

“Laughing at you or with you”: The Role of Sarcasm in Shaping the Disagreement Space

Debanjan Ghosh^{*1} Ritvik Shrivastava^{*2,3} Smaranda Muresan³

¹Educational Testing Service

²MindMeld, Cisco Systems

³Data Science Institute, Columbia University

dghosh@ets.org, ritvshri@cisco.com, {rs3868, smara}@columbia.edu

Abstract

Detecting arguments in online interactions is useful to understand how conflicts arise and get resolved. Users often use figurative language, such as sarcasm, either as persuasive devices or to attack the opponent by an ad hominem argument. To further our understanding of the role of sarcasm in shaping the disagreement space, we present a thorough experimental setup using a corpus annotated with both argumentative moves (agree/disagree) and sarcasm. We exploit joint modeling in terms of (a) applying discrete features that are useful in detecting sarcasm to the task of argumentative relation classification (agree/disagree/none), and (b) multi-task learning for argumentative relation classification and sarcasm detection using deep learning architectures (e.g., dual Long Short-Term Memory (LSTM) with hierarchical attention and Transformer-based architectures). We demonstrate that modeling sarcasm improves the argumentative relation classification task (agree/disagree/none) in all setups.

1 Introduction

User-generated conversational data such as discussion forums provide a wealth of naturally occurring arguments. The ability to automatically detect and classify argumentative relations (e.g., agree/disagree) in threaded discussions is useful to understand how collective opinions form, how conflict arises and is resolved (van Eemeren et al., 1993; Abbott et al., 2011; Walker et al., 2012b; Misra and Walker, 2013; Ghosh et al., 2014; Rosenthal and McKeown, 2015; Stede and Schneider, 2018). Linguistic and argumentation theories have thoroughly studied the use of sarcasm in argumentation, including its effectiveness as a persuasive device or as a means to express an ad hominem

*Equal Contribution.

Arg. Rel.	Turn Pairs
Agree	Prior Turn: Today, no informed creationist would deny natural selection.
	Current Turn: Seeing how this was proposed over a century and a half ago by Darwin, what took the creationists so long to catch up?
Disagree	Prior Turn: Personally I wouldn't own a gun for self defense because I am just not that big of a sissy.
	Current Turn: Because taking responsibility for ones own safety is certainly a sissy thing to do?
	Prior Turn: I'm not surprised that no one on your side of the debate would correct you, but wolves and dogs are both members of the same species. The Canid species.
	Current Turn: Wow, you 're even wrong when you get away from your precious Bible and try to sound scientific.
None	Prior Turn: The hand of God kept me from serious harm. Maybe He has a plan for me.
	Current Turn: You better hurry up . Are n't you like 113 years old.

Table 1: Sarcastic turns that disagree, agree or have no argumentative relation with their prior turns.

fallacy (attacking the opponent instead of her/his argument) (Tindale and Gough, 1987; van Eemeren and Grootendorst, 1992; Gibbs and Izett, 2005; Averbeck, 2013). We propose an experimental setup to further our understanding of the role of sarcasm in shaping up the disagreement space in online interactions. The disagreement space, defined in the context of the dialogical perspective on argumentation, is seen as the speech acts initiating the difference of opinions that argumentation is intended to resolve (Jackson, 1992; van Eemeren et al., 1993). Our study is based on the Internet Argument Corpus (IAC) introduced by Abbott et al. (2011) that contains online discussions annotated for the presence/absence and the type of an argumentative move (agree/disagree/none) as well as the presence/absence of sarcasm. Consider the dialogue turns from IAC in Table 1, where the **current turn** (henceforth, *ct*) is a sarcastic response to the

prior turn (henceforth, *pt*). These dialogue moves can be argumentative (agree/disagree) or not argumentative (none). The argumentative move can express agreement (first example) or disagreement (the second example is an undercutter, while the third example is an ad hominem attack). The fourth example, although sarcastic, it is not argumentative. It can be noticed that none of the current turns contain explicit lexical terms that could signal an argumentative relation with the prior turn. Instead, the argumentative move is being implicitly expressed using sarcasm.

We study whether modeling *sarcasm* can improve the detection and classification of *argumentative relations* in online discussions. We propose a thorough experimental setup to answer this question using feature-based machine learning approaches and deep learning models. For the former, we show that *combining* features that are useful to detect sarcasm (Joshi et al., 2015; Muresan et al., 2016; Ghosh and Muresan, 2018) with state-of-the-art argument features leads to better performance for the argumentative relation classification task (agree/disagree/none) (Section 5). For the deep learning approaches, we hypothesize that *multitask learning*, which allows representations to be shared between multiple tasks (e.g., here, the tasks of argumentative relation classification and sarcasm detection), lead to better generalizations. We investigate the impact of multitask learning for a dual Long Short-Term Memory (LSTM) Network with hierarchical attention (Ghosh et al., 2017) (Section 4.2) and BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019), including an optional joint multitask learning objective with uncertainty-based weighting of task-specific losses (Kendall et al., 2018) (Section 4.3). We demonstrate that multitask learning improves the performance of the argumentative relation classification task for all settings (Section 5). We provide a detailed qualitative analysis (Section 5.1) to give insights into when and how modeling sarcasm helps. We make the code from our experiments publicly available.¹ The Internet Argument Corpus (*IAC*) (Walker et al., 2012b) can be found for public access here:²

¹https://github.com/ritvikshrivastava/multitask_transformers

²<https://nlds.soe.ucsc.edu/iac2>

2 Related Work

Argument mining is a growing area of research in computational linguistics, focusing on the detection of argumentative structures in a text (see Stede and Schneider (2018) for an overview). This paper focuses on two subtasks: argumentative relation identification and classification (i.e., agree/disagree/none). Some of the earlier work on argumentative relation identification and classification has relied on feature-based machine learning models, focusing on online discussions (Abbott et al., 2011; Walker et al., 2012b; Misra and Walker, 2013; Ghosh et al., 2014; Wacholder et al., 2014) and monologues (Stab and Gurevych, 2014, 2017; Persing and Ng, 2016; Ghosh et al., 2016). Stab and Gurevych (2014) proposed a set of lexical, syntactic, semantic, and discourse features to classify them. On the same essay dataset, Nguyen and Litman (2016) utilized contextual information to improve the accuracy. Both Stab and Gurevych (2017) and Persing and Ng (2016) used Integer Linear Programming (ILP) based joint modeling to detect argument components and relations. Rosenthal and McKeown (2015) introduced sentence similarity and accommodation features, whereas Menini and Tonelli (2016) presented how entailment between text pairs can discover argumentative relations. Our argumentative features in the feature-based model are based on the above works (Section 4.1). We show that additional features that are useful in sarcasm detection (Joshi et al., 2015; Ghosh and Muresan, 2018) enhance the performance on the argumentative relation identification and classification tasks.

In addition to feature-based models, deep learning models have been recently used for these tasks. Potash et al. (2017) proposed a pointer network, and Hou and Jochim (2017) offered LSTM+Attention network to predict argument components and relations jointly, whereas (Chakrabarty et al., 2019) exploited adaptive pre-training (Gururangan et al., 2020) for BERT to identify argument relations. We use two multitask learning objectives (argumentative relation identification/classification and sarcasm detection), as our goal is to investigate whether identifying sarcasm can help in modeling the disagreement space. Majumder et al. (2019); Chauhan et al. (2020) used multitask learning for sarcasm & sentiment and sarcasm, sentiment, & emotion, respectively, where a direct link between the corresponding tasks is

evident.

Finally, analyzing the role of sarcasm and verbal irony in argumentation has a long history in linguistics (Tindale and Gough, 1987; Gibbs and Izett, 2005; Averbeck, 2013; van Eemeren and Grootendorst, 1992). We propose joint modeling of argumentative relation detection and sarcasm detection to empirically validate sarcasm’s role in shaping the disagreement space in online conversations.

While the focus of our paper is not to provide a state-of-the-art sarcasm detection model, our feature-based models, along with the deep learning models for sarcasm detection are based on state-of-the-art approaches. We implemented discrete features such as pragmatic features (González-Ibáñez et al., 2011; Muresan et al., 2016), diverse sarcasm markers (Ghosh and Muresan, 2018), and incongruity detection features (Riloff et al., 2013; Joshi et al., 2015). The LSTM models are influenced by Ghosh and Veale (2017); Ghosh et al. (2018), where the function of contextual knowledge is used to detect sarcasm. Lastly, transformer models such as BERT and RoBERTa have been used in the winning entries for the recent shared task on sarcasm detection (Ghosh et al., 2020). In our research, for both kinds of deep-learning models, the best results are obtained by using the multitask setup, showing that multitask learning indeed helps improve both tasks.

3 Data

Our *training* and *test* data are collected from the Internet Argument Corpus (*IAC*) (Walker et al., 2012a). This corpus consists of posts from conversations in online forums on a range of controversial political and social topics such as Evolution, Abortion, Gun Control, and Gay Marriage (Abbott et al., 2011, 2016). Multiple versions of *IAC* corpora are publicly available, and we use a particular subset, marked as *IAC_{orig}*, collected from Abbott et al. (2011). This consists of around 10K pairs of conversation turns (i.e., prior turn *pt* and the current turn *ct*) that were annotated using Mechanical Turk for argumentative relations (agree/disagree/none) and other characteristics such as sarcasm/non-sarcasm, respect/insult, nice/nastiness. Median Cohen’s κ is 0.5 across all topics.

For agree/disagree/none relations the annotation was a scalar judgment on an 11 point scale [-5,5] where “-5” indicates a high disagreement move, “0” indicates none relation, and “5” denotes a high

Arg. Rel.	Sarcasm	# of turns
<i>A</i>	<i>S</i>	315 (33%)
	<i>NS</i>	638 (67%)
<i>D</i>	<i>S</i>	2207 (57%)
	<i>NS</i>	1696 (43%)
<i>N</i>	<i>S</i>	2285 (44%)
	<i>NS</i>	2841 (56%)

Table 2: Dataset statistics; A (Agree), D (Disagree), N (None); S (Sarcasm), NS (Non-Sarcasm)

agreement move. We converted the scalar values to three categories: disagree (*D*) for values between [-5, -2], none (*N*) for values between [-1,1], and agree (*A*) for values between [2,5], where the scalar partitions ([]) follow prior work with *IAC* (Misra and Walker, 2013; Rosenthal and McKeown, 2015).

Each “current turn” that is part of a <pt,ct> pair is also labeled with a Sarcasm (*S*) or Non-Sarcasm (*NS*) label. Table 2 shows the data statistics in terms of argumentative relations (*A/D/N*) and sarcasm (*S/NS*). We split the dataset into *training* (80%; 7,982 turn pairs), *test* (10%; 999 turn pairs), and *dev* (10%; 999 turn pairs) sets where each set contains a proportional number of instances (i.e., 80% of 315 (=252) sarcastic turns (*S*) with argument relation label *A* (agree) appears in the training set). The *dev* set is used for parameter tuning.

4 Experimental Setup

We present the computational approaches to investigate whether modeling *sarcasm* can help detect argumentative relations. As our goal is to provide a comprehensive empirical investigation of sarcasm’s role in argument mining rather than propose new models, we explore three separate machine learning approaches well-established for studying argumentation and figurative language. First, we implement a Logistic Regression method that exploits a combination of state-of-the-art features to detect argumentative relations as well as sarcasm (Section 4.1). Second, we present a dual LSTM architecture with hierarchical attention and its multitask learning setup (Section 4.2). Third, we discuss experiments using the pre-trained BERT models and our multitask learning architectures based on it (Section 4.3).

4.1 Logistic Regression with Discrete Features

We use a Logistic Regression (LR) model that uses both argument-relevant ($ArgF$) and sarcasm-relevant ($SarcF$) features. Unless mentioned, all features were extracted from the current turn ct .

Argument-relevant features ($ArgF$). We first evaluate the features that are reported as being useful for identifying and classifying argumentative relations: (a) n -grams (e.g., unigram, bigram, trigram) created based on the full vocabulary of the *IAC* corpus; (b) *argument lexicons*: two lists of twenty words representing agreement (e.g., “agree”, “accord”) and disagreement (e.g., “differ”, “oppose”), respectively (Rosenthal and McKeown, 2015) (c) *sentiment lexicons* such as MPQA (Wilson et al., 2005) and opinion lexicon (Hu and Liu, 2004) to identify sentiment in the turns; (d) *hedge features*, since they are often used to mitigate speaker’s commitment (Tan et al., 2016); (e) *PDTB discourse markers* because *claims* often start with discourse markers such as *therefore*, *so*. We discard markers from the temporal relation; (f) *modal verbs* because they signal the degree of certainty when expressing a claim (Stab and Gurevych, 2014); (g) *pronouns*, since they dialogically point to the previous speaker’s stance; (h) *textual entailment*: captures whether a position expressed in the prior turn is accepted in the current turn (Cabrio and Villata, 2012; Menini and Tonelli, 2016)³; (i) *lemma overlap* to determine topical alignment between the prior and current turn (Somasundaran and Wiebe, 2010). We compute lemma overlap of noun, verbs, and adjectives between the turns, and (j) *negation* to extract explicit negation cues (e.g., “not”, “don’t”) that often signal disagreement.

Sarcasm-relevant features ($SarcF$). As sarcasm-relevant features we use: (a) Linguistic Inquiry Word Count (*LIWC*) (Pennebaker et al., 2001) features to capture the linguistic, social, individual, and psychological processes; (b) measuring *sentiment incongruity*, that is, capturing the number of times the difference in sentiment polarity between the prior turn pt and the current turn ct occurs and number of positive and negative sentiment words in turns (Joshi et al., 2015); (c) *sarcasm markers* used by Ghosh and Muresan (2018), such as *capitalization*, *quotation marks*,

³We used the textual entailment toolkit (AllenNLP) (Gardner et al., 2017).

punctuation, *exclamations* that emphasize a sense of surprisal, *tag questions*, *interjections* because they seem to undermine a literal evaluation, *hyperbole* because users frequently overstate the magnitude of an event in sarcasm, and *emoticons & emojis*, since they often emphasize the sarcastic intent.

We use *SKLL*, an open-source Python package that wraps around the Scikit-learn tool (Pedregosa et al., 2011).⁴ We perform the feature-based experiment using the Logistic Regression model from Scikit-learn.

In the experimental runs, LR_{ArgF} (i.e., model that uses just the $ArgF$ features) denotes the *individual* model and $LR_{ArgF+SarcF}$ (i.e., model that uses both $ArgF$ and $SarcF$ features) is the *joint* model.

4.2 Dual LSTM and Multitask Learning

LSTMs are able to learn long-term dependencies (Hochreiter and Schmidhuber, 1997) and have been shown to be effective in Natural Language Inference (NLI) research, where the task is to establish the *relationship* between multiple inputs (Rocktäschel et al., 2015). This type of architecture is often denoted as the *dual architecture* since one LSTM models the premise and the other models the hypothesis (in Recognizing Textual Entailment (RTE) tasks). Ghosh et al. (2018) used the dual LSTM architecture with hierarchical attention (HAN) (Yang et al., 2016) for sarcasm detection to model the conversation context, and we use their approach in this paper to model the current turn ct and the prior turn pt . HAN implements attention both at the word level and sentence level. The distinct characteristics of this attention is that the word/sentence-representations are weighted by measuring similarity with a word/sentence level context vector, respectively, which are randomly initialized and jointly learned during training (Yang et al., 2016). We compute the vector representation for the current turn ct and prior turn pt and concatenate vectors from the two LSTMs for the final softmax decision (i.e., A , D or N for argumentative relation detection). Henceforth, this dual LSTM architecture is denoted as $LSTM_{attn}$.

To measure the impact of *sarcasm* in argumentative relation detection, we use a multitask learning approach. Multitask learning aims to leverage useful information in multiple related tasks to im-

⁴<https://pypi.org/project/skll/>

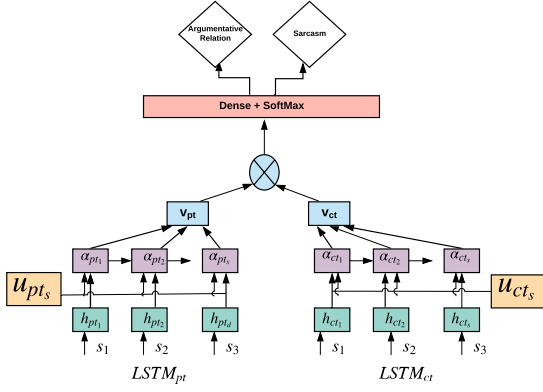


Figure 1: Sentence-level Multitask Attention Network for prior turn pt and current turn ct . Figure is inspired by Yang et al. (2016).

prove each task’s performance (Caruana, 1997; Liu et al., 2019). We use a simple hard parameter sharing network. The architecture is a replica of the $LSTM_{attn}$, with a modification of employing two loss functions, one for sarcasm detection (i.e., training using the S and NS labels) and another for the argumentative relation classification task (i.e., training using the A , D , and N labels).

Figure 1 shows the high-level architecture of the dual LSTM and multitask learning ($LSTM_{MT}$). The prior turn pt (left) and the current turn ct (right) are read by two separate LSTMs (i.e., $LSTM_{pt}$ and $LSTM_{ct}$). In case of $LSTM_{MT}$ the concatenation of v_{pt} and v_{ct} is passed through a dense+Softmax layer for the MTL as shown in Figure 1. Similar to the LR models, $LSTM_{attn}$ now represents the *individual* model (i.e., predicts only the argumentative relation) whereas $LSTM_{MT}$ represents the *joint* model.

Dynamic Multitask Loss. In addition to simply adding the two losses, we also employed *dynamic weighting* of task-specific losses during the training process, based on the homoscedastic uncertainty of tasks, as proposed in Kendall et al. (2018):

$$L = \sum_t \frac{1}{2\sigma_t^2} L_t + \log \sigma_t^2 \quad (1)$$

where L_t and σ_t depict the task-specific loss and its variance, respectively, over training instances. We denote this variation as $LSTM_{MT_{uncert}}$.

4.3 Pretrained BERT and Multitask Learning

BERT (Devlin et al., 2019), a bidirectional transformer model, has achieved state-of-the-art per-

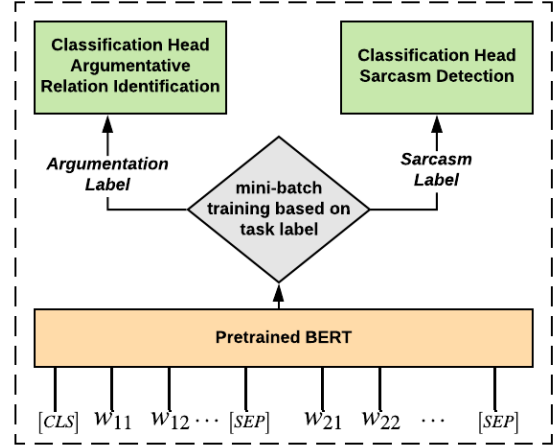


Figure 2: Alternating mini-batch training based on the task type ($BERT_{ALT}$).

formance for many NLP tasks. BERT is initially trained on masked token prediction and next sentence prediction tasks over large corpora (English Wikipedia and Book Corpus). During its training, a special token “[CLS]” is added to the beginning of each training instance, and the “[SEP]” tokens are added to indicate the end of utterance(s) and separate, in case of two utterances (e.g., pt and ct). During the evaluation, the learned representation for the “[CLS]” token is processed by an additional layer with nonlinear activation. In its standard form, pre-trained BERT (“bert-base-uncased”) can be used for transfer learning by fine-tuning on a downstream task, i.e., argument relation detection where training instances are labeled as A , D , and N . We denote the BERT baseline model as $BERT_{orig}$ that is fine-tuned over the *training* partition of only the argumentative relation data (i.e., individual task training). Unless mentioned otherwise, we use the BERT predictions available via the “[CLS]” token. To this end, we propose a couple of variations in the multitask learning settings, and they are briefly described in the following sections.

Multitask Learning with BERT. The first model we use for multitask learning is denoted as $BERT_{MT}$ (i.e., BERT Multitask Learning). Here, we pass the BERT output embeddings to two classification heads - one for each task (i.e., detection of argumentative relation and sarcasm), and the relevant gold labels are passed to them. Each classification head is a linear layer (size=3 and 2 for # of labels for argumentative relation and sarcasm detection, respectively) applied on top of the pooled BERT output. The losses from these individual

heads are added and propagated back through the model. This allows BERT to model the nuances of both tasks and their interdependence simultaneously.

Dynamic Loss: Similar to the LSTM architecture, here, too, we experiment with dynamic multitask loss. We denote this variation as $BERT_{MT_{uncert}}$.

Alternate Multitask Learning. We employ another multitask learning technique where we attempt to enrich the learning with fine-tuning of labeled *additional* material from the sarcasm detection task. Notably, we exploit “sarcasm V2”, a sarcasm detection dataset that was also curated from the original corpus of *IAC* and was released by Oraby et al. (2016). We pre-process the “sarcasm V2” dataset by removing duplicates that appear in IAC_{orig} and we end up selecting 3513 $training_{v2}$ instances and 423 dev_{v2} instances balanced between S/N/S categories for experiments and merged them to the sarcasm dataset ($training$ and dev , respectively) from IAC_{orig} . Note, unlike the original multitask setting, this time we have more sarcastic instances (a total of 11,495) than instances labeled with argumentative roles (7,982 instances as before) for the training purpose, while keeping the $test$ set from IAC_{orig} unchanged.

Since the $training$ data is now unequal between the two tasks of argumentative relation and sarcasm detection, we create mini-batches so that each batch consists of instances with only one task label (i.e., either argumentative labels or sarcasm labels). The batches from the two tasks are interleaved uniformly, i.e., the BERT model is only passed to one of the two tasks’ specific classification heads, and the related loss is used to update the parameters in that iteration. This way, the model trains both tasks but alternates between the two tasks per mini-batch iteration while the extra batches of sarcasm data from the “sarcasm V2” dataset are managed at the end together. This model is denoted as $BERT_{ALT}$ (see Figure 2).

For brevity, all models’ parameter tuning description (e.g., Logistic Regression, Dual LSTM, BERT) is in the supplemental material.

5 Results and Discussion

Table 3 presents the classification results on the $test$ set. We report F1 scores for each class (A , D and N) and Micro-F1 overall score ($F1_{micro}$) (used to account for multi-class and class imbalance).

Model	$F1_{micro}$	A	D	N
LR_{ArgF}	53.5	22.4	57.2	56.3
$LR_{ArgF+SarcF}$	56.4 ^{α^*}	31.0	58.4	58.9
$LSTM_{Attn}$	51.8	28.0	49.4	59.2
$LSTM_{MT}$	53.1	30.0	53.2	56.5
$LSTM_{MT_{uncert}}$	54.6 ^{α^*}	33.1	54.5	58.5
$BERT_{orig}$	62.2	41.8	63.3	64.4
$BERT_{MT}$	63.2	44.5	64.1	65.4
$BERT_{MT_{uncert}}$	65.3 ^{α^*}	44.6	66.2	67.5
$BERT_{ALT}$	63.4	40.1	62.2	66.9

Table 3: Results for argumentative relation detection ($F1_{micro}$ and F1 scores/category) on the $test$ set of IAC_{orig} . ^{α^*} depict significance on $p \leq 0.05$ (measured via McNemar’s test) against the corresponding individual model (e.g., LR_{ArgF} , $LSTM_{Attn}$, $BERT_{orig}$, respectively). Highest scores per group of models are in **bold**.

The LR model using both the $SarcF$ and $ArgF$ features performs better than the model that uses $ArgF$ features alone, improving the overall performance by an absolute 2.9% $F1_{micro}$, and showing a huge impact on the agreement class (A) (8.6% absolute improvement). Table 4 shows the *top* discrete features for argumentative relation identification. From $ArgF$ features (first column), we notice discourse expansion (“particularly”), contrast (“although”) and agree/disagree lexicon getting high feature weights. We also notice *pronouns* receive large feature weights because argumentative text often refers to personal stance (e.g., “you think”, “I believe”). However, when analyzing $ArgF + SarcF$ features we find various sarcasm markers, such as tag questions, hyperbole, multiple punctuation, or sarcasm characteristics such as sentiment incongruity receive the highest weights.

For **LSTM models**, we see that multitask learning helps, $LSTM_{MT_{uncert}}$ showing a 2.8% improvement over the single model $LSTM_{Attn}$, which is statistically significant. Moreover, we notice that the improvement for the agree (A) and disagree (D) classes is 5.1%, with just a small reduction for the none (N) class (0.7%).

For **BERT**, we notice better results when performing multitask learning, while the best performing model is obtained from $BERT_{MT_{uncert}}$ where we experimented with the dynamic weighting of task-specific losses during the training process (Kendall et al., 2018). The performance increase is consistent across all three classes. The difference in performance among each setup is sta-

LR _{ArgF}	LR _{ArgF+SarcF}
<i>pronouns</i> : I, my (both <i>A</i>), your(s) (<i>D</i>); <i>discourse</i> : so, because, for (all <i>A</i>), incidentally, particularly, although (all <i>D</i>); <i>disagree.Lexicon</i> : disagree, differ (both <i>D</i>); <i>agree.Lexicon</i> : agreed (<i>A</i>); <i>entailment relation</i> : negation (<i>D</i>)	<i>pronouns</i> : mine, my (both <i>A</i>), you (<i>D</i>); <i>discourse</i> : then (<i>A</i>), though, however (both <i>D</i>); <i>modal</i> : will (<i>A</i>); <i>punctuation</i> : multiple question marks (both <i>A</i> and <i>D</i>); <i>tag question</i> : “are you”, “do you” (both <i>D</i>); <i>hyperbole</i> : wonderful (<i>A</i>), nonsense, biased (both <i>D</i>); <i>LIWC dimensions</i> : anxiety, assent, certainty (all <i>D</i>); <i>sentiment incongruity</i> (<i>D</i>); <i>interj</i> : so, agreed (both <i>A</i>)

Table 4: Top discrete features from LR_{ArgF} and LR_{ArgF+SarcF} models, respectively. *A* and *D* depict the argumentative relations (agree and disagree) for the particular feature.

tistically significant, as shown in Table 3. Moreover, BERT_{MT_{uncert}} model improves the $F1_{micro}$ by a large margin when compared to the LR and the LSTM models. However, adding more data for the auxiliary task (i.e., sarcasm detection) as presented in BERT_{ALT} did not provide any significant improvement, only a 0.2 improvement of $F1_{micro}$ over BERT_{MT} (however it does show improvement over the single task model). The reason could be that although “sarcasm V2” is a subset of the original *IAC* corpus, it was annotated by a different set of Turkers than *IAC_{orig}* with different annotation guidelines.

Between the three classes - *A*, *D*, and *N* - we observe the lowest performance on the *A* class. This is unsurprising, given the highly unbalanced setting of the *training* data (*A* occurs less than 10% of times in the *IAC_{orig}*, see Table 2).

In sum, these improvements through multitask learning over single task argumentative relation detection indicate that modeling sarcasm is useful in modeling the disagreement space in online discussions. This provides an empirical justification to existing theories that study sarcasm’s impact in modeling argumentation, persuasion, and argument fallacies such as ad hominem attacks. Finally, we notice that multitask learning also improves the performance on the sarcasm detection task (results are presented in the Appendix).

5.1 Qualitative Analysis

To further investigate the effect of multitask learning, we present qualitative analysis studies to:

1. Understand the models’ performance by look-

ing at the turns correctly classified by the multitask models and misclassified by the corresponding individual single task model. We analyze the turns in terms of sarcastic characteristics - whether they depict incongruity, humor, or sarcasm indicators (i.e., markers).

2. Understand when both multitask and individual model made incorrect predictions.

We compare the predictions between the multitask and the individual models for different settings to address the first issue. For example, BERT_{MT_{uncert}} correctly identifies 6 *A*, 50 *D*, and 60 *N* instances more than BERT_{Orig} (out of 91, 398 and 510 instances, respectively). Two of the authors independently investigated a random sample of 100 instances (*qual* set) chosen from the union of the *test* instances that are correctly predicted only by the multitask models (LR_{ArgF+SarcF}, LSTM_{MT_{uncert}}, BERT_{MT_{uncert}}, and BERT_{ALT}) and not by the corresponding individual models (LR_{ArgF}, LSTM_{attn}, and BERT_{Orig}). For both Transformer and LSTM-based models, we explore how attention heads behave and whether common patterns exist (e.g., attending words with opposite meaning when incongruity occurs). We display the heat maps of the attention weights for a pair of prior and current turns (LSTM-based models) (Figure 3) whereas for BERT we display word-to-word attentions (Figures 4, 5, 6, 7, and 8) using visualization tools (Vig, 2019; Yang and Zhang, 2018).⁵ All the examples presented in this section are argumentative moves (i.e., turns with *A* or *D*) correctly identified by our multitask learning models but wrongly predicted as none (*N*) by the individual models. Moreover, the multitask learning models also correctly predict that these turns are instances of sarcasm.

Incongruity between prior turn and current turn. Semantic incongruity, which can appear between conversation context *pt* and the current turn *ct* is an inherent characteristic of sarcasm (Joshi et al., 2015). This characteristic highlights the inconsistency between *expectations* and *reality*, making sarcasm or irony highly effective in persuasive communication (Gibbs and Izett, 2005).

⁵Clark et al. (2019) have probed different layers and attention heads in BERT to find patterns, e.g., whether a token consistently attends a fixed token in a specific layer. To avoid confusion and bias, we select attention examples from only the middle (layer=6) layer.

I 'm not saying the primates changed all of a sudden , but what all of a sudden made them want to change .	I 'm not saying the primates changed all of a sudden , but what all of a sudden made them want to change .
You actually think Evolution works by what creators want ? You think they just got up one day and said , “ ya know Bob , I wan na evolve . ” !! Oops , there goes my tail	You actually think Evolution works by what creators want ? You think they just got up one day and said , “ ya know Bob , I wan na evolve . ” !! Oops , there goes my tail

Figure 3: Attention heatmap of a particular turn pair from $LSTM_{attn}$ (left) and $LSTM_{MT_{uncert}}$ (right) showing higher weights on sarcasm marker such as “Oops” and “!!” for $LSTM_{MT_{uncert}}$ (disagree relation)

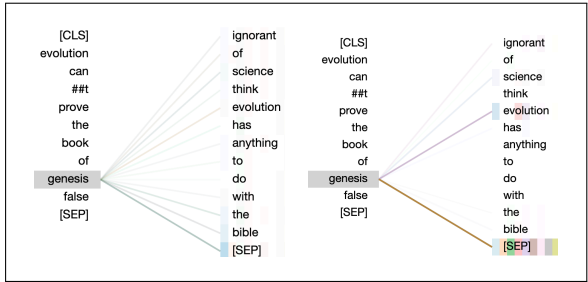


Figure 4: $BERT_{MT_{uncert}}$ (right) attending contrasting words more in word-level attention in comparison to $BERT_{Orig}$ (left) (disagree relation)

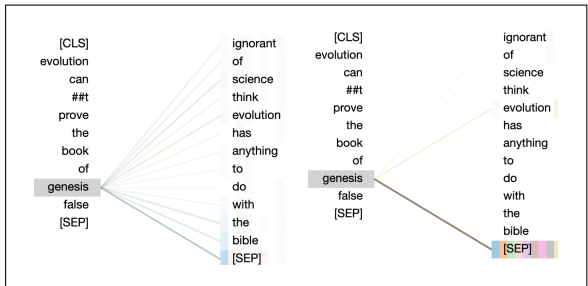


Figure 5: $BERT_{ALT}$ (right) attending only contrasting words in comparison to $BERT_{Orig}$ (left) (disagree relation). However, the strength of the contrast in the case of $BERT_{ALT}$ is lower than $BERT_{MT_{uncert}}$ for the same example turns.

In the case of BERT, Figure 4 presents the turns “evolution can’t prove the book of genesis false” (*pt*) ↔ “ignorant of science think evolution has anything to do with the bible” (*ct*). Here, $BERT_{MT_{uncert}}$ shows more attention between incongruous terms (“genesis” ↔ “science”, “evolution”) as well as to the mocking word “ignorance”. Likewise, Figure 6 presents two turns “you are quite anti religious it seems” (*pt*) ↔ “anti ignorance and superstition . . . this is religion” (*ct*). We notice the word “religious” is attending “anti” and “ignorance” with high weights in case of $BERT_{MT_{uncert}}$ (from *pt* to *ct*) whereas $BERT_{Orig}$ only attends to the word “religious”

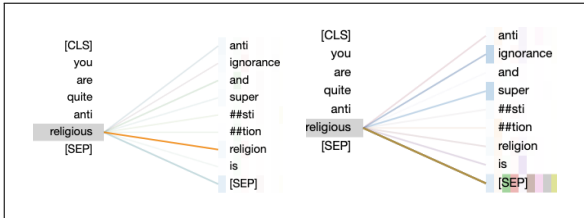


Figure 6: $BERT_{MT_{uncert}}$ (right) attending contrasting words more than $BERT_{Orig}$ (left) (disagree relation)

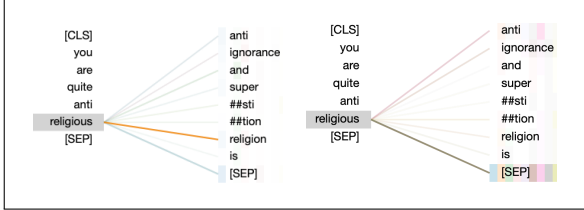


Figure 7: $BERT_{ALT}$ (right) attending only the contrasting words in comparison to $BERT_{Orig}$ (left) (disagree relation)

from the *pt* to *ct* turn. By modeling sarcasm, the multitask learning models can better predict argumentative moves that are expressed implicitly.

We also evaluate the $BERT_{ALT}$ model for the examples presented in Figure 4 and Figure 6. Figure 5 shows that although $BERT_{ALT}$ is attending (from *pt* to *ct*) incongruous terms “genesis” ↔ “evolution”, the strength of the relation (i.e., attention weight) is comparatively lower than $BERT_{MT_{uncert}}$ (See Figure 4). On the contrary, between Figure 6 and Figure 7, $BERT_{MT_{uncert}}$ model is attending multiple words in *ct* from the word “religious” in *pt*, but the $BERT_{ALT}$ model attends only two words ‘anti’ and “ignorance”, with high weights from “religious” (*pt* to *ct*).

Humor by word repetition. Often the current turn *ct* sarcastically taunts the prior turn *pt* by word repetition and rhyme, imposing a humorous comic effect, also regarded as the phonetic style of humor (Yang et al., 2015). For the pair, “genetics

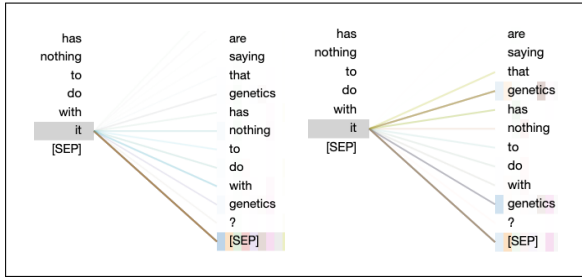


Figure 8: $BERT_{MT_uncert}$ (right) attending co-referenced words in a humorous example missed by the $BERT_{Orig}$ model (left) (disagree relation)

has nothing to do with it” (pt) \leftrightarrow “are saying that genetics has nothing to do with genetics?” (ct), we notice in $BERT_{MT_uncert}$ the token “it” in pt correctly attends to both occurrences of “genetics” in ct where the second occurrence is the co-reference of “it” (Figure 8), which is missed by the individual model $BERT_{Orig}$.

Role of sarcasm markers. Sarcasm markers are indicators that alert if an utterance is sarcastic (Attardo, 2000). While comparing the logistic regression models between $LR_{ArgF+SarcF}$ and LR_{ArgF} , we observe markers such as multiple punctuations (“????”), tag question (“are you”), upper case (“NOT”) have received the highest features weights (Table 4). In Figure 3, while the individual model $LSTM_{attn}$ attends the words almost equally, we notice in the multitask variation several sarcasm markers such as “ya”, “oops”, and numerous exclamations (“!!!”) receive larger attention weights.

Addressing the second issue (i.e., when both multitask and single tasks models make the wrong predictions), we notice that over 100 examples of none (N) class were classified as argumentative by both $BERT_{MT_uncert}$ and $BERT_{Orig}$. For the none N class, one of the most common instances of wrong predictions is when the current turn ct sarcastically takes a “different stance” on a topic from pt in a narrow context but the whole turn is not argumentative. In the following example: “does he just say the opposite of everything <name> says?” (pt) \leftrightarrow “using <name> as a 180 compass is just fine by me” (ct), $BERT_{MT_uncert}$, $BERT_{Orig}$, $LSTM_{MT_uncert}$, and $LSTM_{attn}$ models make disagree D prediction (since ct is sarcastic on “<name>”) where the gold label is none N . Looking closely at this pair of turns, it seems that the ct presents a case of ad hominem attack (on the person’s “<name>”) rather than a none relation.

In the case of argumentative turns (agree and disagree) that are wrongly classified as none by all models, we found two common patterns: the use of concessions (e.g., “it’s a consideration, *but* I doubt we should be promoting this . . .”) and arguments with uncommitted beliefs (e.g., “it is *possible* that”, “that could *probably* be”, “*possibly*, I must admit”).

6 Conclusion and Future Work

Linguistic and argumentation theories have studied the use of sarcasm in argumentation, including its effectiveness as a persuasive device or as a means to express an ad hominem fallacy. We present a comprehensive experimental study for argumentative relation identification and classification using sarcasm detection as an additional task. First, in discrete feature space, we show that sarcasm-related features, in addition to argument-related features, improve the accuracy of the argumentative relation identification/classification task by 3%. Next, we show that multitask learning using both a dual LSTM framework and BERT helps improve performance compared to the corresponding single model by a statistically significant margin. In both cases, the dynamic weighting of task specific losses performs best. We provide a detailed qualitative analysis by investigating a large sample manually and show what characteristics of sarcasm are attended to, which might have guided the correct prediction on the identification of the argumentative relation/classification task. In the future, we aim to study this synergy further by looking at sarcasm as well as the persuasive strategies (e.g., ethos, pathos, logos), and argument fallacies (e.g., ad hominem attack that was also noticed by Habernal et al. (2018)).

Acknowledgements

The authors thank the anonymous reviewers and Tuhin Chakrabarty for helpful comments.

References

- Rob Abbott, Brian Ecker, Pranav Anand, and Marilyn Walker. 2016. 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)*, pages 4445–4452.
- Rob Abbott, Marilyn Walker, Pranav Anand, Jean E Fox Tree, Robeson Bowmani, and Joseph King.

2011. How can you say such things?!?: Recognizing disagreement in informal political argument. In *Proceedings of the Workshop on Languages in Social Media*, pages 2–11. Association for Computational Linguistics.
- Salvatore Attardo. 2000. Irony markers and functions: Towards a goal-oriented theory of irony and its processing. *Rask*, 12(1):3–20.
- Joshua M. Averbeck. 2013. Comparisons of ironic and sarcastic arguments in terms of appropriateness and effectiveness in personal relationships. *Argumentation and Advocacy*, 50(1):47–57.
- Elena Cabrio and Serena Villata. 2012. Combining textual entailment and argumentation theory for supporting online debates interactions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 208–212.
- Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1):41–75.
- Tuhin Chakrabarty, Christopher Hidey, Smaranda Muresan, Kathy McKeown, and Alyssa Hwang. 2019. AMPERSAND: Argument mining for PER-SuAsive oNline discussions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2933–2943, Hong Kong, China. Association for Computational Linguistics.
- Dushyant Singh Chauhan, SR Dhanush, Asif Ekbal, and Pushpak Bhattacharyya. 2020. Sentiment and emotion help sarcasm? a multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4351–4360.
- Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT’s attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Frans H. van Eemeren and Rob Grootendorst. 1992. *Argumentation, communication, and fallacies: a pragma-dialectical perspective*. Lawrence Erlbaum Associates, Inc.
- Frans Hendrik van Eemeren, Rob Grootendorst, Sally Jackson, Scott Jacobs, et al. 1993. *Reconstructing argumentative discourse*. University of Alabama Press.
- Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. *Allennlp: A deep semantic natural language processing platform*.
- Aniruddha Ghosh and Tony Veale. 2017. Magnets for sarcasm: Making sarcasm detection timely, contextual and very personal. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 482–491.
- Debanjan Ghosh, Alexander R Fabbri, and Smaranda Muresan. 2018. Sarcasm analysis using conversation context. *Computational Linguistics*, 44(4):755–792.
- Debanjan Ghosh, Alexander Richard Fabbri, and Smaranda Muresan. 2017. The role of conversation context for sarcasm detection in online interactions. In *Proceedings of the 18th Annual SIG-Dial Meeting on Discourse and Dialogue*, pages 186–196, Saarbrücken, Germany. Association for Computational Linguistics.
- Debanjan Ghosh, Aquila Khanam, Yubo Han, and Smaranda Muresan. 2016. **Coarse-grained argumentation features for scoring persuasive essays**. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 549–554, Berlin, Germany. Association for Computational Linguistics.
- Debanjan Ghosh and Smaranda Muresan. 2018. “with 1 follower i must be awesome :p”. exploring the role of irony markers in irony recognition. *Proceedings of ICWSM*.
- Debanjan Ghosh, Smaranda Muresan, Nina Wacholder, Mark Aakhus, and Matthew Mitsui. 2014. Analyzing argumentative discourse units in online interactions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 39–48.
- Debanjan Ghosh, Avijit Vajpayee, and Smaranda Muresan. 2020. A report on the 2020 sarcasm detection shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 1–11, Online. Association for Computational Linguistics.
- Raymond W Gibbs and Christin Izett. 2005. Irony as persuasive communication. *Figurative language comprehension: Social and cultural influences*, pages 131–151.
- Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. 2011. **Identifying sarcasm in twitter: A closer look**. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages

- 581–586, Portland, Oregon, USA. Association for Computational Linguistics.
- Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don’t stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.
- Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 386–396, New Orleans, Louisiana. Association for Computational Linguistics.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Yufang Hou and Charles Jochim. 2017. [Argument relation classification using a joint inference model](#). In *Proceedings of the 4th Workshop on Argument Mining*, pages 60–66, Copenhagen, Denmark. Association for Computational Linguistics.
- Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.
- S Jackson. 1992. Virtual standpoints’ and the pragmatics of conversational argument. *Argumentation illuminated*, page 260–269.
- Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. 2015. Harnessing context incongruity for sarcasm detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 757–762, Beijing, China. Association for Computational Linguistics.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Herve Jégou, and Tomas Mikolov. 2016. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.
- Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491.
- Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.
- Navonil Majumder, Soujanya Poria, Haiyun Peng, Niyati Chhaya, Erik Cambria, and Alexander Gelbukh. 2019. Sentiment and sarcasm classification with multitask learning. *IEEE Intelligent Systems*, 34(3):38–43.
- Stefano Menini and Sara Tonelli. 2016. Agreement and disagreement: Comparison of points of view in the political domain. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2461–2470.
- Amita Misra and Marilyn A Walker. 2013. Topic independent identification of agreement and disagreement in social media dialogue. In *Proceedings of the SIGDIAL 2013 Conference*, pages 41–50. Association for Computational Linguistics.
- Smaranda Muresan, Roberto Gonzalez-Ibanez, Debanjan Ghosh, and Nina Wacholder. 2016. [Identification of nonliteral language in social media: A case study on sarcasm](#). *Journal of the Association for Information Science and Technology*.
- Huy Nguyen and Diane Litman. 2016. Context-aware argumentative relation mining. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1127–1137.
- Shereen Oraby, Vrindavan Harrison, Lena Reed, Ernesto Hernandez, Ellen Riloff, and Marilyn Walker. 2016. Creating and characterizing a diverse corpus of sarcasm in dialogue. In *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 31.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.
- James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71:2001.
- Isaac Persing and Vincent Ng. 2016. [Modeling stance in student essays](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2174–2184, Berlin, Germany. Association for Computational Linguistics.
- Peter Potash, Alexey Romanov, and Anna Rumshisky. 2017. Here’s my point: Joint pointer architecture for argument mining. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1364–1373, Copenhagen, Denmark. Association for Computational Linguistics.

- Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 704–714. Association for Computational Linguistics.
- Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiský, and Phil Blunsom. 2015. Reasoning about entailment with neural attention. *arXiv preprint arXiv:1509.06664*.
- Sara Rosenthal and Kathleen McKeown. 2015. I couldn't agree more: The role of conversational structure in agreement and disagreement detection in online discussions. In *16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 168.
- Swapna Somasundaran and Janyce Wiebe. 2010. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 116–124. Association for Computational Linguistics.
- Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *EMNLP*, pages 46–56.
- Christian Stab and Iryna Gurevych. 2017. [Parsing argumentation structures in persuasive essays](#). *Computational Linguistics*, 43(3):619–659.
- Manfred Stede and Jodi Schneider. 2018. Argumentation mining. *Synthesis Lectures on Human Language Technologies*, 11(2):1–191.
- Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of WWW*.
- Christopher W Tindale and James Gough. 1987. The use of irony in argumentation. *Philosophy & rhetoric*, pages 1–17.
- Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy. Association for Computational Linguistics.
- Nina Wacholder, Smaranda Muresan, Debanjan Ghosh, and Mark Aakhus. 2014. Annotating multiparty discourse: Challenges for agreement metrics. *LAW VIII*, page 120.
- Marilyn A Walker, Pranav Anand, Robert Abbott, and Ricky Grant. 2012a. 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*, pages 592–596. Association for Computational Linguistics.
- Marilyn A Walker, Jean E Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012b. A corpus for research on deliberation and debate. In *LREC*, pages 812–817.
- Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics.
- Diyi Yang, Alon Lavie, Chris Dyer, and Eduard Hovy. 2015. Humor recognition and humor anchor extraction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2367–2376.
- Jie Yang and Yue Zhang. 2018. [Ncrf++: An open-source neural sequence labeling toolkit](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.
- Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of NAACL-HLT*, pages 1480–1489.

A Supplemental Material

A.1 Parameter Tuning

Logistic Regression (LR) experiment: A Logistic Regression model with L_2 penalty is employed where the class weights are proportional to the number of instances for A , D and N classes. The regularization strength C is searched over a grid using the *dev* data. Following values were tried for c : [.0001, .001, .01, .1, 1, 10, 100, 1000, 10000].

Dual LSTM and Multi-task Learning experiment: For LSTM networks based experiments we searched the hyper parameters over the *dev* set. Particularly we experimented with different mini-batch size (e.g., 8, 16, 32), dropout value (e.g., 0.3, 0.5, 0.7), number of epochs (e.g., 40, 50), hidden state of different sized-vectors (100, 300) and the Adam optimizer (learning rate of 0.01). Embeddings were generated using FastText vectors (300 dimensions) (Joulin et al., 2016). Any token occurring less than five times were replaced by a special UNK token where the UNK vector is created based on random samples from a normal (Gaussian) distribution between 0.0 and 0.17. After tuning we use the following hyper-parameters for the *test* set: mini-batch size of 16, hidden state of size 300, number of epochs = 50, and dropout value of 0.5. Task-specific losses for the dynamic multitask version was learned during training.

BERT based models: We use the *dev* partition for hyperparameter tuning such as different mini-batch size (e.g., 8, 16, 32, 48), number of epochs (3, 5, 6), learning rate of 3e-5) and optimized networks with the Adam optimizer. The training partitions were fine-tuned for 5 epochs with batch size = 16. Each training epoch took between 08:46 ~ 9 minutes over a K-80 GPU with 48GB vRAM.

A.2 Results on the Sarcasm Detection Task

Although improving sarcasm detection is not the focus our paper, we observe that multi-task learning improves the performance on this task as well, when compared to the single task model. We present results for the deep learning models in Table 5. The multi-task models (both for LSTM and BERT) outperform the corresponding single task models (by 6.9 F1 and 6.4 F1 for LSTM and BERT models, respectively). We note that the results on this particular dataset are much lower than on other datasets used for sarcasm detection. For example,

the LSTM_{Attn} which is the best model used by Ghosh et al. (2018) obtained only 52.9 F1 score on this dataset, while it obtained 70.34 F1 on Sarcasm V2 (derived also from IAC but using different annotation guidelines), 74.96 F1 on a Twitter dataset and 75.41 F1 on a Reddit dataset (Ghosh et al., 2018).

Model	Precision	Recall	F1
LSTM _{Attn}	52.9	52.8	52.9
LSTM _{MT}	59.5	59.3	59.4
BERT _{orig}	57.4	57.4	57.4
BERT _{MT}	61.8	61.7	61.8
BERT _{MT_{uncert}}	64.1	63.5	64.0

Table 5: Evaluations of sarcasm detection on the *test* set of IAC_{orig}.