ABSTRACT

Controversial claims are abundant in online media and discussion forums. A better understanding of such claims requires analyzing them from different perspectives. Stance classification is a necessary step for inferring these perspectives in terms of supporting or opposing the claim. In this work, we present a neural network model for stance classification leveraging BERT representations and augmenting them with a novel consistency constraint. Experiments on the Perspectrum dataset, consisting of claims and users’ perspectives from various debate websites, demonstrate the effectiveness of our approach over state-of-the-art baselines.

1 Introduction

There is an abundance of contentious claims on the Web including controversial statements from politicians, biased news reports, rumors, etc. People express their perspectives about these controversial claims through various channels like editorials, blog posts, social media, and discussion forums. To achieve a deeper understanding of these claims, we need to understand users’ perspectives and stance towards the claims. Recent research (FNC-1, 2016; Baly et al., 2018; Chen et al., 2019) has shown stance classification to be a critical step for information credibility and automated fact-checking.

Prior Work and Limitations: Prior approaches for stance classification proposed in Somasundaran and Wiebe (2010); Anand et al. (2011); Walker et al. (2012); Hasan and Ng (2013, 2014); Srithar et al. (2015); Sun et al. (2018) rely on various linguistic features, e.g., n-grams, dependency parse tree, opinion lexicons, and sentiment to determine the stance of perspectives regarding controversial topics. Ferreira and Vlachos (2016) further incorporate natural language claims and propose a logistic regression model using the lexical and semantic features of claims and perspectives. SemEval tasks (Mohammad et al., 2016; Kochkina et al., 2017) and other approaches (Chen and Ku, 2016; Lukasik et al., 2016; Sobhani et al., 2017) have focused on determining stance only in Tweets.

Bar-Haim et al. (2017) propose classifiers based on hand-crafted lexicons to identify important phrases in perspectives and their consistency with the claim to predict the stance. However, their model critically relies on manual lexicons and assumes that the important phrases in claims are already identified.

Neural-network-based approaches for stance classification learn the claim and perspective representations separately and later combine them with conditional LSTM encoding (Augenstein et al., 2016), attention mechanisms (Du et al., 2017) or memory networks (Mohtarami et al., 2018). Some neural network models also incorporate lexical features (Riedel et al., 2017; Hanselowski et al., 2018). None of these approaches leverage knowledge acquired from massive external corpora.

Approach and Contributions: To overcome the limitations of prior works, we present STANCY, a neural network model for stance classification. Given an input pair of a claim and a user’s perspective, our model predicts whether the perspective is supporting or opposing the claim. For example, the claim “You have nothing to worry about surveillance, if you have done nothing wrong” is supported by the user perspective “Information gathered through surveillance could be used to fight terrorism” and opposed by another user perspective “With surveillance, the user privacy will go away!”.

Our model for stance classification leverages representations from the BERT (Bidirectional
Encoder Representations from Transformers) neural network model (Devlin et al., 2019). BERT is trained on huge text corpora and serves as background knowledge. We fine-tune BERT for our task which also allows us to jointly model claims and perspectives. Furthermore, we enhance our model by augmenting it with a novel consistency constraint to capture agreement between the claim and perspective.

Key contributions of this paper are:

- **Model**: A neural network model for stance classification leveraging BERT representations learned over massive external corpora and a novel consistency constraint to jointly model claims and perspectives.
- **Interpretability**: A simple approach to interpret the contribution of perspective tokens in deciding their stance towards the claim.
- **Experiments**: Experiments on a recent dataset, Perspectrum, highlighting the effectiveness of our approach with error analysis.

2 BERT-based Approaches

In this section, first we describe the base model, BERTBASE, that is adapted for the stance classification (Chen et al., 2019). Thereafter, we present our consistency-aware model, BERTCONS.

2.1 Adapting BERT for Stance Classification

The goal of the stance classification task is to determine the stance of the user Perspective (P) with respect to the Claim (C). Since this task involves a pair of sentences (C and P), we follow the approach for sentence pair classification task as proposed in Devlin et al. (2019); Chen et al. (2019).

In order to obtain the representation $X^{P|C}$ of P with respect to C, this sentence pair is fused into a single input sequence by using a special classification token ([CLS]) and a separator token ([SEP]): $X^{P|C}$. The input sequences are tokenized using Word-Piece tokenization. The final hidden state representation corresponding to the [CLS] token is used as $X^{P|C}$. The classification probability is given by passing this representation through the softmax layer:

$$\hat{y} = \text{softmax}(X^{P|C}W^T)$$  \hspace{1cm} (1)

where softmax layer weights $W \in R^{H \times K}$ and $K$ is the number of stance (classification) labels. All the parameters of BERT and W are fine-tuned jointly by minimizing the cross-entropy loss ($\text{loss}_{ce}$). The architecture of this model, BERTBASE, is shown in Figure 1a.

2.2 Consistency-aware Stance Classification

In this setting, we want to incorporate the consistency between the claim (C) and perspective (P) representations. We hypothesize that the latent representations of claim and perspective should be **dissimilar** if the perspective opposes the claim, whereas their representations should be **similar** if the claim is supported by the perspective. We capture this with the following components.

- **Claim Representation**: To capture the latent representation of the claim, we use only the claim text as the input sequence to BERT, i.e., [CLS] $C$ [SEP]. The final hidden state of the first input token ([CLS]) is used as the claim’s representation $X^{C} \in R^{H}$.

- **Perspective Representation**: Latent representation of the perspective (with respect to the claim) is captured by fusing the two sequences as described in Section 2.1. We pack the claim and perspective pair as a single input sequence and use the final hidden state of the first input token as the perspective representation $X^{P|C} \in R^{H}$. 

![Figure 1: BERT-based methods for determining the stance of the perspective with respect to the claim.](image)
Table 1: Perspectrum data statistics.

<table>
<thead>
<tr>
<th>Split</th>
<th>Supporting Pairs</th>
<th>Opposing Pairs</th>
<th>Total Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>train</td>
<td>3603</td>
<td>3404</td>
<td>7007</td>
</tr>
<tr>
<td>dev</td>
<td>1051</td>
<td>1045</td>
<td>2096</td>
</tr>
<tr>
<td>test</td>
<td>1471</td>
<td>1302</td>
<td>2773</td>
</tr>
<tr>
<td>Total</td>
<td>6125</td>
<td>5751</td>
<td>11876</td>
</tr>
</tbody>
</table>

Capturing Consistency: To incorporate the consistency between claim and perspective representations, we use the cosine embedding loss:

\[
loss_{cos} = \begin{cases} 
1 - \cos(X^C, X^{P/C}) & y_{sim} = 1 \\
\max(0, \cos(X^C, X^{P/C})) & y_{sim} = -1 
\end{cases}
\]

where \(\cos(.)\) is the cosine similarity function. \(y_{sim}\) is equal to 1 if the perspective is supporting the claim (similar representations), and -1 if the claim is opposed by the perspective (dissimilar representations).

Joint Loss: The classification probabilities are determined by concatenating \(X^{P/C}\) and \(\cos(X^C, X^{P/C})\) and passing it through a softmax layer. However, unlike the BERT\(_{BASE}\) configuration, parameters of the consistency-aware model are learned by optimizing the joint loss function: \(loss = loss_{ce} + loss_{cos}\). With this joint loss function, we enforce consistency between latent representations of the claim and perspective. The architecture of this consistency-aware model, BERT\(_{CONS}\), is shown in Figure 1b.

3 Experimental Setup

For our experiments, we consider the base version of BERT\(^1\) with 12 layers, 768 hidden size, and 12 attention heads. We fine-tune BERT-based models using the Adam optimizer with learning rates \(\{1, 3, 5\} \times 10^{-5}\) and training batch sizes \(\{24, 28, 32\}\). We choose the best parameters based on the development split of the dataset. For measuring the performance, we use per-class and macro-averaged Precision/Recall/F1.

3.1 Dataset

We evaluate our approach on the Perspectrum dataset (Chen et al., 2019). Perspectrum contains claims and users’ perspectives from various online debate websites like idebate.com, debatewise.org, and procon.org. Each claim has different perspectives along with the stance (supporting or opposing the claim). We use the same train/dev/test split as provided in the released dataset. Statistics of the dataset is shown in Table 1.

3.2 Baselines

We use the following baselines:

- **LSTM**: A long short-term memory (LSTM) model, in which we pass the claim and perspective word representations (using GloVE-6B word embeddings of size 300) through a bidirectional LSTM. Then we concatenate the final hidden states of the claim and perspective, and pass it through dense layers with ReLU activations.

- **ESIM**: An enhanced sequential inference model (ESIM) for natural language inference proposed in Chen et al. (2017).

- **MLP**: Multi-layer perceptron (MLP) based model using lexical and similarity-based features – presented as a simple but tough-to-beat baseline for stance detection in Riedel et al. (2017).

- **WordAttn**: Our implementation of word-by-word attention-based model using long short-term memory networks (Rocktäschel et al., 2016).

- **LangFeat**: A random forest classifier using linguistic lexicons like NRC lexicon (Mohammad and Turney, 2010), hedges (e.g., possibly, might, etc.), positive/negative sentiment words (Hu and Liu, 2004), MPQA subjective lexicon (Wilson et al., 2005) and bias lexicon (Recasens et al., 2013) along with sentiment scores as features.

- **BERT\(_{BASE}\)**: Approach proposed in Chen et al. (2019) (as described in Section 2.1).

- **Human**: Human performance on this task as reported in Chen et al. (2019).

4 Results and Discussion

Stance classification performance of our model and the baselines on the test split of the Perspectrum dataset are presented in Table 2. Our consistency-aware model BERT\(_{CONS}\) outperforms all the other baselines. It achieves a performance improvement of about 2 points in F1-score over the strong baseline corresponding to the BERT\(_{BASE}\) model (p-value of 4.985e−4 as per the McNemar test). This highlights the value addition achieved by incorporating consistency cues. Since the BERT-based models incorporate the knowledge acquired from massive external corpora, our model, BERT\(_{CONS}\), captures better semantics and outperforms the other baselines.

\(^{1}\)BERT implementation: https://git.io/fhbJQ
### Approach Supporting Opposing Overall (Macro)

<table>
<thead>
<tr>
<th>Approach</th>
<th>Prec.</th>
<th>Recall</th>
<th>F1</th>
<th>Prec.</th>
<th>Recall</th>
<th>F1</th>
<th>Prec.</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM</td>
<td>63.42</td>
<td>58.80</td>
<td>61.02</td>
<td>56.99</td>
<td>61.67</td>
<td>59.24</td>
<td>60.20</td>
<td>60.24</td>
<td>60.13</td>
</tr>
<tr>
<td>ESIM</td>
<td>64.38</td>
<td>61.32</td>
<td>62.81</td>
<td>58.53</td>
<td>61.67</td>
<td>60.06</td>
<td>61.46</td>
<td>61.50</td>
<td>61.44</td>
</tr>
<tr>
<td>MLP</td>
<td>64.53</td>
<td>60.98</td>
<td>62.71</td>
<td>58.50</td>
<td>62.14</td>
<td>60.26</td>
<td>61.51</td>
<td>61.56</td>
<td>61.48</td>
</tr>
<tr>
<td>WordAttn</td>
<td>64.43</td>
<td>63.43</td>
<td>63.93</td>
<td>59.40</td>
<td>60.45</td>
<td>59.92</td>
<td>62.07</td>
<td>62.03</td>
<td>62.04</td>
</tr>
<tr>
<td>LangFeat</td>
<td>63.74</td>
<td>75.05</td>
<td>68.94</td>
<td>64.75</td>
<td>51.77</td>
<td>57.53</td>
<td>64.24</td>
<td>63.41</td>
<td>63.23</td>
</tr>
<tr>
<td>BERT\textsubscript{BASE}</td>
<td>78.43</td>
<td>80.08</td>
<td>79.25</td>
<td>76.95</td>
<td>75.12</td>
<td>76.02</td>
<td>77.60</td>
<td>77.60</td>
<td>77.62</td>
</tr>
<tr>
<td>BERT\textsubscript{CONS}</td>
<td>79.05</td>
<td>84.64</td>
<td>81.75</td>
<td>81.14</td>
<td>74.65</td>
<td>77.76</td>
<td>80.09</td>
<td>79.65</td>
<td>79.95</td>
</tr>
<tr>
<td>Human</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>91.3</td>
<td>90.6</td>
</tr>
</tbody>
</table>

Table 2: Comparison of our approach BERT\textsubscript{CONS} with different baseline models for stance classification.

### Table 3: Top phrases for determining stance.

<table>
<thead>
<tr>
<th>Opposing Class</th>
<th>Supporting Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>unauthorized, falsely, even</td>
<td>enabling, ensuring, prevail, positive discrimination, gains, help reduce, would improve, right, would allow, encourage, more effective, ...</td>
</tr>
</tbody>
</table>

Negations: One of the major challenges in solving this task is understanding negations and their scope. For example, given the claim “College education is worth it”, the perspective “Many college graduates are employed in jobs that do not require college degrees” is opposing the claim. However, our model is not able to capture that the negation phrase ‘do not require’ opposes the claim. On the other hand, the presence of negation in the perspective does not necessarily imply that it is opposing the claim. Contrast this with the claim “Chess must be at the Olympics” and perspective “Chess is currently not an Olympic sport, but it should be” – where the negation is merely a part of the statement and the stance is given by the discourse segment following ‘but’.

Commonsense: Determining the stance may require commonsense knowledge. For example, the claim “Chess must be at the Olympics” is opposed by the perspective “Olympic sports are supposed to be physical”. To understand this, the model should have the background knowledge that chess is not a physical sport.

Semantics: Understanding the stance also involves a deeper understanding of semantics. For example, given the claim “Make all museums free of charge” is opposed by the perspective “State funding should be used elsewhere”. Here, the word ‘elsewhere’ is the key cue which determines the stance. However, the presence of the word ‘elsewhere’ does not necessarily imply that the perspective is opposing the claim. For instance, the perspective “We could spend the money elsewhere” is supporting the claim “The EU should significantly reduce the amount it spends on agricultural production subsidies”. Hence, the polarity of the word ‘elsewhere’ is determined by the context and semantics of the statement.

#### 4.1 Interpreting Token-level Contribution

Due to the massive structure of BERT with a complex attention mechanism, it is difficult to interpret the significance of different lexical units in the text. Therefore, we propose a simple technique to interpret the contribution of each token in the text in determining the stance.

Given the claim \((C)\) and perspective \((P)\) pair, we tokenize \(P\) into phrases. We record the change in stance classification probabilities by adding one perspective phrase at a time to the input:

\[ \Delta_i = |BERT\textsubscript{CONS}(C, P_i) - BERT\textsubscript{CONS}(C, P_{i-1})| \]

where \(P_i\) is the prefix of \(P\) up to the \(i^{th}\) phrase. This helps us in understanding the contribution of each perspective phrase towards determining the stance – the larger the change in the classification probabilities, the larger the contribution. For this analysis, we consider unigrams and chunks from a shallow parser as phrases. The top contributing phrases for the supporting and opposing classes are shown in Table 3.

#### 4.2 Error Analysis

In this section, we analyze why the task of stance classification is challenging and why the performance of the best model configuration is far from human performance as observed by the performance gap in Table 2.
5 Conclusion

In this work, we propose a consistency-aware neural network model for stance classification. Our model leverages representations from the BERT model trained over massive external corpora and a novel consistency constraint to jointly model claims and perspectives. Our experiments on a recent benchmark highlight the advantages of our approach. We also study the gap in human performance and the performance of the best model for stance classification.

Acknowledgments

This research was partly supported by the ERC Synergy Grant “imPACT” (No. 610150).

References


Kazi Saidul Hasan and Vincent Ng. 2014. Why are you taking this stance? identifying and classifying reasons in ideological debates. In EMNLP.


Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kociský, and Phil Blunsom. 2016. Reasoning about entailment with neural attention. In ICLR.


