Stance Prediction for Contemporary Issues: Data and Experiments

Marjan Hosseinia
University of Houston
mhosseinia@uh.edu

Eduard Dragut
Temple University
edragut@temple.edu

Arjun Mukherjee
University of Houston
arjun@cs.uh.edu

Abstract

We investigate whether pre-trained bidirectional transformers with sentiment and emotion information improve stance detection in long discussions of contemporary issues. As a part of this work, we create a novel stance detection dataset covering 419 different controversial issues and their related pros and cons collected by procon.org in nonpartisan format. Experimental results show that a shallow recurrent neural network with sentiment or emotion information can reach competitive results compared to fine-tuned BERT with $20\times$ fewer parameters. We also use a simple approach that explains which input phrases contribute to stance detection.

1 Introduction

Stance detection identifies whether an opinion is in favor of an idea or opposes it. It has a tight connection with sentiment analysis; however, stance detection usually investigates the two-sided relationship between an opinion and a question. For example, 'should abortion be legal?' or 'is human activity primarily responsible for global climate change?'

Contemporary debatable issues, even though non-political, usually carry some political weight and controversy. For example, legislators may allow soda vending machines in our school or consider obesity as a health issue that directly impacts soda manufacturers and insurance companies respectively. On a larger scale, an issue such as climate change is being discussed in US presidential debates constantly. Meanwhile, information about these issues is mostly one-sided and provided by left or right partisan resources. Such information forms public beliefs, has persuasive power, and promotes confirmation bias (Stanojevic et al., 2019), the humans' tendency to search for the information which confirms their existing beliefs [1]. Confirmation bias permits internet debates and promote discrimination, misinformation, and hate speech, all of which are emerging problems in user posts of social media platforms.

Although there are many attempts to automatic identification and removal of such contents from online platforms, the need for accessing bi-partisan information that cultivates critical thinking and avoids confirmation bias remains. In this regard, a few web sources, such as procon.org, present information in a non-partisan format and being used as a resource for improving critical thinking in educational training by teachers [2].

Here, we aim to improve such resources by automatic stance detection of pro or con-perspectives regarding a debatable issue. We extend our previous work (Hosseinia et al., 2019) by creating a new dataset from procon.org with 419 distinct issues and their two-sided perspectives annotated by its experts [3]. Then, we leverage external knowledge to identify the stance of a perspective towards an issue that is mainly represented in the form of a question.

The latest progress in pre-trained language models (Howard and Ruder, 2018) and transformers (Devlin et al., 2019; Yang et al., 2019) allows one to create general models with less amount of effort for task-specific text classification. In this work, we show that bidirectional transformers can produce competitive results even without fine-tuning by leveraging auxiliary sentiment and emotion information (Dragut et al., 2010). Experimental results show the effectiveness of our model and its remarkable performance. The model has a signif-

---

[1] www.procon.org/education.php
[2] https://www.procon.org/view.background-resource.php?resourceID=004241
[3] https://github.com/marjanhs/procon20/

icantly smaller size compared to the BERT-base model.

The main contributions of this work are as following:

- Proposing a simple but efficient recurrent neural network that leverages sentence-wise sentiment or token-level emotion of input sequence with BERT representation for detecting the stance of a long perspective against its related question.

- Creating a novel dataset for stance detection with more than 6K instances.

- Explaining the word/phrase contribution of input sequence using max-pooling engagement score for stance detection.

2   Related Works

We group stance detection methods based on underlying data and approaches as follows:

- *Tweets* are collected from SemEval 2016, Task 6, (Mohammad et al., 2016) and organized in two categories. The first category, which represents a supervised setting, includes tweets that cover opinions about five topics, "Atheism", "Climate Change", "Feminist Movement", "Hillary Clinton", and "Legalization of Abortion". The second category, which represents weakly supervised settings, includes tweets that cover one topic, but the training data is unlabeled.

- *Claims* are obtained from Wikipedia in (Bar-Haim et al., 2017). Each claim is defined as a brief statement that is often part of a Wikipedia sentence. The claim dataset contains 55 different topics.

- *Debates* are gathered from various online debate resources, including idebate, debatewise and procon in the form of perspective, claim, and evidence for substantiated perspective discovery. 49 out of its 947 claims are from procon (Chen et al., 2019). Claims and perspectives are short sentences and have been used for stance detection in (Popat et al., 2019).

Current approaches on stance detection use different types of linguistic features, including word/character n-grams, dependency parse trees,

and lexicons (Sun et al., 2018; Sridhar et al., 2015; Hasan and Ng, 2013; Walker et al.). There are also end-to-end neural network approaches that learn topics and opinions independently while joining them with memory networks (Mohtarami et al., 2018), bidirectional conditional LSTM (Augenstein et al., 2016), or neural attention (Du et al., 2017). There are also some neural network approaches that leverage lexical features (Riedel et al., 2017; Hanselowski et al., 2018). A consistency constraint is proposed to jointly model the topic and opinion using BERT architecture (Popat et al., 2019). It trains the whole massive network for label prediction. None of these approaches incorporate bidirectional transformers with sentiment and emotion in a shallow neural network as we propose in this paper. Additionally, our focus is to find the stance of 100-200 words long discussions, which are commonly present in nonpartisan format.

3   Dataset

We collect data from `procon.org`, a non-profit organization that presents opinions on controversial issues in a nonpartisan format. Issues (questions) and their related responses are professionally researched from different online platforms by its experts. The dataset covers 419 different detailed issues ranging from politics to sport and healthcare. The dataset instances are pairs of issues, in the form of *questions*, and their corresponding *perspectives* from proponents and opponents. Each *perspective* is either a pro or a con with 100-200 words that supports its claim with compelling arguments. Table 1 provides some examples of the questions from the dataset. The dataset statistics are also presented in Table 2. We may use the words *opinion* and *perspective* interchangeably as both refer to the same concept in this work.

4   Model

Utilizing pre-trained models has been widely popular in machine translation and various text classification tasks. Prior efforts were hindered by the lack of labeled data (Zhang et al., 2019). With the growth of successful pre-trained models, a model fine-tuned on a small portion of data can compete with models trained on $10\times$ more training data without pre-training (Howard and Ruder, 2018). Recently, transformer models trained on both directions of language simultaneously, such as BERT (Devlin et al., 2019) and XLNet (Yang et al., 2019),
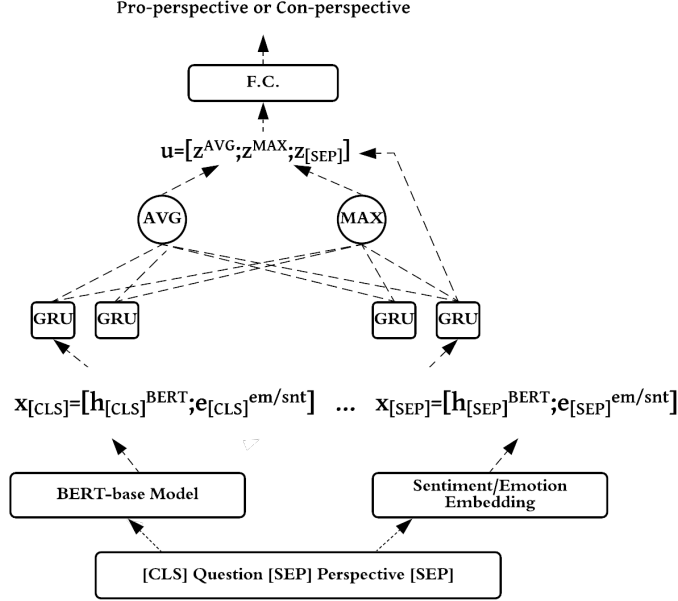
Figure 1: Stance detection architecture; snt: sentiment; em: emotion

The figure shows the architecture with: "Pro-perspective or Con-perspective" at top, an F.C. box, $u=[z^{AVG};z^{MAX};z_{[SEP]}]$, AVG and MAX circles, GRU boxes, $x_{[CLS]}=[h_{[CLS]}^{BERT};e_{[CLS]}^{em/snt}] \quad ... \quad x_{[SEP]}=[h_{[SEP]}^{BERT};e_{[SEP]}^{em/snt}]$, BERT-base Model and Sentiment/Emotion Embedding boxes, and "[CLS] Question [SEP] Perspective [SEP]".

HEALTH and MEDICINE 1- Should euthanasia or physician-assisted suicide be legal? 2- Is vaping with e-cigarettes safe?
EDUCATION 1-Should parents or other adults be able to ban books from schools and libraries? 2- Should public college be tuition-free?
POLITICS 1- Should recreational marijuana be legal? 2- Should more gun control laws be enacted?
SCIENCE and TECHNOLOGY 1- Is cell phone radiation safe? 2- Should net neutrality be restored?
ENTERTAINMENT and SPORTS 1- Are social networking sites good for our society? 2- Do violent video games contribute to youth violence?

Table 1: Procon dataset questions

overcome previous unidirectional language models (e.g. ULMFiT (Howard and Ruder, 2018)) or models trained on two independent directions (ELMo) (Peters et al., 2018) significantly. So, we build our baselines based on BERT architecture in two different ways: single and pair of inputs. A question and its related opinion are concatenated for single inputs. However, for input pairs, the question and the opinion are being separated with the BERT separator tag [SEP]. This approach has been used for question-answering applications (Devlin et al., 2019).

Opinion is connected with sentiment and emotion (Schneider and Dragut, 2015). Moreover, prior efforts show the successful employment of linguistic features, extracted with external tools, in neural networks for emotional cognition (Yang et al., 2017). So, we leverage sentiment and emotion information separately with BERT representations obtained from the last BERT-base layer to form the input of a shallow recurrent neural network. In the following, we provide the details.

- Employing sentiment: We analyze how the sentiment of sentences in proponents' and opponents' opinions can affect stance detection. Accordingly, we use a rule-based sentiment tool, VADER (Hutto and Gilbert, 2014), for obtaining the sentiment of a sentence. VADER translates its compound sentiment score, ranging from -1 to $+1$, into negative sentiment labels for scores $\leqslant -0.05$, positive labels for scores $\geqslant +0.05$, and neutral for the scores between -0.05 and $+0.05$.

Here, we compute sentence-wise sentiment using VADER to let the model learn the flow of sentiment across the opinion. So, each token borrows the sentiment of its correspond-

34

Set	#of Topics	#of Words	#of Pro-perspectives	#of Con-perspectives	Total
train	417	127	2,140	2,125	4,265
dev	265	125	326	284	610
test	336	123	613	606	1,219

Table 2: Procon dataset statistics

ing sentence. Then, an embedding layer converts the discrete labels into d-dimensional vectors ($d = 768$) using a randomly initialized matrix $W^s_{3 \times d}$; These representations are concatenated with BERT token embeddings to form the bidirectional Gated Recurrent Units ($\overleftrightarrow{\text{GRU}}$) input ($x_t$ for token t):

$$x_t = [h_t^{\text{BERT}}; e_t^{\text{snt}}],$$
$$z_t = \overleftrightarrow{\text{GRU}}(x_t),$$
$$u = [\text{avg-pool}(Z); \text{max-pool}(Z); z_T],$$
$$y = \text{softmax}(Wu + b)$$

For an input sequence with T tokens, h_t^{BERT} is the hidden state of the last BERT-base layer corresponding to the input token at time t, e_t^{snt} is sentiment embedding of the token, [;] denotes concatenation operator, $Z = [z_i]_{i=1}^T$, and W, b are parameters of a fully connected layer.

Recall that our task is to identify the stance of long opinions; So, important information towards the final stance might be anywhere in the opinion. Because of that, we collect such information from the recurrent hidden states of all input tokens using max and average-pooling. Max-pooling returns a vector with maximum weights across all hidden states of input tokens for each dimension. In this way, the input tokens with higher weights will be engaged for stance prediction. Aside from that, the last hidden state of the recurrent network (z_T) is concatenated with the pooled information (u). Finally, a dense layer transforms vector u into the class dimension. Figure 1 shows the model architecture.

We refer to this model as VADER-Sent-GRU and report the experimental results in Section 6.

- Employing emotion: We take a similar approach to engage emotion information for

stance detection using the NRC emotion lexicon (Mohammad and Turney, 2013). The Lexicon is collected by crowdsourcing and consists of English words with their eight basic emotions including anger, fear, anticipation, trust, surprise, sadness, joy, and disgust. So, the GRU input is a concatenation of BERT representation with emotion embedding (gained from a $9 \times d$ matrix with random initialization; one dimension is added for neutral emotion). Here, we use unidirectional $\overrightarrow{\text{GRU}}$ as it shows more stable results in our pilot experiments.

5 Experiments

In this section, we describe the corresponding baselines followed by the training setup.

5.1 Baselines

We use the following baselines utilized in opinion mining including sentiment analysis and stance detection:

- BERT (Devlin et al., 2019) followed by a non-linear transformation on a dense layer is used for downstream stance detection. Here, the whole network is fine-tuned and all 12 BERT-base layers' weights will be updated in back-propagation. The information is pooled from the final hidden state of the classification token ($h_{\text{[cls]}}^{\text{BERT}}$) after passing a fully connected layer with non-linear activation ($\tanh$). Then, a classifier layer shrinks the activations to a binary dimension.

$$x = \tanh(W^p h_{\text{[cls]}}^{\text{BERT}} + b^p),$$
$$y = W^c x + b^c$$

where W^c, W^p, b^p, and b^c are the layers' parameters.

- BERT$_{\text{CONS}}$ is a BERT base model that considers two different inputs using a perspective and its respective claim (Popat et al., 2019). The first input is similar to BERT sentence model's, *[CLS] claim [SEP] perspective*

[SEP], and the second one is the sequence of *[CLS] claim [SEP]*. Each input will be given to the BERT model separately. The goal is to incorporate the consistency between the representation of the perspective and claim using cosine distance of the two inputs. Accordingly, the following loss ($loss_c$) is added to the regular cross-entropy loss of the BERT model:

$$loss_c = \begin{cases} 1 - cos(X^{[C]}, X^{[C;P]}), \text{y=pro} \\ max(0, cos(X^{[C]}, X^{[C;P]})), \text{y=con} \end{cases}$$

where $X^{[C]}$ and $X^{[C;P]}$ are the final hidden state representations corresponding to the [CLS] token of the BERT model for the specified input. In our experiments, we replace the underlying question of a perspective with the claim in the two input sequences.

- XML-CNN model consists of three convolution layers with kernel size$= (2, 4, 8)$. With a dynamic max-pooling layer, crucial information is extracted across the document. XML-CNN was able to beat most of its deep neural network baselines in six benchmark datasets (Liu et al., 2017). We use, BERT, Word2vec, and FastText (Mikolov et al., 2018) embeddings for input tokens.

- AWD-LSTM is a weight-dropped LSTM that deploys DropConnect on hidden-to-hidden weights as a form of recurrent regularization (Merity et al., 2017). Word2vec Embedding is used for its input.

We define the corresponding hidden states of the last BERT layer as BERT embedding/representation of input sequence for both single and pair of inputs mode.

5.2 Training

We develop our code based on the Hedwig[4] implementation and train the models on 30 epochs with batch size$=8$. We apply early stopping technique to avoid overfitting during training. Training is stopped after 5 consequent epochs of no improvement of the highest F1 score. We inspect the test set on the model with the best F1 score of development set and keep the settings for BERT the same as the

---

[4]https://github.com/castorini/hedwig

Model	P.	R.	F1
Pair of Input			
BERT	**76.49**	75.37	75.92*†
BERT$_{CONS}$	70.34	81.24	75.40
XML-CNN(BERT)	68.48	82.22	74.72
VADER-Sent-$\overleftrightarrow{GRU}$	69.14	**86.62**	**76.90**†
NRC-Emotion-$\overrightarrow{GRU}$	73.79	79.45	76.51*
Unary Input			
BERT	73.89	76.18	75.02
AWD-LSTM(Word2Vec)	65.93	73.25	69.40
XML-CNN(Word2Vec)	58.30	83.03	68.51
XML-CNN(FastText)	66.85	77.32	71.71
XML-CNN(BERT)	70.30	79.93	74.81
VADER-Sent-$\overleftrightarrow{GRU}$	66.36	82.38	73.51
NRC-Emotion-$\overrightarrow{GRU}$	68.46	83.20	75.11

Table 3: Evaluation results; P.:Precision, R.:Recall, *: p-value ≤ 0.001; †: p-value ≤ 0.0001

BERT-base-uncased model. Adam optimizer with the learning rate of $2e - 5$ (for BERT) and $2e - 4$ (for other models) is used. We see a dramatic drop in BERT performance with some other learning rates. Scikit-learn (Pedregosa et al., 2011) library is employed for evaluation measures.

6 Results and Discussion

Experimental results are provided in Table 3. It was expected that fine-tuning BERT with a pair of input achieves a competitive performance among other baselines; but it shows that even with a shallow concatenation of the question and perspective (unary input), BERT can achieve consistent results. Moreover, models that take BERT representation in feature selection mode (without fine-tuning), e.g. XML-CNN(BERT), show better stance detection performance than other token embeddings.

We apply McNemar's test to measure whether the disagreement between the predictions of the two models is statistically significant.

Among the models with pairs of input, VADER-Sent-GRU gains the highest recall and F1 score. It indicates that the external knowledge gained from a massive corpus, fine-tuned on $20\times$ fewer parameters and enriched with sentiment information can compete with the original architecture (75.92 vs 76.90, $p < 0.0001$). As the model is significantly smaller, it trains faster and needs fewer resources for training. NRC-Emotion-GRU, highlighted in gray, achieves the second-highest F1 score among the models; It reveals that adding emotion information improves stance detection (75.92 vs 76.51,

36

Model	P.	R.	F1
VADER-Sent-$\overleftrightarrow{GRU}$	69.14	86.62	76.90
$\overleftrightarrow{GRU}$	72.18	78.30	75.12 (**1.78** ↓)
NRC-Emotion-$\overrightarrow{GRU}$	73.79	79.45	76.51
$\overrightarrow{GRU}$	69.14	83.69	75.72 (**0.79** ↓)

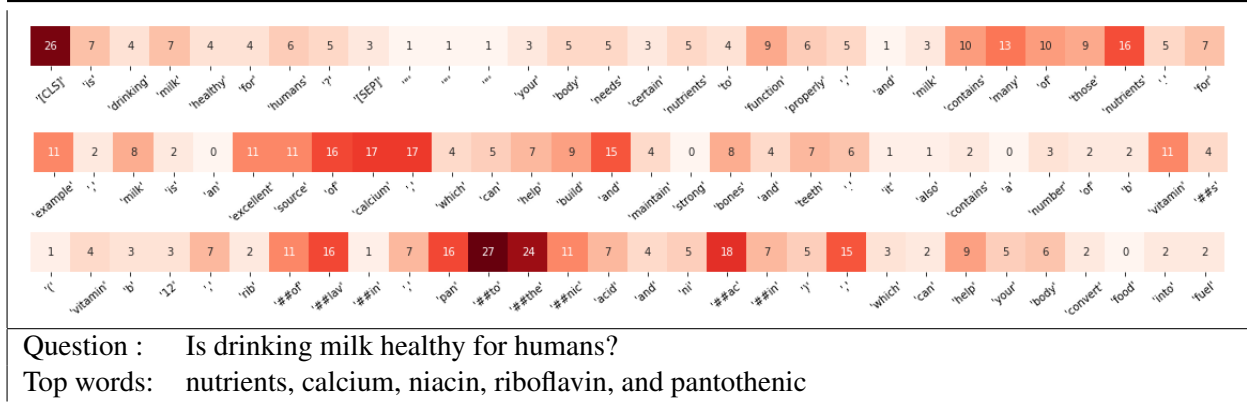Table 4: Effect of sentiment and emotion in our models with pair of input

Question :	Is drinking milk healthy for humans?
Top words:	nutrients, calcium, niacin, riboflavin, and pantothenic

Table 5: A pro-perspective

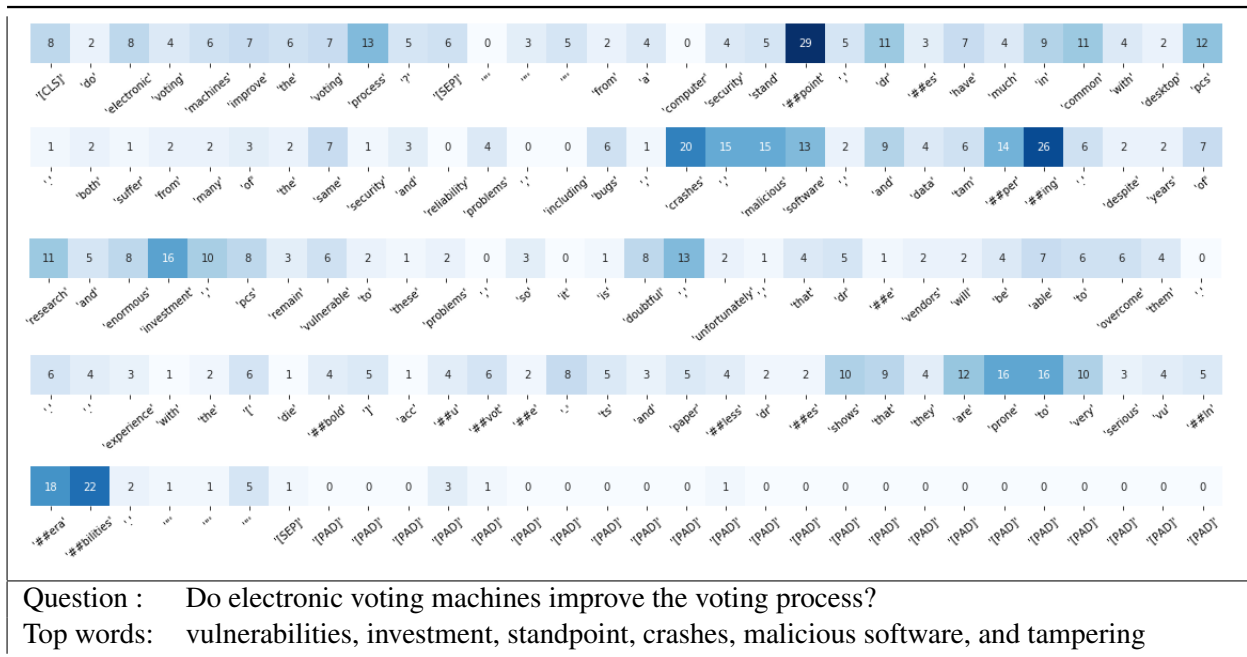| Question :	Do electronic voting machines improve the voting process?
Top words:	vulnerabilities, investment, standpoint, crashes, malicious software, and tampering

Table 6: A con-perspective

$p < 0.001$). However, employing sentiment information is more helpful than emotion in detecting the stance of opinions with compelling arguments (76.51 vs 76.90, $p < 0.0001$).

Unlike the superiority of BERT$_{CONS}$ over BERT reported in (Popat et al., 2019), we do not see a similar performance here. BERT$_{CONS}$ uses cosine similarity between the BERT representations of [claim] and [perspective; claim] in the loss function such that their representations become similar when per-

spective supports the claim and dissimilar when it opposes the claim. This method works for claims and perspectives of the Perspectrum dataset where the two input components are short sentences with $5 - 10$ words long. However, in our dataset, we have a *question* and its perspective that spans multiple sentences. So, forcing the model to make the BERT representations of [question] and [perspective; question] similar or dissimilar, according to the stance, harms the model training. Because the input components have different characteristics utilizing this method results in lower performance than the base model (BERT).

Next, we present some experiments to better understand the model's units.

6.1 Effect of Sentiment and Emotion

As stated in Section 4, our recurrent models (VADER-Sent-GRU and NRC-Emotion-GRU) employ sentiment and emotion information of tokens respectively. To see the effect of learning the flow of sentiment and emotion across an opinion, we lift their embeddings from the input of the models. So, $\overrightarrow{\text{GRU}}$ and $\overleftrightarrow{\text{GRU}}$ are unidirectional and bidirectional Gated Recurrent Units network respectively, followed by pooling and classification layers:

$$
\begin{aligned}
x_t &= h_t^{\text{BERT}}, \\
z_t &= \text{GRU}(x_t), \\
u &= [\text{avg-pool}(Z); \text{max-pool}(Z); z_T], \\
y &= \text{softmax}(Wu + b)
\end{aligned}
$$

Similarly, for an input sequence with T tokens, h_t^{BERT} is the hidden state of the last BERT layer corresponding to the input token at time t, $Z = [z_i]_{i=1}^T$, and W, b are parameters of a fully connected layer.

According to the results in Table 4, both precision and F1 score reduce for the model without emotion ($\overrightarrow{\text{GRU}}$); however, we see a reduction in recall and F1 in the model after lifting sentiment ($\overleftrightarrow{\text{GRU}}$) indicating that integrating sentence-wise sentiment and token-level emotion impact stance detection. We also provide the average sentiment score of the perspectives regarding five different questions in Figure 2. The figure shows the difference between the sentiment of the two stance classes in each issue resulting in a better stance classification. In the next part, we analyze the effect of pooling.

6.2 Pooling Explanation

In (Popat et al., 2019), authors find the most important phrases of input by removing phrases from the sequence and finding the ones with maximum effect on misclassification. In our model, we find the crucial information engaged in identifying the stance of a perspective using the max-pooling operation applied to the output sequence of recurrent neural networks (see Section 4). We hypothesize that the more a token is engaged in max-pooling, the more critical the token is for final stance prediction.

Tables 5 and 6 show the heatmap plots of two test instances. The number in each square is the engagement score, the frequency of the presence of a token in max-pooling operation. Darker colors show a higher frequency and indicate how the model identifies the stance across the perspective towards a question. The underlying question in Table 5 asks 'Is drinking milk healthy for humans?' According to its figure, we find sub-tokens of *nutrients*, *calcium*, *niacin*, *riboflavin*, and *pantothenic* with high scores. All of these words are positively aligned with the final (pro) stance; Specifically, the last three words are a type of Vitamin B. In another example in Table 6, the question is 'Do electronic voting machines improve the voting process?' Its corresponding heatmap displays sub-tokens of *vulnerabilities*, *investment*, *standpoint*, *crashes*, *malicious software*, and *tampering* with high scores; all of which are almost consistent with the perspective's (con) stance.

Similarly, we find the most important words/phrases, regarding their engagement score, for a few other examples of the test set that are correctly classified. The sub-tokens of these phrases have the highest frequency in max-pooling operation. We add (pro) or (con) at the end of each phrase list to indicate the stance of their respective perspective.

- Should students have to wear school uniforms? uniforms restrict students' freedom of expression (con)

- Are social networking sites good for our society? lead to stress and offline relationship (con)

- Should recreational marijuana be legal? legalization, odious occasion (con)
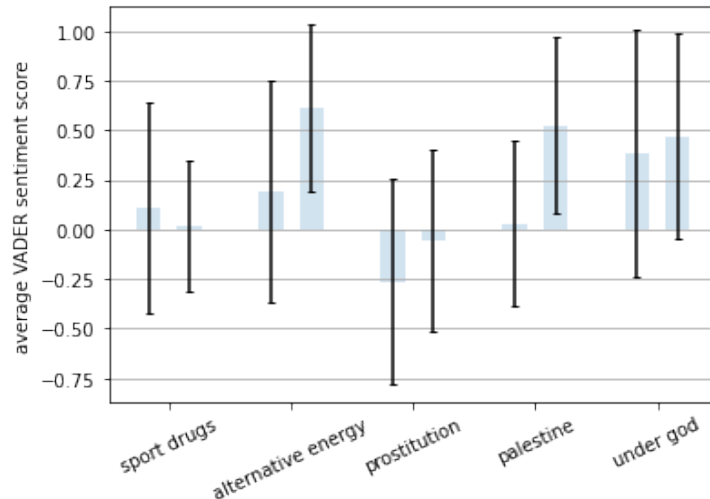
38

Figure 2: Average VADER sentiment scores across five different issues. In each issue the first bar belongs to proponents and the second bar belongs to the opponents

- What are the pros and cons of milk's effect on cancer? dairy consumption is linked with rising death rates from prostate cancer (con)

- Is human activity responsible for climate change? significant, because, (likely greater than 95 percent probability) (pro)

- Is obesity a disease? no question that obesity is a disease, blood sugar is not functioning properly, dysregulation, diabetes (pro)

- Is the death penalty immoral? anymore, failed policy (pro)

The above list shows that the stance-related phrases have been well identified by the model in the pooling step.

7 Conclusion

We propose a model that leverages BERT representation with sentiment or emotion information for stance detection. We create a new dataset for the perspectives that are as long as a paragraph covering a wide variety of contemporary topics. The experiments on our benchmark dataset highlight the effect of emotion and sentiment in stance prediction. The model can improve BERT base performance with significantly fewer parameters. We also explain the contribution of essential phrases of perspectives in detecting their stance using max-pooling operation.

References

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. In *Proceedings of the 2016 Conference on EMNLP*, pages 876–885, Austin, Texas. ACL.

Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. 2017. Stance classification of context-dependent claims. In *Proceedings of the 15th Conference of EACL: Volume 1, Long Papers*, pages 251–261, Valencia, Spain. ACL.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing things from a different angle:discovering diverse perspectives about claims. In *NAACL*, pages 542–557, Minneapolis, Minnesota. ACL.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT)*, pages 4171–4186, Minneapolis, Minnesota. ACL.

Eduard C. Dragut, Clement T. Yu, A. Prasad Sistla, and Weiyi Meng. 2010. Construction of a sentimental word dictionary. In *CIKM*, pages 1761–1764.

Jiachen Du, Ruifeng Xu, Yulan He, and Lin Gui. 2017. Stance classification with target-specific neural attention networks. IJCAI.

Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M. Meyer, and Iryna Gurevych. 2018. A retrospective analysis of the fake news challenge stance-detection task. In *ACL*, pages 1859–1874, Santa Fe, New Mexico, USA. ACL.

Kazi Saidul Hasan and Vincent Ng. 2013. Stance classification of ideological debates: Data, models, features, and constraints. In *Proceedings of the Sixth IJCNLP*, pages 1348–1356, Nagoya, Japan. AFNLP.

Marjan Hosseinia, Eduard Dragut, and Arjun Mukherjee. 2019. Pro/con: Neural detection of stance in argumentative opinions. In *SBP-BRiMS*, pages 21–30.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of ACL*, pages 328–339, Melbourne, Australia. ACL.

Clayton J. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *ICWSM*.

Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. 2017. Deep learning for extreme multi-label text classification. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 115–124. ACM.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2017. Regularizing and optimizing lstm language models. *CoRR*, abs/1708.02182.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of LREC 2018*.

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

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. 29(3):436–465.

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 NAACL: Human Language Technologies*, pages 767–776, New Orleans, Louisiana. ACL.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of NAACL: Human Language Technologies*, pages 2227–2237, New Orleans, Louisiana. ACL.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2019. STANCY: Stance classification based on consistency cues. In *Proceedings of the 2019 Conference on EMNLP-IJCNLP*, pages 6413–6418, Hong Kong, China. Association for Computational Linguistics.

Benjamin Riedel, Isabelle Augenstein, Georgios P. Spithourakis, and Sebastian Riedel. 2017. A simple but tough-to-beat baseline for the Fake News Challenge stance detection task. *CoRR*, abs/1707.03264.

Andrew T. Schneider and Eduard C. Dragut. 2015. Towards debugging sentiment lexicons. In *ACL*, pages 1024–1034.

Dhanya Sridhar, James Foulds, Bert Huang, Lise Getoor, and Marilyn Walker. 2015. Joint models of disagreement and stance in online debate. In *Proceedings of the 53rd ACL and the 7th IJCNLP*, pages 116–125, Beijing, China. ACL.

Marija Stanojevic, Jumanah Alshehri, Eduard C. Dragut, and Zoran Obradovic. 2019. Biased news data influence on classifying social media posts. In *NEwsIR@SIGIR*, volume 2411, pages 3–8.

Qingying Sun, Zhongqing Wang, Qiaoming Zhu, and Guodong Zhou. 2018. Stance detection with hierarchical attention network. In *ACL*, pages 2399–2409, Santa Fe, New Mexico, USA. ACL.

Marilyn Walker, Pranav Anand, Rob Abbott, and Ricky Grant. Stance classification using dialogic properties of persuasion.

Fan Yang, Arjun Mukherjee, and Eduard Dragut. 2017. Satirical news detection and analysis using attention mechanism and linguistic features. In *Proceedings of the 2017 Conference on EMNLP*, pages 1979–1989, Copenhagen, Denmark. ACL.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5754–5764.

Shanshan Zhang, Lihong He, Eduard Dragut, and Slobodan Vucetic. 2019. How to invest my time: Lessons from human-in-the-loop entity extraction. In *SIGKDD*, page 2305–2313.