pages 219–222, Boulder, CO, USA.

Nicole Gruber and Alfred Jockisch. 2020. Are GRU Cells More Specific and LSTM Cells More Sensitive in Motive Classification of Text? *Frontiers in Artificial Intelligence*, 3.

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):391–415.

Felix Hamborg, Karsten Donnay, and Bela Gipp. 2021. Towards Target-dependent Sentiment Classification in News Articles. In *Proceedings of the 16th iConference*, pages 1–9, Beijing, China (Virtual Event). Springer.

Marjan Hosseinia, Eduard Dragut, and Arjun Mukherjee. 2020. Stance Prediction for Contemporary Issues: Data and Experiments. In *Proceedings of*

the Eighth International Workshop on Natural Language Processing for Social Media, pages 32–40, Stroudsburg, PA, USA. Association for Computational Linguistics.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04*, page 168, New York, New York, USA. ACM Press.

Qingnan Jiang, Lei Chen, Ruifeng Xu, Xiang Ao, and Min Yang. 2019. A Challenge Dataset and Effective Models for Aspect-Based Sentiment Analysis. 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 6279–6284, Stroudsburg, PA, USA. Association for Computational Linguistics.

Daniel Kahneman and Amos Tversky. 1984. Choices, values, and frames. *American Psychologist*, 39(4):341–350.

Sung-Hee Kim, Hyokun Yun, and Ji Soo Yi. 2012. How to filter out random clickers in a crowdsourcing-based study? In *Proceedings of the 2012 BELIV Workshop on Beyond Time and Errors - Novel Evaluation Methods for Visualization - BELIV '12*, pages 1–7, New York, New York, USA. ACM Press.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv: 1412.6980*.

Svetlana Kiritchenko, Xiaodan Zhu, Colin Cherry, and Saif Mohammad. 2014. NRC-Canada-2014: Detecting Aspects and Sentiment in Customer Reviews. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 437–442, Dublin, Ireland. Association for Computational Linguistics.

Xin Li, Lidong Bing, Piji Li, and Wai Lam. 2019. A Unified Model for Opinion Target Extraction and Target Sentiment Prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:6714–6721.

Bing Liu. 2012. Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

Pengfei Liu, Shafiq Joty, and Helen Meng. 2015. Fine-grained Opinion Mining with Recurrent Neural Networks and Word Embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1433–1443, Stroudsburg, PA, USA. Association for Computational Linguistics.

Saif Mohammad and Peter Turney. 2010. Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34, Los Angeles, CA. Association for Computational Linguistics.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016. SemEval-2016 Task 4: Sentiment Analysis in Twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1–18, San Diego, CA, USA. Association for Computational Linguistics.

Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. 2013. SemEval-2013 Task 2: Sentiment Analysis in Twitter. In *Second Joint Conference on Lexical and Computational Semantics (SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 312–320, Atlanta, GA, USA. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2005. Seeing stars. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics - ACL '05*, pages 115–124, Morristown, NJ, USA. Association for Computational Linguistics.

Minh Hieu Phan and Philip O. Ogunbona. 2020. Modelling Context and Syntactical Features for Aspect-based Sentiment Analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3211–3220, Stroudsburg, PA, USA. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 Task 12: Aspect Based Sentiment Analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, Stroudsburg, PA, USA. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 Task 4: Aspect Based Sentiment Analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Stroudsburg, PA, USA. Association for Computational Linguistics.

James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*, 1 edition. O'Reilly Media, Inc., Sebastopol, CA, US.

Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. 2009. *Dataset Shift in Machine Learning*. The MIT Press.

Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca

Bogoni, and Linda Moy. 2010. Learning From Crowds. *The Journal of Machine Learning Research*, 11:1297–1322.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic Models for Analyzing and Detecting Biased Language. In *Proceedings of the 51st Annual Meeting on Association for Computational Linguistics*, pages 1650–1659, Sofia, BG. Association for Computational Linguistics.

Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. 2019. Adapt or Get Left Behind: Domain Adaptation through BERT Language Model Finetuning for Aspect-Target Sentiment Classification. *arXiv preprint arXiv:1908.11860*.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 Task 4: Sentiment Analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv: 1910.01108*.

Youwei Song, Jiahai Wang, Tao Jiang, Zhiyue Liu, and Yanghui Rao. 2019. Targeted Sentiment Classification with Attentional Encoder Network. In *Artificial Neural Networks and Machine Learning - ICANN 2019: Text and Time Series*, pages 93–103, Cham, US. Springer International Publishing.

Ralf Steinberger, Stefanie Hegele, Hristo Tanev, and Leonida Della Rocca. 2017. Large-scale news entity sentiment analysis. In *RANLP 2017 - Recent Advances in Natural Language Processing Meet Deep Learning*, pages 707–715. Incoma Ltd. Shoumen, Bulgaria.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019a. Utilizing BERT for Aspect-Based Sentiment Analysisvia Constructing Auxiliary Sentence. In *Proceedings of the 2019 Conference of the North*, pages 380–385, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019b. How to Fine-Tune BERT for Text Classification? In *Chinese Computational Linguistics*, pages 194–206. Springer International Publishing, Cham, US.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826. IEEE.

Yla R. Tausczik and James W. Pennebaker. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT '05*, pages 347–354, Morristown, NJ, USA. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *Preprint*.

Heng Yang. 2020. LC-ABSA.

Da Yin, Tao Meng, and Kai-Wei Chang. 2020. SentiBERT: A Transferable Transformer-Based Architecture for Compositional Sentiment Semantics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3695–3706, Stroudsburg, PA, USA. Association for Computational Linguistics.

Biqing Zeng, Heng Yang, Ruyang Xu, Wu Zhou, and Xuli Han. 2019. LCF: A Local Context Focus Mechanism for Aspect-Based Sentiment Classification. *Applied Sciences*, 9(16):1–22.

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, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4).

A Appendices

Imagine you are a journalist asked to write a news article about a given topic. Depending on your own attitude towards the topic or the people involved in the news story, you may portray people more positively and other more negatively. For example, by using rather positive of negative words, e.g., 'freedom fighters' vs. 'terrorists' or 'cross the border' vs. 'invade,' or by describing positive or negative aspects, e.g., that a person did something negative.

In the sentence below, what do you think is the **attitude of the sentence's author towards the underlined subject**? Consider the attitude only towards the underlined subject, not the event itself or other people. FYI: further assignments may show the same sentence but with a different underlined subject than the subject shown below.

Subject: the president

The comments come after McConnell expressed his frustrations with the president for having "excessive expectations" for his agenda.

The attitude of the sentence's author towards the underlined subject is…

strongly negative attitude (dislike) | neutral / can't tell | strongly positive attitude (like)

Figure 2: Final version of the annotation instructions shown on MTurk.