2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

A Pre-training Approach for Stance Classification in Online Forums

Jean Marie Tshimula,[1] Belkacem Chikhaoui,[1,2] Shengrui Wang[1]
[1]Prospectus Lab, Université de Sherbrooke, QC J1K 2R1, Canada
[2]LICEF Research Center, Université TÉLUQ, QC H2S 3L5, Canada
{kabj2801, shengrui.wang}@usherbrooke.ca, belkacem.chikhaoui@teluq.ca

***Abstract*—Stance detection is the task of automatically determining whether the author of a piece of text is in favor of, against, or neutral towards a target such as a topic, entity, or claim. In this paper, we propose a method based on RoBERTa to classify stances by capturing the context of the discussion through the examination of pairs of stances and relational structures of debates specific to each topic within the defined window of each forum participant's interventions. Furthermore, we examine the degree of disagreement and neutrality in various debate topics to measure divergence of opinion in the course of the debate and estimate the emotional state manifested in different debate topics. We conduct extensive experiments using two publicly available datasets and demonstrate that our method considers more stance classes, provides better results and yields statistical improvements over existing techniques. Our quantitative analysis of model performance yields F-1 scores of over 0.745. Interestingly, we obtained the highest F-1 score, 0.814, on a stance class which was not taken into consideration in prior work. We report that none of the metrics utilized to measure divergence of opinion yield values exceeding 50% and the correlations between the same topics over 10-fold cross-validation are statistically significant for the majority of them ($p < 0.005$). Several future research avenues are proposed.**

***Index Terms*—opinion, sentence-pair, divergence of opinion**

I. Introduction

Online forums are Internet-based group communities that provide an environment in which numerous topics can be discussed with other people who may be like-minded or hold opposing views. Since online discussion forums may assemble participants who have diversified convictions and beliefs on the various matters covered in the community, there can be a strong divergence of opinion in debates and some difficulty in reaching complete unanimity on a debate topic.

Online debates may contain discussions on several topics involving many participants, in which each intervention is either a response to a preceding post or the root of the discussion. The work reported here focuses on classifying the stances expressed by forum participants. Stance classification can be considered as the task of inferring from the text whether a particular forum participant agrees with an opinion expressed by another participant, disagrees with it, or has a neutral point of view towards it [7], [16], [19], [42]. Early work [2], [34] considered this issue as a binary classification task and focused on feature representations, and demonstrated that stance classification in the context of online forums is a very challenging problem.

Understanding stance can be of practical interest to many stakeholders, including companies and governments since it can provide critical insight into the theoretical foundation of discourse, argumentation, and sentiment [8], [35]. Such knowledge can be used for multiple purposes, such as predicting behavioral deterioration [41], detecting affinity relationships [39], [40], revealing misinformation [26], identifying fickle-minded people and weathervanes, recognizing logical fallacies like strawman arguments, targeting public awareness and advocacy campaigns [36], adapting users' information preferences to their beliefs and ideologies, conducting personality tests [15], [27], [30], [46] and online background checks, discerning the divergence of online discussion [12], [29], [44], and so on.

Prior studies proposed various techniques for detecting and classifying stance in a set of real-world texts. For instance, Sridhar *et al.* (2014) [36] introduced a collective classification method that captures the debate structure tree by modeling the dependencies between forum participants and their posts. Li *et al.* (2018) [19] used the structural dependencies of debate dialogues by measuring the similarity between embedding representations of a post and a given stance label. To determine the stance label, Sridhar *et al.*'s approach exploits manually written predicates and probabilistic soft logic to model reply links, and Li *et al.*'s approach relies exclusively on inference over the relationships between the learned representations of a post of interest; while other approaches merely detect the stance of participants from analysis of the text of a single post [5], [7], [8], [16], [19], [28], [37], [42], [47].

To overcome the limitations of the research discussed above, we propose a method that extracts the context of the discussion. The rationale behind context extraction is to capture relational structures of the discussion specific to each topic in order to classify proper pairs of posts. To classify two posts, we use RoBERTa [21], a Transformer-based pre-trained language model that carefully tunes hyper-parameters and trains data size, leading to significantly improved results on language understanding. It should be noted that RoBERTa is one of the top pre-trained language models and has yielded state-of-the-art results on many NLP downstream tasks and benchmarks (see SuperGLUE[1]), including sentence-

[1] https://super.gluebenchmark.com/leaderboard

IEEE/ACM ASONAM 2020, December 7-10, 2020
978-1-7281-1056-1/20/$31.00 © 2020 IEEE

pair classification. We use RoBERTa to generate features and then heuristically map entailment-type class labels onto an `Agreement-Disagreement-Neutral` relational structure to train secondary classifiers. Furthermore, we investigate the degree of disagreement and neutrality in different debate topics to measure divergence of opinion within the debate. Specifically, this work makes the following contributions:

- We propose a method that extracts the context of the discussion by exploring possible combinations of pairs of posts specific to each topic within the window between previous and next opinions expressed by each forum participant.
- We suggest a new transformation of the discussion to capture relational structures of the debate and simplify it to build our classifiers by means of RoBERTa-based sentence-pair labels, topics and new features extracted from initial annotations of the experimental datasets.
- We explore topics-based graphs to measure divergence of opinion throughout the discussion.
- We conduct extensive experimentation using real-world datasets to validate stance classification, measure divergence of opinion within the debate and assess emotional states manifested in different debate topics.

The rest of this paper is organized as follows. In Section II, we briefly outline some related work. Section III describes the proposed method. We conduct experiments and discuss the outcomes in Section IV. Finally, we conclude and present future directions in Section V.

II. Related Work

Prior work on stance classification has focused on linguistic features for identifying clues from oppositional speakers [2], [34], structural features for modeling agreement or disagreement with forum posts by inferring their labels [19], [35], [36], [43], sentiment [16], [20], [33], [38] and neural attention network approaches [5], [7], [37], [47].

To classify stances in tweets, Tutek *et al.* (2016) [42] designed both lexical and task-specific features to train and fine-tune several classifiers using a genetic algorithm. Ebrahimi *et al.* (2016) [8] proposed a probabilistic approach that discriminates sentiment- and target-specific features and then regularizes this on a single classifier. Krejzl and Steinberger (2016) [16] constructed a maximum entropy classifier based on surface-level, sentiment, and domain-specific features. The limitation of this work is that it classifies stances without comparing them to one another to capture the context in which these stances were expressed.

More recently, neural attention network techniques have been applied to classify stance more efficiently. These have achieved competitive results and helped stance classification make an important stride forward to investigate new avenues. For instance, Sun *et al.* (2018) [37] proposed a hierarchical attention network to weigh the importance of various linguistic information and learn the mutual attention between the document and the linguistic information. Zarrella and Marsh (2016) [47] suggested a transfer-learning-based method using large unlabeled datasets to learn sentence representations. Du *et al.* (2017) [7] introduced a neural attention model to extract target-specific related information for classifying stance in texts. These efforts were handicapped because they relied heavily on comparing post content with annotated labels, rather than classifying from the context of the stance of the hypothesis with respect to the premise.

We drew our inspiration from the work cited above, including [19] and [36] discussed in the previous section. These authors used probabilistic soft logic to model post stance by leveraging both the local linguistic features and the observed network structure of the posts [36], and introduced an approach for representing the structural dependencies of debate dialogues using graphical models and joint relational embeddings [19]. In contrast, our work captures relational structures to understand the context of the discussion and classify stances based on pairs of posts/sentences.

III. Model

To illustrate our model, we use some highly simplified notation. Let $F = \{f_1, f_2, \dots, f_N\}$ and $T{=}\{T_1, T_2, \dots, T_M\}$ respectively denote the sets of forum participants and topics, where each f_c represents a forum participant and each topic T_i consists of a set of sentences $\{s_{i,1}, s_{i,2}, \dots, s_{i,m_i}\}, \forall c \in \{1, \dots, N\}$ and $i \in \{1, \dots, M\}$. Each sentence s_j of topic T_i is mapped to its author f_c, and $\alpha_c^{T_i}$ represents the set of sentences belonging to T_i written by f_c. A forum participant f_c participates in a topic T_i if and only if $s_j \in \alpha_c^{T_i}$. The sentence s_j is the stance expressed by the forum participant f_c at time t and the sentence s_l is the stance expressed by the same person at time $t{+}n$ $(n > 0)$ in the course of the debate. The sentence s_k denotes the stance expressed by another forum participant before s_l is voiced $(j < k < l)$.

Respecting the timestamps of each sentence, we follow the flow of the discussion and learn the language inference between sentences to determine the stance class. To this end, we make combinations of sentence pairs between s_j and s_l. The rationale behind this is to identify the relationship between the meanings of a sentence pair by verifying whether s_k agrees with s_j, contradicts it or is simply neutral vis-a-vis the topic (T_i) that is being discussed. Specifically, we consider s_j as the premise and all possible s_k as the hypothesis for deciding the stance class. Let L denote a set of stance classes in the discussion that characterizes the position of the sentence s_k towards the sentence s_j, that is, the couple (s_j, s_k): the premise-hypothesis relationship. The class of the couple (s_j, s_k) corresponds to z, $\forall z \in L$ and $\forall s_j, s_k \in T_i$, where $L{=}\{\texttt{neutral}, \texttt{agreement}, \texttt{disagreement}\}$ and $k \in]j, l[$. A sentence without a predecessor could be considered as the root of the discussion and might not directly depend on any premise. We, therefore, lack couples for such cases.

To perform the sentence-pair classification task, we utilize RoBERTa [21], which is already fine-tuned on the MultiNLI (Multi-genre Natural Language Inference) corpus. The MultiNLI is a crowdsourced collection of 433K sentence pairs annotated with textual entailment information [45]. To

TABLE I: An example to demonstrate the functioning of our approach. We suppose a discussion forum of 10 sentences that involves three forum participants, F=$\{f_1, f_2, f_3\}$, who debate on a given topic. The sentence s_1 is the root of the discussion; it does not depend on a premise. It should be noted that the class labels of the couples are extracted separately using RoBERTa.

(a) Sentences by f_c		(b) Stance combinations
F	Sentence	Sentence pairs (premise, hypothesis)
f_1	s_1	–
f_2	s_2	(s_1, s_2)
f_3	s_3	$(s_1, s_3), (s_2, s_3)$
f_2	s_4	$(s_1, s_4), (s_2, s_4), (s_3, s_4)$
f_1	s_5	$(s_1, s_5), (s_4, s_5), (s_3, s_5)$
f_3	s_6	$(s_5, s_6), (s_4, s_6), (s_3, s_6)$
f_1	s_7	$(s_5, s_7), (s_4, s_7), (s_6, s_7)$
f_3	s_8	$(s_7, s_8), (s_4, s_8), (s_6, s_8)$
f_2	s_9	$(s_7, s_9), (s_4, s_9)$
f_1	s_{10}	(s_7, s_{10})

TABLE II: An example to illustrate feature extraction. Suppose that s_1 and s_2 are respectively annotated beforehand as Agreement and Disagreement in Table I(a) and RoBERTa yields a Neutral for the couple (s_1, s_2). Clearly, s_1 represents the premise and s_2 the hypothesis. Therefore, we take Agreement as the premise label and Disagreement as the hypothesis label.

Sentence pair	Premise	Hypothesis
(s_1, s_2)	s_1	s_2
(s_1, s_3)	s_1	s_3
(s_2, s_3)	s_2	s_3
(s_1, s_4)	s_1	s_4
(s_2, s_4)	s_2	s_4
(s_3, s_4)	s_3	s_4
(s_1, s_5)	s_1	s_5
(s_4, s_5)	s_4	s_5
(s_3, s_5)	s_3	s_5
(s_5, s_6)	s_5	s_6
(s_4, s_6)	s_4	s_6
(s_3, s_6)	s_3	s_6
(s_5, s_7)	s_5	s_7
(s_4, s_7)	s_4	s_7
(s_6, s_7)	s_6	s_7
(s_7, s_8)	s_7	s_8
(s_4, s_8)	s_4	s_8
(s_6, s_8)	s_6	s_8
(s_7, s_9)	s_7	s_9
(s_4, s_9)	s_4	s_9
(s_7, s_{10})	s_7	s_{10}

classify two sentences, RoBERTa generates fixed-size sentence embeddings, where the feature representations of sentences are obtained from the trained encoders, and then passes them to a softmax classifier to derive the final label: i.e., contradiction, Neutral or entailment. We obtain (i) entailment when the hypothesis has a similar meaning to the premise, (ii) contradiction when the hypothesis has a contradictory meaning, and (iii) neutral when the hypothesis has mostly the same lexical items as the premise but bears a different meaning. In this paper, we have chosen to use Agreement and Disagreement to simplify the terms entailment and contradiction, respectively.

Table I illustrates the functioning of our approach. We assume that each sentence in Table I(a) solely addresses a single topic (T_1) which is argued by three forum participants. Table I(b) shows how we generate possible combinations of sentence pairs based on the logic described above. To graphically represent the flow of the discussion, we take the relational structure depicted by Table I(b) to construct topic-based graphs. More formally, $G = (V, A, T_i)$ denotes a directed multigraph for the topic T_i, where V is the set of vertices corresponding to forum participants and A is the set of arcs indicating stance labels. Recall that stance labels are results of sentence-pair classification yielded by RoBERTa, and a directed multigraph may have several arcs with the same origin and destination vertices. We explore G to measure the divergence of opinion throughout the discussion.

To classify stances, we design additional features from the dataset annotations, namely the premise and hypothesis labels. We assume that sentences in the dataset are annotated beforehand. For each couple (s_j, s_k), we derive the sentence-pair label z using RoBERTa ($z \in L$). Additionally, we collect the true labels of these sentences as annotated, and then follow the position of each sentence in the couple to properly assign premise and hypothesis labels to them (see Table II).

IV. Experiments

To empirically evaluate our method, we conducted extensive experiments with two publicly available online forum datasets: Annotated Coarse Discourse and Internet Argument Corpus v2. We will now describe these datasets, introduce the techniques used as a baseline for comparison, and present the evaluation metric and details of the training process for our method.

A. Data Description

The Internet Argument Corpus v2 (IAC2) dataset is a collection of corpora of political debate topics on online forums [1]. Initially, it should be noted that IAC2 is composed of three different datasets: 4forums, which comprises over 3.5K participants and 414K posts (with an average of 340 users per topic and 19 posts per user); ConvinceMe (65K posts) and CreateDebate (3K posts). Of these, we opted to use 4forums because of its size and the number of users it contains, as well as for its topic annotations and response characterization. Notably, 4forums has crowdsourced annotations with a high inter-annotator agreement for stances of users in each topic and disagreement between users who reply to one another, and it spans many topics. In our experiments, we limited ourselves to five topics, as these are most prominent in the dataset: `Evolution`, `Gay Marriage`, `Abortion`, `Gun Control`, and `Death Penalty`. On 4forums, `agreement/disagreement` scores are given on an 11-point scale [-5,5]. Scores ≤ 0 indicate higher inter-annotator confidence for disagreement, whereas scores ≥ 1 denote agreement. Sridhar *et al.* (2014) [36] removed all posts for which annotations belong to the interval [0,1] due to uncertainty about the agreement. In contrast, we opted to keep these posts, since RoBERTa is fine-tuned to detect cases where the stance of the hypothesis sentence is neutral.

The Annotated Coarse Discourse (ACD) dataset is a large corpus of discourse annotations and relations collected from Reddit by Zhang *et al.* (2017) [48]. Its goal is to allow a better understanding of online discussions at scale. It contains over 61K participants and 9,000 threads comprising over 101,000 comments, manually annotated. Basically, the discourse-act annotation scheme was developed to highlight comments that include agreement, appreciation, disagreement and negative reactions. In contrast to IAC2, we assume that ACD solely covers one topic, given that it does not include topic annotations.

B. Model Evaluation

To validate the performance of the proposed method, we compared it with the following baseline methods: PSL [36], UWB [16], MITRE [47], SRL [19] and BERT [6]. It should be recalled that we plainly discussed some limitations of the first four methods in Section II.

- PSL [36] uses probabilistic soft logic to capture relational information in the network of authors and posts. The intuition of PSL is that the class `Agreement` or `Disagreement` between users correlates to their stance towards a topic.
- UWB [16] is based on a maximum entropy classifier with mainly surface-level, sentiment and domain-specific features.
- MITRE [47] maximizes the value of limited training data by transferring features from other systems trained on large, unlabeled datasets.
- SRL [19] uses the structural dependencies of the discussion and measures the similarity between embedding representations of the post and a given stance label.
- BERT [6] is a bidirectional Transformer-based pre-trained contextual representation trained using masked language modeling objective and next sentence prediction tasks. It exploits a multi-layer bidirectional Transformer encoder, where each layer contains multiple attention heads. More specifically, we utilize a BERT-large model fine-tuned on the MultiNLI [6], [13].

Feature sets. To build our stance classifiers, we used four different features. The feature `sentence-pair label` is our response variable and refers to the stance label yielded by RoBERTa. This portrays the stance label of a sentence pair, i.e., the stance of a given hypothesis sentence towards a premise sentence. The features `premise label` and `hypothesis label` stem from human-annotated labels in the experimental datasets: we consider the manually annotated stance label for each sentence as ground truth. Finally, the feature `topic` denotes the topics discussed by the forum participants.

Model performance. To evaluate model performance, we conducted stratified ten-fold cross-validation to split our training and testing sets. We trained three distinct classifiers: logistic regression (LR), support vector machine (SVM) and random forest (RF). For SVM, we set the value γ of the radial basis function kernel to 0.5 and for RF, we built a model with 100 trees. We replicated the same logic for BERT to generate features, then trained an SVM classifier, BERT+SVM. We computed the F-1 score (harmonic mean of precision and recall) to measure the accuracy of our classifiers and to quantitatively compare them with the baseline techniques.

Divergence metrics. To quantify divergence of opinion within topic debates, we used four measures of probability divergence: Kullback–Leibler (KL), Jensen-Shannon (JS), Hellinger distance (HD) and Bhattacharyya distance (BD). These divergence metrics measure the discrepancy/similarity between two probability distributions [4], [14], [17], [31], [32].

Let us look at two discrete probability distributions $P = \{p_i\}_{i\in[n]}$ and $Q = \{q_i\}_{i\in[n]}$ supported on $[n]$. KL is a directed divergence that measures the discrepancy between the two, with the meaning being dependent on which direction was computed (see eq. (1)). Equation (1) determines how the Q distribution is different from the P distribution. KL is a non-negative, asymmetric distance (i.e., $\mathrm{KL}(P\|Q) \neq \mathrm{KL}(Q\|P)$); it is zero if the two distributions are identical and can potentially equal infinity [32]. JS is a symmetrized, smoothed version of KL which measures the total KL divergence from the average mixture distribution, $M = \frac{(P+Q)}{2}$ (see eq. (2)). Some salient features of JS are that it is always defined, bounded and symmetric, and only vanishes when $P = Q$ [4]. HD is a probabilistic analog of the Euclidean distance and satisfies the triangle inequality. The

$\sqrt{2}$ in equation (3) is to ensure that HD$(P,Q){\leq}1$ for all probability distributions. One advantage of HD is that it serves to provide the lower bounds for Bayes risk in non-regular situations [31]. BD is defined as the negative logarithm of the Bhattacharyya coefficient [14]. Clearly, BD does not satisfy the triangle inequality, $0{\leq}\text{BD}(P,Q){\leq}{+\infty}$ (see eq. (4)). The Bhattacharyya measure has a simple geometric interpretation as the cosine of the angle between two position vectors in n-dimensional space $(\sqrt{p_1},\ldots,\sqrt{p_n})^\top$and $(\sqrt{q_1},\ldots,\sqrt{q_n})^\top$, where $\cos(\theta)=\sum_{i\in[n]}\sqrt{p_i\times q_i}$. Consequently, if the P and Q distributions are identical, $\cos(\theta){=}1$, corresponding to $\theta{=}0$.

$$\text{KL}(P\|Q)=\sum_{i\in[n]}p_i\times\log{(\frac{p_i}{q_i})} \tag{1}$$

$$\text{JS}(P\|Q)=\frac{1}{2}\text{KL}(P\|M)+\frac{1}{2}\text{KL}(Q\|M) \tag{2}$$

$$\text{HD}(P,Q)=\frac{1}{\sqrt{2}}\sqrt{\sum_{i\in[n]}(\sqrt{p_i}-\sqrt{q_i})^2} \tag{3}$$

$$\text{BD}(P,Q)=-\ln(\sum_{i\in[n]}\sqrt{p_i\times q_i}) \tag{4}$$

We assume that one is more likely to encounter divergences of opinion in sentence pairs for which the class label is `Disagreement` or `Neutral`. We, therefore, take each of these sentence pairs, apply the word2vec skip-gram model to embed the words of each sentence of the pair as vectors in a low-dimensional space [23], and finally encode them as probability densities [3]. Note that the densities represent the distributions over the possible significations of a word. We calculate the divergence metrics between the distributions of sentence pairs. To estimate divergence of opinion on the whole topic, we calculate the arithmetic mean of the results obtained from all sentence pairs of interest.

C. Results and Discussion

Stance classification. Table III presents the performance results for three-class classification in different experimental settings. It should be noted that the baseline techniques only classify two classes (`Agreement` and `Disagreement`), except for BERT+SVM which classifies the three classes (`Agreement`, `Disagreement` and `Neutral`). Our classifiers and BERT+SVM are able to accurately distinguish between the three classes and achieve better performance than the other baseline techniques.

We observe that the F-1 scores are significantly higher than 0.5 for stance classification, although PSL achieved poor performance on ACD. However, this poor performance can be partly explained by PSL's inability to capture the context of the discussion. Our method yields statistically significant improvements over PSL, surpassing it by 0.292 for `Agreement` and 0.289 for `Disagreement` on ACD. We empirically show that our method outperforms the baselines by a considerable margin for the 2-classes and yields good classification performance on the label `Neutral`. Further investigation found that the arithmetic mean of our method's F-1 scores on the 2-classes also surpassed the baselines (with 0.796 for `Agreement` and 0.788 for `Disagreement` on IAC2, and 0.776 for `Agreement` and 0.771 for `Disagreement` on ACD). We note that UWB and SRL achieved roughly similar performances on the two experimental datasets.

Compared with MITRE, BERT+SVM attained better performance on both datasets, but still lagged behind our classifiers. Specifically, the average of our best performances yields F-1 scores higher than those for the strongest baseline, BERT+SVM, with improvements of about 0.039 and 0.027 over IAC2 and ACD, respectively. BERT+SVM achieved performances closer to those of our smallest classifier for the class `Neutral`: 0.004 on IAC2 and 0.003 on ACD. It should be noted that, aside from BERT+SVM, MITRE is the baseline that comes closest to our best F-1 scores. We observe that our method improves upon MITRE by (0.047, 0.059) and (0.047, 0.041) for `Agreement` and `Disagreement` on (IAC2, ACD), (p-value of $p<0.0005$ as per the McNemar test [22] on IAC2 and $p<0.0003$ on ACD). MITRE performs similarly on the 2-classes on both datasets, and once outperformed our smallest classifier and BERT+SVM on the class `Disagreement`, i.e., (BERT+SVM $<$ Ours+RF $<$ MITRE $<$ Ours+LR $<$ Ours+SVM). It should be noted that MITRE is a transfer-learning method trained on large unlabeled datasets to generate features using word embeddings, and then learn sentence representations from these features to classify stances. MITRE retains the knowledge acquired in solving one case and subsequently applies it to a different but related case. This explains its good performance.

We have shown clear benefits and strong evidence that capturing the context and relational structures of debates can provide better performance on the task of stance classification. We achieved our best F-1 scores on IAC2 with LR and ACD with RF. The highest F-1 score, 0.814, was achieved for the class `Neutral` on IAC2, and the smallest F-1 score obtained by the proposed method is much greater than the F-1 scores of all baseline techniques except for MITRE; that is, (MITRE $>$ Ours+RF $>$ BERT+SVM $>$ UWB $>$ SRL $>$ PSL).

Divergence within debate topics. Table V shows the results of the divergence metrics we utilized to measure the divergence of opinion on each debate topic addressed in the forum. The metrics used yielded approximately similar results, even though JS achieved the best performance on the majority of topics tackled in the experimental datasets. We note that the values for divergence of opinion yielded by the experimental metrics do not exceed 0.50. The significance of the overall divergence value may be somewhat difficult to interpret, whereas divergence of opinion between two different viewpoints can be understood and explained. However, analyzing the motives and sentiments behind divergences of opinion which fueled heated discussions normally requires further inquiry.

We observe that `Gun Control` is a topic that sparked a relatively large divergence of opinion between proponents and opponents of the right to keep and bear arms, and yielded

TABLE III: Stance classification results for the proposed method and baselines on the two experimental datasets. The F-1 score metric is used to gauge model performance. Bold font indicates the best results for each class label. PSL, UWB, MITRE and SRL merely classified two classes, i.e., `Agreement` and `Disagreement`.

	Label	Ours+LR	Ours+SVM	Ours+RF	BERT+SVM	PSL	UWB	MITRE	SRL
IAC2	`Agreement`	**0.803**	0.796	0.788	0.781	0.572	0.687	0.756	0.671
	`Disagreement`	0.807	**0.812**	0.746	0.737	0.56	0.675	0.753	0.664
	`Neutral`	0.799	0.805	**0.814**	0.795	–	–	–	–
ACD	`Agreement`	**0.787**	0.772	0.768	0.762	0.495	0.648	0.74	0.626
	`Disagreement`	0.751	0.778	**0.785**	0.746	0.496	0.654	0.744	0.631
	`Neutral`	0.776	0.767	**0.780**	0.764	–	–	–	–

TABLE IV: Prediction performance (Pearson's r) based on 10-fold cross-validation using LIWC features (positive and negative emotions) extracted from different topics addressed on IAC2 and ACD datasets. All features are significant at $p < 0.005$, except for the negative emotion on `Gay Marriage` and the positive emotion on `Coarse`, for which p is not statistically significant.

LIWC	Evolution	Gay Marriage	Abortion	Gun Control	Death Penalty	Coarse
`Positive emotion`	0.301	0.225	0.26	0.425	0.218	–0.151
`Negative emotion`	0.287	0.163	0.271	0.409	0.43	0.467

TABLE V: Quantifying divergence of opinion by topic.

	Topic	JS	KL	HD	BD
IAC2	`Evolution`	0.114	0.089	0.103	0.077
	`Gay Marriage`	0.206	0.193	0.198	0.185
	`Abortion`	0.337	0.323	0.305	0.294
	`Gun Control`	0.442	0.437	0.421	0.405
	`Death Penalty`	0.253	0.268	0.244	0.231
ACD	`Coarse`	0.414	0.406	0.409	0.382

a higher divergence value than other topics on IAC2 (with 0.442 over JS and 0.426 as the arithmetic mean of the divergence values of all metrics). We notice that KL and HD performed similarly on `Gay Marriage` on IAC2. We found that `Evolution` is the topic with the smallest divergence value, followed by `Gay Marriage`. Moreover, the topic `Abortion` generated a greater divergence of opinion than `Evolution` and `Gay Marriage` combined ($0.337 > 0.32$ with JS, $0.323 > 0.282$ with KL, $0.305 > 0.301$ with HD, $0.294 > 0.262$ with BD).

We note that KL achieved better performance than JS on `Death Penalty` ($0.268 > 0.253$). (Proponents of this topic argue that capital punishment is beneficial even if it has no deterrent effect, while opponents alternatively suggest life imprisonment.) We obtained an arithmetic mean of 0.196 on the whole discussion on `Gay Marriage`. Somehow, this value is greatly inferior to that for cases where there is no divergence (a difference of 1 – 0.196, i.e., 0.804), suggesting that this topic may have triggered relatively few emotional debates.[2] Finally, we observe that JS and BD achieved the highest and lowest divergence values, respectively, on ACD.

Psychological processes. To measure emotional state manifested [9] by forum participants who addressed the topics that we studied above, we utilize Linguistic Inquiry and Word Count (LIWC) [25], a dictionary that is widely employed in computational linguistics as a source of features for psychological and psycholinguistic analysis. LIWC comprises words that have very clear, pre-labeled meanings. The dictionary includes words in various categories, notably linguistic dimensions, psychological processes and personal concerns. Each category is found to be correlated with several psychological traits and outcomes [10], [11]. Within the psychological processes category, we find the emotion sub-dictionaries, that is, positive and negative emotions. We focus on the psychological processes category in order to explore the linguistic usage in user viewpoints. It should be noted that the positive and negative emotions are not two ends of a scale, since a point of view can include the two. We leverage each opinion (each sentence, see Table I) and measure the proportion of word tokens that fall into negative and positive emotions.

To predict the emotion associated with opinions by topic, we treat each topic separately and stratify each topic's data by 10-fold cross-validation to split our training and testing sets. We utilize linear regression with three different regularization methods: LASSO (Least Absolute Shrinkage and Selection Operator), ridge and elastic net. The elastic net yielded

[2] Based on the results yielded in Table IV, we find that the correlation of negative emotion-related words over `Gay Marriage` is not statistically significant. This could be considered as strong evidence to argue that this topic may have sparked few emotional statements.

marginally higher performance over the two other techniques. The performance was measured using the Pearson correlation (*r*) [24]. Table IV indicates the correlations between words containing positive and negative emotions over different topics addressed in the two experimental datasets. It can be seen that `Coarse` yielded the strongest correlation with negative emotion, 0.467 ($p < 0.001$) and `Gun Control` produced the highest correlation with positive emotion, 0.425 ($p < 0.001$). More importantly, we find that both positive and negative emotion features for all topics are significant at $p < 0.005$, apart from the *p* for negative emotions on `Gay Marriage` (p = 0.756) and positive emotions on `Coarse` (p = 0.725) and `Death Penalty` (p = 0.618), respectively, which are not statistically significant. We observe that positive emotions appear to have a stronger effect on `Gay Marriage` (r = 0.225, $p < 0.001$) and `Evolution` (r = 0.301, $p < 0.001$).

V. Conclusion

We present a RoBERTa-based method to classify stances by capturing the context and relational structures of the debate. Our method shows statistically significant improvements over existing methods in terms of F-1 score performance on this task and provides good results for the class `Neutral`, which was not considered in prior work. We report that `Neutral` yields performance surpassing 0.75 on the two experimental datasets. Furthermore, we examine the degree of disagreement and neutrality to measure the divergence of opinion on topics addressed in the debate. We note that none of the metrics utilized yields values surpassing 0.5. We limit ourselves to reporting the observed divergence values rather than explaining the motives and sentiments that fueled the debate so that we have divergence of opinion among individuals; this aspect normally requires further analysis. We measure the emotional state manifested in topics addressed in different debates. To this end, we resort to the LIWC dictionary, especially the psychological processes category, to calculate the proportion of opinion-related words that fall into positive and negative emotions. We find that the majority of features extracted from all topics addressed are statistically significant. Additionally, we indicate the topics addressed that include the highest and lowest correlations of positive and negative emotions.

This study provides a framework for further research about stance classification in different settings in online discussion forums. Specifically, we aim to exploit pre-trained language models to classify stances based on hypotheses related to multiple independent premise sentences [18] and thereafter detect some logical fallacies in debates, including strawman, red herring, *tu quoque*, hasty generalization and slippery slope arguments. Furthermore, we would like to study the effects of emotional reactions on divergent opinions, investigate at the user- and debate-levels in order to discern the motives behind divergent opinions, and predict whether the intensity of emotional reaction in divergent opinions is likely to grow as the debate moves forward.

References

[1] R. Abbott, B. Ecker, P. Anand, and M.A. Walker, “Internet Argument Corpus 2.0: An SQL schema for Dialogic Social Media and the Corpora to go with it,” In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16), pp. 4445–4452, 2016.

[2] P. Anand, M. Walker, R. Abbott, J.E.F. Tree, R. Bowmani, and M. Minor, “Cats Rule and Dogs Drool!: Classifying Stance in Online Debate,” In Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, pp. 1–9, 2011.

[3] A. Brazinskas, S. Havrylov, and I. Titov, “Embedding Words as Distributions with a Bayesian Skip-gram Model,” In Proceedings of the 27th International Conference on Computational Linguistics, pp. 1775–1789, 2018.

[4] J. Briët, and P. Harremoës, “Properties of classical and quantum Jensen-Shannon divergence,” Phys. Rev. A 79, 052311, 2009.

[5] W.F Chen, and L.W. Ku, “UTCNN: A Deep Learning Model of Stance Classification on Social Media Text,” In Proceedings of the 26th International Conference on Computational Linguistics, pp. 1635–1645, 2016.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “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, pp. 4171–4186, 2019.

[7] J. Du, R. Xu, Y. He, and L. Gui, “Stance Classification with Target-Specific Neural Attention Networks,” In Proceedings of the 26th International Joint Conference on Artificial Intelligence, pp. 3988–3994, 2017.

[8] J. Ebrahimi, D. Dou, and D. Lowd, “A Joint Sentiment-Target-Stance Model for Stance Classification in Tweets,” In Proceedings of the 26th International Conference on Computational Linguistics, pp. 2656–2665, 2016.

[9] A. Giachanou, P. Rosso, I. Mele, and F. Crestani, “Emotional Influence Prediction of News Posts,” In Proceedings of the Twelfth International AAAI Conference on Web and Social Media, pp. 592–595, 2018.

[10] S.C. Guntuku, R. Schneider, A. Pelullo, J. Young, V. Wong, L. Ungar, D. Polsky, K.G. Volpp, and R. Merchant, “Studying Expressions of Loneliness in Individuals using Twitter: An Observational Study,” BMJ Open, 9:e030355, 2019.

[11] S.C. Guntuku, D.B. Yaden, M.L. Kern, L.H. Ungar, and J.C. Eichstaedt, “Detecting Depression and Mental Illness on Social Media: An Integrative Review,”, Current Opinion in Behavioral Sciences, 18:43-49, 2017.

[12] D. Jian-hua, W. Ya-li and D. Yi, “A New Method of Defining Divergent Opinions in Group Decision-Making,” In 2013 International Conference on Management Science and Engineering 20th Annual Conference Proceedings, pp. 447–451, 2013.

[13] N. Jiang and M.-C. de Marneffe, “Evaluating BERT for Natural Language Inference: A Case Study on the CommitmentBank,” In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 6086–6091, 2019.

[14] T. Kailath, “The Divergence and Bhattacharyya Distance Measures in Signal Selection,” IEEE Transactions on Communication Technology, 15(1):52–60, 1967.

[15] V. Kaushal, and M. Patwardhan, “Emerging Trends in Personality Identification Using Online Social Networks—A Literature Survey,” ACM Transactions on Knowledge Discovery from Data, 12(2):15, 2018.

[16] P. Krejzl, and J. Steinberger, “UWB at SemEval-2016 Task 6: Stance Detection,” In Proceedings of the 10th International Workshop on Semantic Evaluation, pp. 408–412, 2016.

[17] K. Krstovski, D.A. Smith, H.M. Wallach, and A. McGregor, “Efficient Nearest-Neighbor Search in the Probability Simplex,” In Proceedings of the 2013 Conference on the Theory of Information Retrieval, pp. 101–108, 2013.

[18] A. Lai, Y. Bisk, and J. Hockenmaier, “Natural Language Inference from Multiple Premises,” In Proceedings of the 8th International Joint Conference on Natural Language Processing,” pp. 100–109, 2017

[19] C. Li, A. Porco, and D. Goldwasser, “Structured Representation Learning for Online Debate Stance Prediction,” In Proceedings of the 27th International Conference on Computational Linguistics, pp. 3728–3739, 2018.

[20] Y. Li, and C. Caragea, "Multi-Task Stance Detection with Sentiment and Stance Lexicons," In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 6299–6305, 2019

[21] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," In arXiv preprint arXiv:1907.11692, 2019.

[22] Q. McNemar, "Note on the Sampling Error of the Difference between Correlated Proportions or Percentages," Psychometrika, 12(2):153–157, 1947.

[23] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," In Proceedings of the 26th International Conference on Neural Information Processing Systems, pp. 3111–3119, 2013.

[24] K. Pearson, "Notes on Regression and Inheritance in the Case of Two Parents," In Proceedings of the Royal Society of London, 58:240–242, 1895.

[25] J.W. Pennebaker, R.L. Boyd, K. Jordan, and K. Blackburn, "The Development and Psychometric Properties of LIWC2015," Austin, TX: University of Texas at Austin, 2015

[26] G. Pennycook, and D.G. Rand, "Fighting Misinformation on Social Media Using Crowdsourced Judgments of News Source Quality," In Proceedings of the National Academy of Sciences, 116(7):2521–2526, 2019.

[27] F. Piedboeuf, P. Langlais, and L. Bourg, "Personality Extraction Through LinkedIn," In: Meurs MJ., Rudzicz F. (eds) Advances in Artificial Intelligence. Canadian AI 2019. Lecture Notes in Computer Science, vol 11489. Springer, Cham. 2019

[28] P. Potash, and A. Rumshisky, "Towards Debate Automation: A Recurrent Model for Predicting DebateWinners," In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2465–2475, 2017.

[29] H. Purohit, Y. Ruan, D. Fuhry, S. Parthasarathy, and A. Sheth, "On Understanding the Divergence of Online Social Group Discussion," In Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media, pp. 396–405, 2014.

[30] D. Quercia, M. Kosinski, D. Stillwell, and J. Crowcroft, "Our Twitter Profiles, Our Selves: Predicting Personality with Twitter," Proceedings of 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust, and IEEE Third International Conference on Social Computing, pp. 180–185, 2012.

[31] A. Shemyakin, "Hellinger Distance and Non-informative Priors," Bayesian Analysis, 9(4):923–938, 2014.

[32] J. Shlens, "Notes on Kullback-Leibler Divergence and Likelihood Theory," In arXiv preprint arXiv:1404.2000, 2014.

[33] P. Sobhani, S. Mohammad, and S. Kiritchenko, "Detecting Stance in Tweets And Analyzing its Interaction with Sentiment," In Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics, 159–169, 2016.

[34] S. Somasundaran, and J. Wiebe, "Recognizing Stances in Ideological On-Line Debates," In Proceedings of the 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, pp. 116–124, 2010.

[35] D. Sridhar, J. Foulds, B. Huang, L. Getoor, and M. Walker, "Joint Models of Disagreement and Stance in Online Debate," In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pp. 116–125, 2015.

[36] D. Sridhar, L. Getoor, and M. Walker, "Collective Stance Classification of Posts in Online Debate Forums," In Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media, pp. 109–117, 2014.

[37] Q. Sun, Z. Wang, Q. Zhu, and G. Zhou, "Stance Detection with Hierarchical Attention Network," In Proceedings of the 27th International Conference on Computational Linguistics, pp. 2399–2409, 2018.

[38] Q. Sun, Z. Wang, S. Li, Q. Zhu, and G. Zhou, "Stance Detection via Sentiment Information and Neural Network Model," Front. Comput. Sci. 13, 127–138, 2019.

[39] J.M. Tshimula, B. Chikhaoui, and S. Wang, "HAR-search: A Method to Discover Hidden Affinity Relationships in Online Communities," In Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 176–183, 2019.

[40] J.M. Tshimula, B. Chikhaoui, and S. Wang, "A New Approach for Affinity Relationship Discovery in Online Forums," Social Network Analysis and Mining, 10, 40. 2020.

[41] J.M. Tshimula, B. Chikhaoui, and S. Wang, "On Predicting Behavioral Deterioration in Online Discussion Forums," In Proceedings of the 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.

[42] M. Tutek, I. Sekulic, P. Gombar, I. Paljak, F. Culinovic, F. Boltuzic, M. Karan, D. Alagic, and J. Snajder, "TakeLab at SemEval-2016 Task 6: Stance Classification in Tweets Using a Genetic Algorithm Based Ensemble," In Proceedings of the 10th International Workshop on Semantic Evaluation, pp. 464–468, 2016.

[43] M. Walker, P. Anand, R. Abbott, and R. Grant, "Stance Classification using Dialogic Properties of Persuasion," In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 592–596, 2012.

[44] M. Wan and J. McAuley, "Modeling Ambiguity, Subjectivity, and Diverging Viewpoints in Opinion Question Answering Systems," In 2016 IEEE 16th International Conference on Data Mining (ICDM), pp. 489–498, 2016.

[45] A. Williams, N. Nangia, and S. Bowman, "A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference," In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1112–1122, 2018.

[46] W. Youyou, M. Kosinski, and D. Stillwell, "Computer-based personality judgments are more accurate than those made by humans," PNAS, 112(4):1036–1040, 2015.

[47] G. Zarrella, and A. Marsh, "MITRE at SemEval-2016 Task 6: Transfer Learning for Stance Detection," In Proceedings of the 10th International Workshop on Semantic Evaluation, pp. 458–463, 2016.

[48] A.X. Zhang, B. Culbertson, and P. Paritosh, "Characterizing Online Discussion Using Coarse Discourse Sequences," In Proceedings of the Eleventh International AAAI Conference on Web and Social Media, pp. 357–366, 2017.