Who Blames or Endorses Whom? Entity-to-Entity Directed Sentiment Extraction in News Text

Kunwoo Park
Soongsil University*
kunwoo.park@ssu.ac.kr

Zhufeng Pan
UCLA
panzhufeng@cs.ucla.edu

Jungseock Joo
UCLA
jjoo@comm.ucla.edu

Abstract

Understanding who blames or supports whom in news text is a critical research question in computational social science. Traditional methods and datasets for sentiment analysis are, however, not suitable for the domain of political text as they do not consider the direction of sentiments expressed between entities. In this paper, we propose a novel NLP task of identifying directed sentiment relationship between political entities from a given news document, which we call *directed sentiment extraction*. From a million-scale news corpus, we construct a dataset of news sentences where sentiment relations of political entities are manually annotated. We present a simple but effective approach for utilizing a pretrained transformer, which infers the target class by predicting multiple question-answering tasks and combining the outcomes. We demonstrate the utility of our proposed method for social science research questions by analyzing positive and negative opinions between political entities in two major events: 2016 U.S. presidential election and COVID-19. The newly proposed problem, data, and method will facilitate future studies on interdisciplinary NLP methods and applications.[1]

1 Introduction

Sentiment analysis is a useful technique for opinion mining from text data. Most existing work either focuses on sentence-level classification (Van Hee et al., 2018; Zampieri et al., 2019), or aims to detect the sentiment polarity towards specific targets (Pontiki et al., 2016; Cortis et al., 2017). These approaches typically do not distinguish sources and targets of the sentiment. They mainly use user-generated content (UGC), such as tweets or restaurant reviews from Yelp, which assumes each user (account holder) is the source of the sentiment, and that the target is also clearly defined or easily identifiable (e.g., the restaurant).

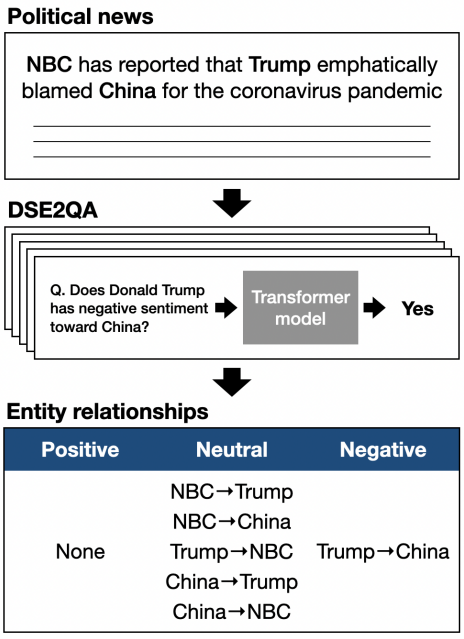

Figure 1: Overview of the directed sentiment extraction

This assumption does not hold for political news analysis where a large number of political actors blame or support each other in news articles. The key interest for political sentiment analysis is to identify "who" blame or support "whom" (Balahur and Steinberger, 2009) rather than simply assigning a global sentiment polarity to a given document or sentence. For example, from a sentence like "X supported Y for criticizing Z," we can infer X is positive toward Y and both X and Y are negative toward Z. However, existing sentiment analysis methods are not suitable to detect such sentiment relationships between entities.

This paper proposes a new NLP task of identifying directed sentiment relationships from a political entity to another in news articles: *directed sentiment extraction*. For this task, we introduce a newly annotated corpus from a million-scale news

This work was done while the first author was a postdoctoral researcher at UCLA.

[1] https://github.com/bywords/directed_sentiment_analysis

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 4091–4102

dataset. As demonstrated in Figure 1, we transform directed sentiment extraction to multiple question-answering tasks (**DSE2QA**) and combine their predictions for making a final prediction. Evaluation results show it outperforms state-of-the-art classification approaches, such as a fine-tuned RoBERT classification model. Going further, we demonstrate the utility of our method through two case studies on how news media in the U.S. portrayed relationships between political entities differently amid the US election and COVID-19 pandemic. The analysis reveals that the left-leaning media present Donald Trump more as the target of blame while the right-leaning media present news stories blaming other entities. This study not only makes a contribution to the NLP community by defining a new problem and approach but also adds the empirical understanding of media bias to the social science community.

2 Related Work

2.1 Sentiment Analysis in Media

Sentiment analysis or polarity detection aims at deciding whether a given text contains positive or negative sentiment (Liu, 2012) or quantifying the degree of sentiments embedded in a text (Gilbert and Hutto, 2014). Previous studies have tackled the problem as a classification task (Maas et al., 2011) and applied the trained model to infer sentiments embedded in various web and social data (Park et al., 2018). Recently, transformer-based models have shown high performance in sentiment classification (Devlin et al., 2019) and aspect-based sentiment analysis (Sun et al., 2019).

Measuring sentiment or tone of political text in news media is a widely used method in computational social science (Young and Soroka, 2012). A stream of work has used social media posts to estimate public opinions about political actors and predict the outcomes of future events by large scale sentiment analysis (O'Connor et al., 2010; Ceron et al., 2014; Wang et al., 2012), while some studies further extend to nonverbal or multimodal dimensions (Joo et al., 2014; Won et al., 2017; Chen et al., 2020a).

2.2 Stance Detection

Stance detection is a relevant NLP problem that aims to predict a claim's stance on reference text (Augenstein et al., 2016) or to infer social media users' view toward an issue (Darwish et al., 2020). Many studies tried to advance the deep learning-based models (Mohtarami et al., 2018), for example, by modeling text with a hierarchical architecture (Sun et al., 2018; Yoon et al., 2019). Unlike stance detection, this study aims at understanding an entity's sentiment toward another entity, both of which appear in the same sentence.

2.3 Relation Extraction

Our target problem is also relevant to relation extraction, which is a task of extracting structured relationships between entities from unstructured text. While the early literature relied on feature-based methods (Zelenko et al., 2003), recent methods actively utilize neural methods (Lin et al., 2016); for example, a study proposed a neural model that jointly learns to perform entity recognition and relation extraction (Bekoulis et al., 2018). Most recently, a study tested the use of the pretrained transformer-based language model for relation extraction (Zhang et al., 2020).

Despite its similarity to directed sentiment extraction, most of the existing datasets only consider *explicit* entity relationships such as EMPLOYEE_OF and CITY_OF_RESIDENCE (Zhang et al., 2017). Understanding sentiment relationships between political entities is more challenging as their sentiment is usually hidden in text.

3 Problem and Dataset

In this section, we introduce our problem formulation and explain the process of our dataset construction and annotation.

3.1 Target Problem

Given a sentence s that contains two entities p and q, the *directed sentiment extraction* problem aims to detect the sentiment relation from p to q among five classes: neutral, p holds a positive or negative opinion towards q, and the reverse direction. For example, in the given sentence in Figure 1, the model should infer that Trump is the source of the negative sentiment toward China, the target. Existing approaches for sentiment analysis cannot be easily adapted to the task, as existing methods aim to detect polarity embedded in a text (sentiment classification), for a specific target (targeted), or with regard to an aspect (aspect-based). These problem setups do not consider the source and target of the sentiment at a time, which cannot identify directed sentiment relationships between political entities.

Class	Count
Neutral	10,604
Positive ($p \rightarrow q$)	1,656
Positive ($p \leftarrow q$)	327
Negative ($p \rightarrow q$)	3,163
Negative ($p \leftarrow q$)	478
Total	16,228

Table 1: Dataset statistics

In the following, we introduce a new annotated corpus for the problem.

3.2 Data Collection

To construct our dataset, we used news articles from the Real News corpus (Zellers et al., 2019), which consists of 32,797,763 real news stories in English published by various outlets over multiple years. Among them, we used 7,127,692 news articles shared by news media that are verified as trustworthy by Media Bias/Fact Check (Media Bias Fact Check, 2015). After splitting each article into multiple sentences, we only took sentences with two or more entities using the named entity recognition tool in Spacy[2]. To focus on the relationships between political entities, we considered named entities identified as people, countries/political groups, organizations, or cities/states[3].

Since most entity relationships in regular sentences are presumably neutral, we took two approaches for sampling sentences that cover diverse relationships: (i) dictionary-based and (ii) random selection approaches. The dictionary-based approach filters in sentences containing positive or negative keywords from a pre-defined dictionary. Starting from the blame-related keywords (Liang et al., 2019), we extended the dictionary by adding their synonyms and antonyms. While this method is effective in sampling from an unbalanced dataset, it excludes sentences that do not explicitly mention a blame or support keyword. Thus, we also randomly drew sentences to improve the dataset coverage.

3.3 Crowdsourced Annotation

We used Amazon Mechanical Turk (AMT) to annotate each sentence. The annotation task asked workers to identify what sentiment does an entity holds toward another in a given sentence. We instructed them to annotate a sentence based only on the sentiment expressed within the sentence, without relying on prior knowledge. There are five options to choose from, neutral, positive ($p \rightarrow q$), positive ($q \rightarrow p$), negative ($p \rightarrow q$), and negative ($q \rightarrow p$). p is the preceding entity of q, and the arrow indicates sentiment direction between the two entities. The detailed instruction used for educating annotator is presented in Table A1 in Appendix.

We hired five workers to annotate each sentence. For ensuring high-quality responses, we only allowed workers to participate in the task when they had at least a 70% acceptance rate for more than 1,000 previous annotations. After completing the initial round of annotation, we filtered out unreliable responses completed within one second or responses by workers who did not pass test questions. We designated workers with at least one unreliable response as untrustworthy and discarded all the other responses submitted by the untrustworthy workers. We repeated the annotation task for the discarded answers until we have five annotations for every sentence. The final set of annotations indicates a Fleiss' kappa value of 0.26, indicating an acceptable level of agreement among annotators. The level of reliability is comparable to studies using subjective text annotations such as hate speech annotation (Ross et al., 2017), subjectivity (Abdul-Mageed and Diab, 2011), and sentiment analysis (Park et al., 2018). By aggregating five responses for each sentence by a majority vote, we obtained the final dataset of 16,228 sentences of which sentiment direction is annotated, as shown in Table 1. While keeping the label distribution almost identical, we split the dataset into 13144, 1461, and 1623 instances for train, validation, and test set, respectively.

4 Methods

This section presents methods for addressing the problem of directed sentiment extraction. We propose a novel approach for solving it by employing augmented inputs in BERT-like transformer models and compare it with classification approaches.

4.1 Classification Approaches

Following the standard in using classification setups for (undirected) sentiment detection (Liu, 2012; Devlin et al., 2019), we construct classification models that predict scores of each sentiment for the directed sentiment extraction.

[2]https://spacy.io

[3]https://spacy.io/api/annotation#named-entities

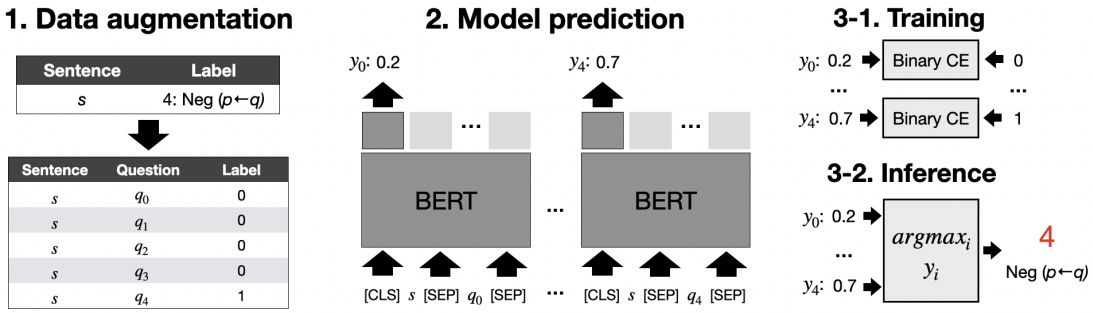

Figure 2: Detailed illustration of the DSE2QA approach

4.1.1 Existing Methods

We first construct previous approaches proposed for blame relationship detection. Liang et al. (2019) proposed three neural models that predict the classes of blame relationships in a given text: $p \rightarrow q$, $p \leftarrow q$, and no relationship. We extend their approaches to the five classes of directed sentiment in our dataset.

Entity prior model exploits the prior knowledge on political entities by using the representation of two entities using a pre-trained word embedding: $\mathbf{e}_p$ and $\mathbf{e}_q$. After concatenation, the two vectors are fed into a fully-connected network with a ReLU hidden layer to make a final prediction. *Context* model utilizes the context of the target sentence where two entities appear. After replacing p and q with special tokens ([ENT1] and [ENT2]) respectively, the model encodes the input text through a bidirectional LSTM and extracts the representation corresponding to the two entities: $\mathbf{lstm}_p$ and $\mathbf{lstm}_q$. The representation gets fed into a fully-connected network. *Combined* model utilizes the concatenated representation of the outputs of the entity prior and context model: $\mathbf{e}_p$, $\mathbf{e}_q$, $\mathbf{lstm}_p$, and $\mathbf{lstm}_q$. Then, the vector is fed into a fully-connected network. We train the existing models by minimizing the cross-entropy loss.

4.1.2 Fine-Tuning a Pretrained Transformer

Fine-tuning a BERT-like pretrained transformer has shown significant performance in many downstream tasks (Devlin et al., 2019). We also evaluate the performance of a fine-tuned transformer. In particular, after replacing the tokens corresponding to p and q with [ENT1] and [ENT2], we train a classification model that predicts the five-class output based on the representation of [CLS][4]. We use the RoBERTa base model (Liu et al., 2019) as backbone, and we refer to the classification model as *RoBERTa* in evaluation experiments. The model is trained to minimize the cross-entropy loss.

4.2 Proposed Approach: Directed Sentiment Extraction to Question-Answering

BERT-like transformer models are pretrained using two inputs including a separator token[5] with varying training objectives. The input configuration allows the model to be successfully transferred to the tasks using an auxiliary input, such as sentence pair classification (sentence 1 and sentence 2) and question answering (reference and question). Inspired by the recent achievements using auxiliary inputs (Sun et al., 2019; Cohen et al., 2020), we propose a simple but effective approach of tackling the directed sentiment extraction problem, which we call **DSE2QA**; we transform the sentiment extraction task into the sub-tasks aiming for answering yes/no questions on whether a target sentiment is embedded in the text. The basic idea is we inquire an intelligent machine who can answer yes/no questions on whether a target sentiment exists and then combine the answers corresponding to the each sentiment class for making a final guess. Figure 2 presents the overall framework, which we elaborate on each step in the following. Technically, taking auxiliary input in BERT-like transformers enables implementing the intelligent machine by making a different prediction with the same sentence input, according to the question fed as additional input. We hypothesize that a large-scale pretrained transformer model on a massive corpus can understand the meaning of the augmented question and thus successfully answer whether a directed sentiment exists in a text.

Note that our question-answering setup is different from standard question-answering tasks in NLP, as represented by well-known benchmark data such as SQuAD (Rajpurkar et al., 2016) and

[4] <s> in RoBERTa

[5] [CLS] in BERT, </s></s> in RoBERTa

Index	Complete questions	Pseudo questions
q_0	Do [Ent1] and [Ent2] have neutral sentiment toward each other?	[Ent1] - [Ent2] - neutral
q_1	Does [Ent1] has positive sentiment toward [Ent2]?	[Ent1] - [Ent2] - positive
q_2	Does [Ent2] has positive sentiment toward [Ent1]?	[Ent2] - [Ent1] - positive
q_3	Does [Ent1] has negative sentiment toward [Ent2]?	[Ent1] - [Ent2] - negative
q_4	Does [Ent2] has negative sentiment toward [Ent1]?	[Ent2] - [Ent1] - negative

Table 2: Auxiliary questions according to the target label

WikiQA (Yang et al., 2015). Given a question and reference text, the standard task aims at generating answers in a natural language form. In contrast, the question-answering process in DSE2QA requires a binary answer, which can be seen as a special type of question-answering.

4.2.1 Data Augmentation for DSE2QA

For each pair of label l and sentence s where the two target entities p and q are masked as [ENT1] and [ENT2] respectively, we augment the training data by transforming the original input into the five tuples using the same sentence and different questions: t_i: (s, q_i, l_i) where i is the index of the target relation class: neutral (0), $p \rightarrow q$ with positive sentiment (1), $p \leftarrow q$ with positive sentiment (2), $p \rightarrow q$ with negative sentiment (3), and $p \leftarrow q$ with negative sentiment (4). l_i becomes 1 if l is i; otherwise l_i is 0.

We design the auxiliary question q_i asking whether the given sentence s is classified as the target sentiment i. The list of questions are presented in Table 2. For example, q_1 asks a model whether a given sentence s contains positive sentiment from p to q. Here, we define two types of questions: complete and pseudo. Complete questions are written in a natural language, and pseudo questions only contain keywords that is sufficient to characterize a sentiment class i while ignoring the syntactic structure.

4.2.2 Model Prediction

We utilize the BERT-like transformer model (Devlin et al., 2019), which can take sentence pairs as input, for making a binary prediction on a given sentence s and question q_i. In particular, the model takes

$$\text{[CLS]}\ s\ \text{[SEP]}\ q_i\ \text{[SEP]}$$

as input[6] and predicts a value y_i from 0 to 1 that indicates the confidence on the target label i.

[6] '<s> s </s> </s> q_i </s> </s>' in RoBERTa

4.2.3 Training and Inference

For the augmented input of t_i, a pretrained BERT-like transformer is trained to predict 1 for t_l and 0 for $t_{i \neq l}$ through the [CLS] representation at the last layer of the transformer model followed by a classification layer. For inference, we made a prediction corresponding to s by $\text{argmax}_i\ y_i$ where y_i is the prediction outcome of t_i. The y_i indicates the confidence on each sentiment i, and therefore we take the class of which the value is maximum.

In the experiments, we utilize the RoBERTa base model for the backbone of DSE2QA and train the model to minimize the binary cross-entropy loss. This approach is different from the RoBERTa classification model that only employs a single sentence input.

5 Performance Evaluation

We evaluate the performance of the proposed DSE2QA approach using our annotated corpus. We compare our method with the current state-of-the-art methods for directional blame detection proposed by Liang et al. (2019) (LNZ) as well as a classification model fine-tuned on a pretrained transformer (RoBERTa).

5.1 Experiment Setups

For the LNZ models (Liang et al., 2019), we set the vocabulary size as 10000. We set the word embedding size, LSTM hidden dimension, and the fully connected layer dimension as 256, 512, and 1024. The search space for the dropout rate is $[0.1, 0.5]$. We train the LNZ models using Adam optimizer with a learning rate of 1e-3 (Kingma and Ba, 2014). We adopt an early stopping strategy with the patience of 5. For training RoBERTa and DSE2QA, we followed the procedure of Liu et al. (2019) using AdamW (Loshchilov and Hutter, 2017). We optimize hyperparameters by randomly choosing ten sets for each model and selecting the model with the best performance on the validation set. The learning rate is set to 2e-5 with the epsilon as

Method	Micro F1	Macro F1	mAP
DSE2QA (Pseudo)	**0.7973**	**0.6766**	**0.7488**
DSE2QA (Complete)	0.7726	0.6617	0.7387
RoBERTa	0.7486	0.6409	0.7319
LNZ (Combined)	0.7055	0.5358	0.5295
LNZ (Context)	0.6371	0.4665	0.4921
LNZ (EntityPrior)	0.5853	0.4063	0.414

Table 3: Evaluated performance on the test set. Top performance for each metric is marked as bold.

Method	0	1	2	3	4
DSE2QA (Pseudo)	**0.855**	**0.6519**	**0.5672**	0.7402	**0.5686**
DSE2QA (Complete)	0.8293	0.6421	0.5672	**0.7416**	0.5283
RoBERTa	0.8054	0.6373	0.5079	0.7184	0.5354
LNZ (Combined)	0.7981	0.443	0.3333	0.5827	0.5217
LNZ (Context)	0.7469	0.4069	0.2817	0.5007	0.3964
LNZ (EntityPrior)	0.7133	0.2629	0.2353	0.4533	0.3667

Table 4: F1-score per class measured on the test set. Top performance for each metric is marked as bold.

1e-6. The weight decay is set to 0.1. We apply random oversampling to the training set to make a balanced dataset against the label. We run the experiment five times with different random seeds and report the average scores.

5.2 Evaluation Results

We utilize three measures for evaluation: micro-f1, macro-f1, and mean average precision (mAP). Micro-f1 is calculated by (#correct)/(#total), which corresponds to the multi-class classification accuracy. Macro-f1 measures an f1-score for each class and averages them with equal importance; therefore, macro-f1 is a more robust measure to a skewed class distribution, such as our annotation data (see Table 1). Similarly, mAP measures the unweighted average of average precision (AP) on each class; AP summarizes a precision-recall curve varying prediction threshold for a target class.

In Table 3, we make three observations. First, among classification approaches (the bottom four rows), RoBERTa outperforms the other approaches across the three measures (0.7486/0.6409/0.7319). The LNZ combined model achieves a fair micro-f1 score of 0.7055 but low scores of macro-f1 (0.5358) and mAP(0.5295). This difference is because the combined model (and other non-transformer models) is poor at classifying non-neutral sentiment, which we will further investigate in Table 4. Second, DSE2QA with complete questions outperforms RoBERTa with a margin of 0.024 by micro F1. The proposed approach also achieves better performance in macro F1 and mAP. Third, the performance of DSE2QA gets further increased with the usage of pseudo questions, up to the micro-f1 score of 0.7973. This observation implies that a BERT-like transformer model may not need a full sentence to utilize the auxiliary input because it also performs well using fewer keywords for the detection task with an augmented input.

Table 4 presents the f1-score measured for each class: neutral (0), positive from the left entity to the right (1), positive from the right to the left (2), negative from the left to the right (3), and negative from the right to the left (4). Here, we make three observations. First, all models perform the best at identifying neutral sentiment (0) compared to the other sentiment classes. Second, non-transformer models (the bottom three rows) are poor at extracting non-neutral sentiment regardless of their direction, which contributes to the decreased macro F1 in Table 3. Third, among the sentiment classes from the left entity to the right entity (1, 3), transformer models better detect negative sentiment than positive sentiment. The finding suggests that positive entity relationships are more difficult to be captured in news articles, which calls for future studies for a better understanding and model improvement. AP per each class also shows a similar trend, as presented in Table A3.

In summary, the proposed approach of solving the directed sentiment extraction task by multiple question-answering tasks outperforms the state-of-the-art classification approaches. The high perfor-

mance suggests that the model can understand the meaning of augmented questions to some extent, which may build on the language understanding ability of the pretrained RoBERTa.

6 Analyzing Entity Relationships in News Articles

To demonstrate the utility of the proposed dataset and model, we conduct two case studies to analyze entity-to-entity sentiment relationships presented in recent news articles on political issues: the 2016 U.S. presidential election and the COVID-19 pandemic. For the analyses, we utilize the DSE2QA model with pseudo questions trained on the annotated corpus, and we confirm that the target news articles are not overlapped with the whole set.

6.1 Case Study 1: 2016 U.S. election

We study how news media covered the entity relation during the 2016 United States presidential election using a public dataset on news articles between Feb. 2016 to Feb. 2017[7]. This dataset consists of about 140K news articles in English from fifteen media companies, including CNN, New York Times, and Guardian. We randomly select 3K articles from each month, 39K articles in total. Then we apply the proposed model to all sentences that contain at least two entities from the top-30 most frequent entities, including Donald Trump and Hillary Clinton.

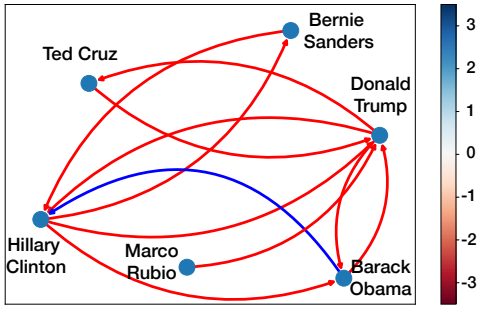

Figure 3: Log ratios of positive (blue) and negative (red) sentiments in directed political relationships in the news articles.

Figure 3 presents the frequently mentioned pairs of politicians. The red color indicates an entity pair tends to contain negative sentiment more than positive, and blue indicates the pairs with the negative sentiment more. The opacity represents the strength of the opinion measured by the log ratio of inferred pairs of positive and negative sentiment. For brevity, we present relations that appear at least 20 times in any sentiment. Overall, there are 4.68 times more negative sentiments found than positive ones, which may be explained due to the negative nature of political media (Soroka, 2014). An interesting observation is the asymmetric relation between Hillary Clinton and Barack Obama. Clinton holds a generally slightly negative opinion towards Obama, while Obama holds a stronger positive opinion towards Clinton. Note that our model does not just memorize the entity relationship in the training dataset and apply it to the target dataset as we replace detected entities with tokens in input sentences.

[7]https://www.kaggle.com/snapcrack/all-the-news

Left	Center	Right
ABC News	Associated Press	Breitbart News
BuzzFeed News	Forbes	Daily Mail
CBS News	NPR	Fox News
CNN	Reuter	National Review
Democracy Now	USA TODAY	New York Post
HuffPost		Reason
MSNBC		The American Spectator
New York Times		theblaze.com
Slate		The Daily Caller
The Atlantic		The Daily Wire
The Guardian		The Epoch Times
The New Yorker		The Federalist
Time Magazine		Washington Times
Vox		Wall Street Journal
Washington Post		

Table 5: The list target media outlets sorted by alphabetical order.

6.2 Case Study 2: the COVID-19 pandemic

In the second study, we investigate how news media portray political entities and their relationships differently according to their political orientations. For example, a right-leaning news outlet may show more negative opinions expressed toward Democrats. To that end, we focused on the recent issue of the COVID-19 pandemic to examine the media bias. We expect the general sentiment about COVID-19 is negative but would like to investigate who blames whom because the messages will have very different meanings according to the sources and targets, differentiated by our data and method.

To collect a recent news article set, we first compiled a list of 35 popular news media outlets that cover American politics in English. We also ensure the list of news outlets was balanced against political bias, according to the media bias ratings in allsides.com. For brevity, we consolidate 'Lean

Entity	Rank (Left)	Rank (Center)	Rank (Right)
Donald Trump	1	2	1
Republican	2	6	6
U.S.	3	1	3
Democrat	4	5	4
Russia	5	6	9
American	6	15	10
China	7	4	5
Obama	8	10	7
Chinese	9	3	2
Joe Biden	9	8	7

Table 6: Frequency rank of top-10 frequent entities targeted with a negative sentiment through an entity-to-entity relationship in the COVID-19 news dataset. The order is sorted by the overall rank in the dataset.

left' and 'Left' into 'Left' and 'Lean right' and 'Right' into 'Right.' Table 5 presents the list of the target media.

For each of the target media outlets, we collected news articles shared throughout 2020 until September from the Common Crawl corpus, which has been collecting web data since 2008[8]. The total number of retrieved news pages are is 256,081; on average, we have 7,113 published news articles for each outlet.

Next, we selected documents containing at least one of the keywords relevant to COVID-19 by following similar practices used for collecting a Twitter dataset (Chen et al., 2020b): *coronavirus*, *covid-19*, *COVID19*, and *corona virus*. We consider sentences containing two or more entities for the target of inference, and every relationship pair is inferred when there are more than two entities in a sentence. The final set consists of 6,180 articles involving 1,078,377 entity pairs for COVID-19.

Table 6 presents the list of 10 frequent entities that are manifested through entity-to-entity relationships with a negative sentiment. While Donald Trump was the most frequent target of blames in the total data, the results show that the right-leaning media tend to express a negative sentiment toward China (#5) and Chinese people (#2) more frequently. To examine the difference systematically, we measure the spearman rank correlation coefficient for the whole list of entities that appeared in the dataset. The rank in the left-leaning media and that in right-leaning media exhibits a highly negative correlation of -0.5722 ($p<0.001$), which suggests that the list of political entities presented with a negative sentiment significantly varies across news media according to their political orientation. Such a high level of negative correlation is also observed in the ranks for the source entity in negative relationships (-0.4129 with $p<0.001$) and the source/target entities in positive relationships (-0.5605/-0.7929 with $p<0.001$).

Going further, we analyze the differences between frequently presented entity pairs with negative sentiment by the left and right-leaning media, respectively. Table 7 presents the rank of each entity pair in the media groups according to their political orientation. In the left-leaning media of our dataset, Donald Trump appears as either source or target in the top-10 frequent negative pairs except for the pair of Republican→Democrat and vice versa. On the contrary, the top-10 pairs in the right-leaning media include the international relationships of Donald Trump to the other countries. In other words, the left-leaning media may try to frame COVID-19 as a **domestic event** focusing on how the President handles it and how people respond to his crisis management, and the right-leaning media focus more on **foreign policies and international relationship** especially between the U.S. and China. For the whole set of entity pairs, the rank correlation between the left and right media is -0.4847 ($p<0.001$) for negative sentiment and -0.7929 ($p<0.001$) for positive sentiment. These negative correlations imply that the news media has a bias in selecting issues to cover (selection bias) and presenting relationships of political entities (presentation bias).

7 Conclusion

Detecting who blames or endorses whom in news articles is a critical ability in understanding opinions and relationship between political actors in news media. This paper provides a computational tool based on natural language processing for facilitating interdisciplinary studies using text in news and social media.

We introduced a new problem of identifying directed sentiment relationships between political entities, called directed sentiment extraction. We constructed a training corpus of which entity relationship is manually annotated for each sentence. This dataset can serve as a benchmark for future studies. A potential future direction is to build an improved version of the dataset where sentiment re-

[8]https://commoncrawl.org/

Frequent pairs in the left-leaning media				Frequent pairs in the right-leaning media			
Entity pairs	Rank (Left)	Rank (Center)	Rank (Right)	Entity pairs	Rank (Left)	Rank (Center)	Rank (Right)
Democrat→ Donald Trump	1	8	1	Democrat→Donald Trump	1	8	1
Republican→Donald Trump	2	66	4	Donald Trump→Democrat	4	3	2
Twitter→Donald Trump	3	67	392	Donald Trump→Chinese	30	17	3
Donald Trump→Democrat	4	3	2	Republican→Donald Trump	2	66	4
Bernie Sander→Donald Trump	5	67	153	Donald Trump→China	9	10	5
Donald Trump→Ted Cruz	6	67	393	Donald Trump→Joe Biden	20	8	6
Republican→Democrat	7	66	47	Democrat→Republican	8	41	7
Democrat→Republican	8	41	7	Joe Biden→Donald Trump	38	5	8
Donald Trump→China	9	10	5	Donald Trump→Russia	23	17	9
Donald Trump→Bush	10	67	393	House→Donald Trump	12	67	10

Table 7: Frequency rank of entity pairs presented with a negative sentiment in the COVID-19 news dataset

lationships between political entities appear across multiple sentences in news articles.

To tackle the problem, we proposed DSE2QA (Directed Sentiment Extraction to Question-Answering), which is a simple yet effective method of utilizing BERT-like pretrained transformers by predicting answers for binary questions on whether a sentiment relationship exists in a given text. Answers for each sentiment class are aggregated to make a final guess. Evaluation experiments show the approach outperforms state-of-the-art classification models, such as the fine-tuned RoBERTa classification model. We hypothesize the language understanding ability of the BERT-like pretrained transformer may contribute to the high performance, combined with its facility of taking auxiliary input. Furthermore, the performance increase with the pseudo questions implies that a few keywords may suffice to make an inquiry. Future research could investigate which kind of pretrained transformer is the most effective for understanding the meaning of the augmented question, as DSE2QA's backbone can be replaced with any BERT-like transformer. Also, this study calls for future studies on advanced methods for directed sentiment extraction. A potential approach could jointly learn entity recognition and directed sentiment extraction as similarly tackled by a study on information extraction (Bekoulis et al., 2018).

As the last step, we conducted case studies by analyzing directed sentiments in news text for the US election and COVID-19 pandemic. The observations not only add empirical understandings to the social science research but also highlight the utility of the proposed problem, dataset, and model for political news analysis. We believe the proposed method can therefore further the current interdisciplinary efforts of the NLP, machine learning, and the social science communities (Grimmer and Stewart, 2013; Roberts et al., 2014; Joo and Steinert-Threlkeld, 2018).

Acknowledgements

We thank anonymous reviewers for their valuable comments. This work was supported by NSF SBE/SMA #1831848 "RIDIR: Integrated Communication Database and Computational Tools".

Ethics and Impact Statement

In online news and social media, people express diverse sentiment toward a target through text, such as blame, support, endorsement, to list a few. Quantifying and understanding the patterns is of significant interest in social science, but the lack of automated methods makes it difficult to handle large-scale data, which can reveal patterns in a comprehensive view. In this light, this study aims to develop automated methods of identifying directed sentiment between entities by defining a new NLP problem: directed sentiment extraction. The newly annotated dataset will facilitate future development of the NLP methods, and the DSE2QA approach will serve as a strong baseline.

The development of an automated method will have a broader impact by tackling real-world challenges such as bias in news reporting against political orientation, with potential collaboration with social science. Moreover, it will enable the discovery of hidden biases with regard to sentiment in online text, which can be mistakenly learned through data-driven methods. A fine-grained understanding of sentiment relationships will broadly contribute to building a fair machine learning model, which is of significant interest in AI ethics.

References

Muhammad Abdul-Mageed and Mona Diab. 2011. Subjectivity and sentiment annotation of modern standard arabic newswire. In *Proceedings of the 5th linguistic annotation workshop*, pages 110–118.

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 876–885.

Alexandra Balahur and Ralf Steinberger. 2009. Rethinking sentiment analysis in the news: from theory to practice and back. *Proceeding of WOMSA*, 9.

Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. 2018. Joint entity recognition and relation extraction as a multi-head selection problem. *Expert Systems with Applications*, 114:34–45.

Andrea Ceron, Luigi Curini, Stefano M Iacus, and Giuseppe Porro. 2014. Every tweet counts? how sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to italy and france. *New media & society*, 16(2):340–358.

Danni Chen, Kunwoo Park, and Jungseock Joo. 2020a. Understanding gender stereotypes and electoral success from visual self-presentations of politicians in social media. In *Joint Workshop on Aesthetic and Technical Quality Assessment of Multimedia and Media Analytics for Societal Trends*, pages 21–25.

Emily Chen, Kristina Lerman, and Emilio Ferrara. 2020b. Covid-19: The first public coronavirus twitter dataset. *arXiv preprint arXiv:2003.07372*.

Amir DN Cohen, Shachar Rosenman, and Yoav Goldberg. 2020. Relation extraction as two-way span-prediction. *arXiv preprint arXiv:2010.04829*.

Keith Cortis, André Freitas, Tobias Daudert, Manuela Huerlimann, Manel Zarrouk, Siegfried Handschuh, and Brian Davis. 2017. SemEval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 519–535.

Kareem Darwish, Peter Stefanov, Michaël Aupetit, and Preslav Nakov. 2020. Unsupervised user stance detection on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 141–152.

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.

CHE Gilbert and Erric Hutto. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, volume 81, page 82.

Justin Grimmer and Brandon M Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, 21(3):267–297.

Jungseock Joo, Weixin Li, Francis F Steen, and Song-Chun Zhu. 2014. Visual persuasion: Inferring communicative intents of images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 216–223.

Jungseock Joo and Zachary C Steinert-Threlkeld. 2018. Image as data: Automated visual content analysis for political science. *arXiv preprint arXiv:1810.01544*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Shuailong Liang, Olivia Nicol, and Yue Zhang. 2019. Who blames whom in a crisis? detecting blame ties from news articles using neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 655–662.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2124–2133.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150.

LLC Media Bias Fact Check. 2015. Media Bias/Fact Check. https://mediabiasfactcheck.com. [Online; accessed 21-Sep-2020].

Mitra Mohtarami, Ramy Baly, James Glass, Preslav Nakov, Lluís Màrquez, and Alessandro Moschitti. 2018. Automatic stance detection using end-to-end memory networks. 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 767–776.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan Routledge, and Noah Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 4.

Kunwoo Park, Meeyoung Cha, and Eunhee Rhim. 2018. Positivity bias in customer satisfaction ratings. In *Companion Proceedings of the The Web Conference 2018*, pages 631–638.

Maria Pontiki, Dimitrios Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *10th International Workshop on Semantic Evaluation*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Margaret E Roberts, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G Rand. 2014. Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58(4):1064–1082.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint arXiv:1701.08118*.

Stuart N Soroka. 2014. *Negativity in democratic politics: Causes and consequences*. Cambridge University Press.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. 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 380–385.

Qingying Sun, Zhongqing Wang, Qiaoming Zhu, and Guodong Zhou. 2018. Stance detection with hierarchical attention network. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2399–2409.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. Semeval-2018 task 3: Irony detection in english tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50.

Hao Wang, Dogan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan. 2012. A system for real-time Twitter sentiment analysis of 2012 U.S. presidential election cycle. In *Proceedings of the ACL 2012 System Demonstrations*, pages 115–120.

Donghyeon Won, Zachary C Steinert-Threlkeld, and Jungseock Joo. 2017. Protest activity detection and perceived violence estimation from social media images. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 786–794.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018.

Seunghyun Yoon, Kunwoo Park, Joongbo Shin, Hongjun Lim, Seungpil Won, Meeyoung Cha, and Kyomin Jung. 2019. Detecting incongruity between news headline and body text via a deep hierarchical encoder. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 791–800.

Lori Young and Stuart Soroka. 2012. Affective news: The automated coding of sentiment in political texts. *Political Communication*, 29(2):205–231.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. 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 1415–1420.

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2003. Kernel methods for relation extraction. *Journal of machine learning research*, 3(Feb):1083–1106.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems*, pages 9051–9062.

Ningyu Zhang, Luoqiu Li, Shumin Deng, Haiyang Yu, Xu Cheng, Wei Zhang, and Huajun Chen. 2020. Can fine-tuning pre-trained models lead to perfect nlp? a study of the generalizability of relation extraction. *arXiv preprint arXiv:2009.06206*.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45.

Instruction:
Based the given sentence, please identify if one entity (person, organization, or country) holds a positive or negative opinion towards another entity. There will be two entities in total. One is in red and the other is in blue.

It is also possible neither positive nor negative opinion exists in the sentence.
Please annotate such cases as neutral.

Please classify the sentence based on what people say instead of what they do.
For example, if a sentence only states the police arrest someone, this sentence should be classified as neutral. Instead, if the police accuses someone of commiting a crime, this sentence should be classified as police holds a negative opinion towards the person.

Examples:
- Earlier on Tuesday, Mr. Trump criticized General Motors for making cars in Mexico.
(Negative: Trump holds a negative opinion towards General Motors)
- Hugo Ras has been accused of killing other people's rhino, and for that South Africans condemn him.
(Negative: South Africans holds a negative opinion towards Hugo Ras)
- DAVID Cameron's accused the Conservatives of failing to devolve essential welfare powers to, as agreed by the cross-party Smith Commission which considered further powers for the Scottish parliament last year.
(Neutral: There is no direct opinions between these two entities.)
- Prime Minister Stephen Harper shakes hands with Petty Harbour, N.L., during a campaign event in Toronto on Sept. 18, 2015. (Neutral: They shake hands just for politeness. No opinions exist.)
- Obama pulled Clinton into his administration after he defeated her in 2008 primary and has effusively praised her tenure as secretary of state. (Positive. Obama holds a positive opinion towards Clinton.)
- Earlier on Tuesday, Mr. Trump has not comments on General Motors making cars in Mexico. (Neutral.)

Note:
There are multiple annotators for each sentence. Your response will be judged as failed when it is different with other annotators. If the percentage of failed response from one annotator is above a threshold, the annotator will NOT get paid for ALL responses. Thanks for your participation.

Table A1: Instruction used for educating annotators in Amazon Mechanical Turk.

Method	Num. parameters	Avg. runtime per epoch	Micro F1	Macro F1	mAP
DSE2QA (Pseudo)	125M + 1536	2154s	**0.8072**	**0.6827**	**0.7528**
DSE2QA (Complete)	125M + 1536	2149s	0.7892	0.6751	0.7724
RoBERTa	125M + 3840	437s	0.7618	0.6516	0.7493
LNZ (Combined)	3.03M	12s	0.694	0.5189	0.4819
LNZ (Context)	2.9M	12s	0.6331	0.4518	0.3908
LNZ (EntityPrior)	2.65M	4.8s	0.5914	0.4427	0.31

Table A2: Model details and evaluated performance on the validation set. Top performance for each metric is marked as bold.

Method	0	1	2	3	4
DSE2QA (Pseudo)	0.9316	0.7157	**0.5952**	0.8358	**0.6658**
DSE2QA (Complete)	**0.9341**	**0.7228**	0.5747	**0.8457**	0.6161
RoBERTa	0.929	0.7299	0.5236	0.8232	0.6536
LNZ (Combined)	0.8452	0.4554	0.2887	0.6273	0.4311
LNZ (Context)	0.8233	0.4261	0.2887	0.568	0.3545
LNZ (EntityPrior)	0.7834	0.2181	0.2248	0.4405	0.4033

Table A3: AP per class measured on the test set. Top performance for each metric is marked as bold.