Issue Framing in Online Discussion Fora

Mareike Hartmann[1]   **Tallulah Jansen**[2]   **Isabelle Augenstein**[1]   **Anders Søgaard**[1]

[1]Dep. of Computer Science, University of Copenhagen, Denmark
`{hartmann,augenstein,soegaard}@di.ku.dk`
[2]Inst. of Cognitive Science, Osnabrück University, Germany
`taljansen@uni-osnabrueck.de`

Abstract

In online discussion fora, speakers often make arguments for or against something, say birth control, by highlighting certain aspects of the topic. In social science, this is referred to as *issue framing*. In this paper, we introduce a new issue frame annotated corpus of online discussions. We explore to what extent models trained to detect issue frames in newswire and social media can be transferred to the domain of discussion fora, using a combination of multi-task and adversarial training, assuming only unlabeled training data in the target domain.

1 Introduction

The *framing* of an issue refers to a choice of perspective, often motivated by an attempt to influence its perception and interpretation (Entman, 1993; Chong and Druckman, 2007). The way issues are framed can change the evolution of policy as well as public opinion (Dardis et al., 2008; Iyengar, 1991). As an illustration, contrast the statement *Illegal workers depress wages* with *This country is abusing and terrorizing undocumented immigrant workers*. The first statement puts focus on the economic consequences of immigration, whereas the second one evokes a morality frame by pointing out the inhumane conditions under which immigrants may have to work. Being exposed to primarily one of those perspectives might affect the publics attitude towards immigration.

Computational methods for frame classification have previously been studied in news articles (Card et al., 2015) and social media posts (Johnson et al., 2017). In this work, we introduce a new benchmark dataset, based on a subset of the 15 generic frames in the *Policy Frames Codebook* by Boydstun et al. (2014). We focus on frame classification in *online discussion fora*, which have be-

Platform: Online discussions
Economic Frame, Topic: Same sex marriage `But as we have seen, supporting same-sex marriage saves money.`
Legality Frame, Topic: Same sex marriage `So you admit that it is a right and it is being denied?`
Platform: News articles
Economic Frame, Topic: Immigration `Study Finds That Immigrants Are Central to Long Island Economy`
Legality Frame, Topic: Same sex marriage `Last week, the Iowa Supreme Court granted same-sex couples the right to marry.`
Platform: Twitter
Legality Frame, Topic: Same sex marriage `Congress must fight to ensure LGBT people have the full protection of the law everywhere in America. #EqualityAct`

Table 1: Example instances from the datasets described in §2 and 3.

come crucial platforms for public dialogue on social and political issues. Table 1 shows example annotations, compared to previous annotations for news articles and social media. Dialogue data is substantially different from news articles and social media, and we therefore explore ways to transfer information from these domains, using multi-task and adversarial learning, providing non-trivial baselines for future work in this area.

Contributions We present a new issue-frame annotated dataset that is used to evaluate issue frame classification in online discussion fora. Issue frame classification was previously limited to news and social media. As manual annotation is expensive, we explore ways to overcome the lack of labeled training data in the target domain with

Frames	1	13	5	6	7
# instances	78	96	234	166	186

Table 2: Class distribution in the online discussion test set. The frame labels correspond to the classes *Economic* (1), *Political* (13), *Legality, Jurisprudence and Constitutionality* (5), *Policy prescription and evaluation* (6) and *Crime and Punishment* (7).

multi-task and adversarial learning, leading to improved results in the target domain.[1]

Related Work Previous work on automatic frame classification focused on news articles and social media. Card et al. (2016) predict frames in news articles at the document level, using clusters of latent dimensions and word-based features in a logistic regression model. Ji and Smith (2017) improve on previous work integrating discourse structure into a recursive neural network. Naderi and Hirst (2017) use the same resource, but make predictions at the sentence level, using topic models and recurrent neural networks. Johnson et al. (2017) predict frames in social media data at the micro-post level, using probabilistic soft logic based on lists of keywords, as well as temporal similarity and network structure. All the work mentioned above uses the generic frames of Boydstun et al. (2014)'s Policy Frames Codebook. Baumer et al. (2015) predict words perceived as frame-evoking in political news articles with hand-crafted features. Field et al. (2018) analyse how Russian news articles frame the U.S. using a keyword-based cross-lingual projection setup. Tsur et al. (2015) use topic models to analyze issue ownership and framing in public statements released by the US congress. Besides work on frame classification, there has recently been a lot of work on aspects closely related to framing, such as subjectivity detection (Lin et al., 2011), detection of biased language (Recasens et al., 2013) and stance detection (Mohammad et al., 2016; Augenstein et al., 2016; Ferreira and Vlachos, 2016).

2 Online Discussion Annotations

We create a new resource of issue-frame annotated online fora discussions, by annotating a subset of the Argument Extraction Corpus (Swanson et al., 2015) with a subset of the frames in the Policy Frames Codebook. The Argument Extraction

Corpus is a collection of argumentative dialogues across topics and platforms.[2] The corpus contains posts on the following topics: *gay marriage*, *gun control*, *death penalty* and *evolution*. A subset of the corpus was annotated with argument quality scores by Swanson et al. (2015), which we exploit in our multi-task setup (see §3).

We collect new issue frame annotations for each argument in the argument-quality annotated data.[3] We refer to this new issue-frame annotated corpus as *online discussion corpus* henceforth. Each argument can have one or multiple frames. Following Naderi and Hirst (2017), we focus on the five most frequent issue frames: *Economic*, *constitutionality and jurisprudence*, *policy prescription and evaluation*, *law and order/crime and justice*, and *political*. See Table 1 for examples and Table 2 for the class distribution in the resulting online discussions test set. Phrases which do not match the five categories are labeled as *Other*, but we do not consider this class in our experiments. The annotations were done by a single annotator. A second annotator labeled a subset of 200 instances that we use to compute agreement as macro-averaged F-score, assuming one of the annotations as gold standard. Results are 0.73 and 0.7, respectively. The averaged Cohen's Kappa is 0.71.

3 Additional Data

The dataset described in the previous section serves as evaluation set for the online discussions domain. As we do not have labeled training data for this domain, we exploit additional corpora and additional annotations, which are described in the next subsection. Statistics of the filtered datasets as well as preprocessing details are given in Appendix A.

Media Frames Corpus The Media Frames Corpus (Card et al., 2015) contains US newspaper articles on three topics: *Immigration*, *smoking* and *same-sex marriage*. The articles are annotated with the 15 framing dimensions defined in the Policy Frames Codebook.[4] The annotations are on

---

[1]Code and annotations are available at https://github.com/coastalcph/issue_framing.

[2]The corpus is a combination of dialogues from http://www.createdebate.com/, and Walker et al. (2012)'s Internet Argument Corpus, which contains dialogues from 4forums.com.

[3]Topic cluster *Evolution* was dropped, because it contained too few examples matching our frame categories.

[4]We discard all instances that do not correspond to the frame categories in the online discussions data.

Model	Task	Domain	Labelset	# classes	# sequences
Baseline	Main task	News articles	Frames	5	10,480
	Target task	Online disc. (test)	Frames	5	692
Multitask	+Aux task	Tweets	Frames	5	1,636
	+Aux task	Online disc.	Argument quality	2	3,785
Adversarial	+Adv task	Online disc. + News articles	Domain	2	4,731 + 10,480
		Online disc. (dev)	Frames	5	176

Table 3: Overview over the data and labelsets for the different tasks. The baseline model trains on the main task and predicts the target task. The multi-task model uses one or both auxiliary tasks in addition to the main task. The adversarial model uses the adversarial task in addition to the main task. All models use the online disc. dev set for model selection.

span-level and can cross sentence boundaries. We convert span annotations to sentence-level annotations as follows: if a span annotated with label l lies within sentence boundaries and covers at least 50% of the tokens in the sentence, we label the sentence with l. We only keep sentence annotations if they are indicated by at least two annotators.

Congressional Tweets Dataset The congressional tweets dataset (Johnson et al., 2017) contains tweets authored by 40 members of the US Congress, annotated with the frames of the Policy Frames Codebook. The tweets are related to one or two of the following six issues: *abortion*, *the Affordable Care Act*, *gun rights vs. gun control*, *immigration*, *terrorism*, and *the LGBTQ community*, where each tweet is annotated with one or multiple frames.

Argument Quality Annotations The corpus of online discussions contains additional annotations that we exploit in the multi-task setup. Swanson et al. (2015) sampled a subset of 5,374 sentences, using various filtering methods to increase likelihood of high quality argument occurrence, and collected annotations for argument quality via crowdsourcing. Annotators were asked to rate argument quality using a continuous slider [0-1]. Seven annotations per sentence were collected. We convert these annotations into binary labels (1 if ≥ 0.5, 0, otherwise) and generate an approximately balanced dataset for a binary classification task that is then used as an auxiliary task in the multi-task setup. Balancing is motivated by the observation that balanced datasets tend to be better auxiliary tasks (Bingel and Søgaard, 2017).

4 Models

The task we are faced with is (multi-label) sequence classification for online discussions. However, we have no labeled training data (and only a small labeled validation set) for the target task in the target domain. Hence, we train our model on a dataset which is labeled with the target labels, but from a different domain. The largest such dataset is the news articles corpus, which we consequently use as main task. Our baseline model is a two-layer LSTM (Hochreiter and Schmidhuber) trained on only the news articles data. We then apply two strategies to facilitate the transfer of information from source to target domain, multi-task learning and adversarial learning. We briefly describe both setups in the following. An overview over tasks and data used in the different models is shown in Table 3.

Multi-Task Learning To exploit synergies between additional datasets/annotations, we explore a simple multi-task learning with hard parameter sharing strategy, pioneered by Caruana (1993), introduced in the context of NLP by Collobert et al. (2011), and to RNNs by Søgaard and Goldberg (2016), which has been shown to be useful for a variety of NLP tasks, e.g. sequence labelling (Rei, 2017; Ruder et al., 2019; Augenstein and Søgaard, 2017), pairwise sequence classification (Augenstein et al., 2018) or machine translation (Dong et al., 2015). Here, parameters are shared between hidden layers. Intuitively, it works by training several networks in parallel, tying a subset of the hidden parameters so that updates in one network affect the parameters of the others. By sharing parameters, the networks regularize each other, and the network for one task can benefit from repre-
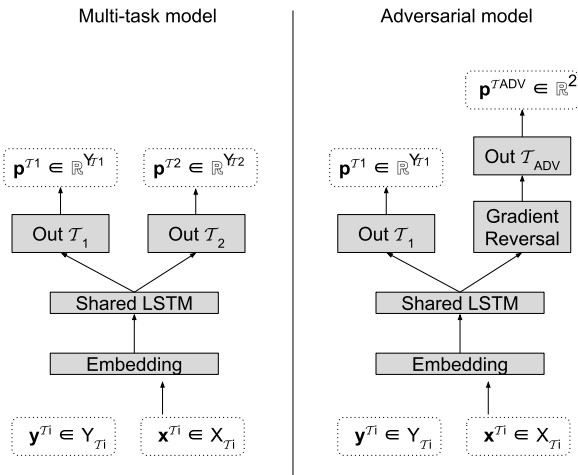
Figure 1: Overview over the multi-task model (left) and the adversarial model (right). The baseline LSTM model corresponds to the same architecture with only one task.
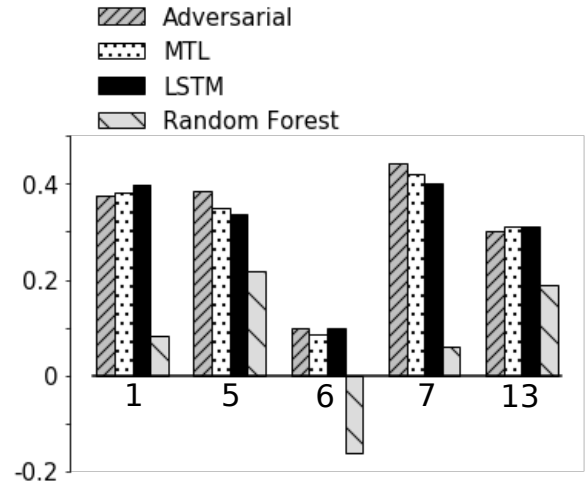

Figure 2: Improvement in F-score over the random baseline by class. The absolute F-scores for the best performing system for classes 1, 5, 6, 7, and 13, are 0.529, 0.625, 0.298, 0.655, and 0.499, respectively.

sentations induced for the others.

Our multi-task architecture is shown in Figure 1. We have N different datasets $\mathcal{T}_1, \cdots, \mathcal{T}_N$. Each dataset $\mathcal{T}_i$ consists of tuples of sequences $x^{\mathcal{T}_i} \in X_{\mathcal{T}_i}$ and labels $y^{\mathcal{T}_i} \in Y_{\mathcal{T}_i}$. A model for task $\mathcal{T}_i$ consists of an input layer, an LSTM layer (that is shared with all other tasks) and a feed forward layer with a softmax activation as output layer. The input layer embeds a sequence $x^{\mathcal{T}_i}$ using pretrained word embeddings. The LSTM layer recurrently processes the embedded sequence and outputs the final hidden state h. The output layer outputs a vector of probabilities $p^{\mathcal{T}_i} \in R^{Y_{\mathcal{T}_i}}$, based on which the loss $\mathcal{L}_i$ is computed as the categorical cross-entropy between prediction $p^{\mathcal{T}_i}$ and true label $y^{\mathcal{T}_i}$. In each iteration, we sample a data batch for one of the tasks and update the model parameters using stochastic gradient descent. If we sample a batch from the main task or an auxiliary task is decided by a weighted coin flip.

Adversarial Learning Ganin and Lempitsky (2015) proposed adversarial learning for domain adaptation that can exploit unlabeled data from the target domain. The idea is to learn a classifier that is as good as possible at assigning the target labels (learned on the source domain), but as poor as possible in discriminating between instances of the source domain and the target domain. With this strategy, the classifier learns representations that contain information about the target class but abstract away from domain-specific features. During training, the model alternates between 1) pre-

dicting the target labels and 2) predicting a binary label discriminating between source and target instances. In this second step, the gradient that is backpropagated is flipped by a Gradient-Reversal layer.[5] Consequently, the model parameters are updated such that the classifier becomes worse at solving the task. The architecture is shown in the right part of Figure 1. In our implementation, the model samples batches from the adversarial task or the main task based on a weighted coinflip.

5 Experiments

We compare the multi-task learning and the adversarial setup with two baseline models: (a) a Random Forest classifier using tf-idf weighted bag-of-words-representations, and (b) the LSTM baseline model. For the multi-task model, we use both the Twitter dataset and the argument quality dataset as auxiliary tasks. For all models, we report results on the test set using the optimal hyper-parameters that we found averaged over 3 runs on the validation set. For the neural models, we use 100-dimensional GloVe embeddings (Pennington et al., 2014), pre-trained on Wikipedia and Gigaword.[6] Details about hyper-parameter tuning and optimal settings can be found in Appendix B.

Results The results in Table 5 show that both the multi-task and the adversarial model improve over

---

[5]In the forward pass, this layer multiplies its input with the identity matrix.

[6]https://nlp.stanford.edu/projects/glove/

Nr.	Gold	Adv	MTL	LSTM	Sentence
(1)	5	5	5	7	But, star gazer, we had guns then when the Constitution was written and enshrined in the BOR and now incorporated into th 14th Civil Rights Amendment.
(2)	6	6	5	1	Gun control is about preventing such security risks.
(3)	7	7	5	1	First, you warn me of the dangers of using violent means to stop a crime .
(4)	5	6	6	6	So I don't see restrictions on handguns in D.C. as being a clear violation of the Second Amendment.

Table 4: Examples for model predictions on the online discussion dev set. The first column shows the gold label and the following columns the prediction made by the adversarial model (Adv), the Multi-Task model (MTL) and the LSTM baseline (LSTM).

Model	P_{ma}	R_{ma}	F_{ma}	F_{mi}
Random Baseline	0.196	0.198	0.189	0.196
Random Forest Baseline	0.496	0.335	0.267	0.279
LSTM Baseline	0.512	0.510	0.503	0.521
Multi-Task	0.526	0.525	0.505	0.534
Adversarial	**0.533**	**0.534**	**0.515**	**0.548**

Table 5: Macro- (ma) and micro-averaged (mi) scores for the online discussion test data averaged over 3 runs. The multi-task model uses the Twitter and argument quality datasets as auxiliary tasks. The micro-average F of a baseline that predicts the majority class is 0.307.

the baselines. The multi-task model achieves minor improvements over the LSTM baseline, with a bigger improvement in the micro-averaged score, indicating bigger improvements with frequent labels. The adversarial model performs best, with an error reduction in micro-averaged F over the LSTM baseline of 5.6%.

Figure 2 shows the system performances for each class. Each bar indicates the difference between the F-score of the respective system and the random baseline. The adversarial model achieves the biggest improvements over the baseline for classes 5 and 7, which are the two most frequent classes in the test set (cf. Table 6). For classes 1 and 13, the adversarial model is outperformed by the LSTM. Furthermore, we see that the hardest frame to predict is the *Policy prescription and evaluation frame* (6), where the models achieve the lowest improvement over the baseline and the lowest absolute F-score. This might be because utterances with this frame tend to address specific policies that vary according to topic and domain of the data, and are thus hard to generalize from source to target domain.

Analysis Table 4 contains examples of model predictions on the dialogue dev set. In Exam-

ple (1), the adversarial and the multi-task model correctly predict a *Constitutionality* frame, while the LSTM model incorrectly predicts a *Crime and punishment* frame. In Examples (2) and (3), only the adversarial model predicts the correct frames. In both cases, the LSTM model incorrectly predicts an *Economic* frame, possibly because it is misled by picking up on a different sense of the terms *means* and *risks*. In Example (4), all models make an incorrect prediction. We speculate this might be because the models pick up on the phrase *restrictions on handguns* and interpret it as referring to a policy, whereas to correctly label the sentence they would have to pick up on the *violation of the Second Amendment*, indicating a *Constitutionality* frame.

6   Conclusion

This work introduced a new benchmark of political discussions from online fora, annotated with issue frames following the Policy Frames Cookbook. Online fora are influential platforms that can have impact on public opinion, but the language used in such fora is very different from newswire and other social media. We showed, however, how multi-task and adversarial learning can facilitate transfer learning from such domains, leveraging previously annotated resources to improve predictions on informal, multi-party discussions. Our best model obtained a micro-averaged F1-score of 0.548 on our new benchmark.

References

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance Detection with Bidirectional Conditional Encoding. In *Proceedings of EMNLP*.

Isabelle Augenstein, Sebastian Ruder, and Anders Søgaard. 2018. Multi-Task Learning of Pairwise Sequence Classification Tasks over Disparate Label Spaces. In *NAACL-HLT*, pages 1896–1906. Association for Computational Linguistics.

Isabelle Augenstein and Anders Søgaard. 2017. Multi-Task Learning of Keyphrase Boundary Classification. In *Proceedings of ACL*.

Eric Baumer, Elisha Elovic, Ying Qin, Francesca Polletta, and Geri Gay. 2015. Testing and Comparing Computational Approaches for Identifying the Language of Framing in Political News. In *Proceedings of HLT-NAACL*, pages 1472–1482. The Association for Computational Linguistics.

Joachim Bingel and Anders Søgaard. 2017. Identifying beneficial task relations for multi-task learning in deep neural networks. In *Proceedings of EACL*.

Amber E. Boydstun, Dallas Card, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2014. Tracking the Development of Media Frames within and across Policy Issues. In *Proceedings of APSA*.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The Media Frames Corpus: Annotations of Frames Across Issues. In *Proceedings of ACL*, pages 438–444.

Dallas Card, Justin Gross, Amber Boydstun, and Noah A Smith. 2016. Analyzing Framing through the Casts of Characters in the News. In *Proceedings of EMNLP*, pages 1410–1420.

Richard Caruana. 1993. Multitask Learning: A Knowledge-Based Source of Inductive Bias. In *Proceedings of ICML*, pages 41–48. Morgan Kaufmann.

Dennis Chong and James Druckman. 2007. Framing Theory. *Annual Review of Political Science*, 10.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural Language Processing (Almost) from Scratch. *JMLR*, 999888:2493–2537.

Frank E. Dardis, Frank R. Baumgartner, Amber E. Boydstun, Suzanna de Boef, and Fuyuan Shen. 2008. Media Framing of Capital Punishment and Its Impact on Individuals' Cognitive Responses. *Mass Communication and Society*, 11(2):115–140.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-Task Learning for Multiple Language Translation. In *Proceedings of ACL*.

Robert M. Entman. 1993. Framing: Toward Clarification of a Fractured Paradigm. *Journal of Communication*, 43(4):51–58.

William Ferreira and Andreas Vlachos. 2016. Emergent: A Novel Data-Set for Stance Classification. In *Proceedings of NAACL HLT*.

Anjalie Field, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. 2018. Framing and Agenda-setting in Russian News: a Computational Analysis of Intricate Political Strategies. In *Proceedings of EMNLP*, pages 3570–3580. Association for Computational Linguistics.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised Domain Adaptation by Backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1180–1189, Lille, France. PMLR.

Sepp Hochreiter and Jürgen Schmidhuber. Long Short-term Memory. *Neural Computation*.

Shanto Iyengar. 1991. *Is Anyone Responsible? How Television Frames Political Issues*. University of Chicago Press.

Yangfeng Ji and Noah Smith. 2017. Neural Discourse Structure for Text Categorization. In *Proceedings of ACL*.

Kristen Johnson, Di Jin, and Dan Goldwasser. 2017. Leveraging Behavioral and Social Information for Weakly Supervised Collective Classification of Political Discourse on Twitter. In *Proceedings of ACL*.

Chenghua Lin, Yulan He, and Richard Everson. 2011. Sentence Subjectivity Detection with Weakly-Supervised Learning. In *Proceedings of IJCNLP*, pages 1153–1161, Chiang Mai, Thailand.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 Task 6: Detecting Stance in Tweets. In *Proceedings of SemEval*.

Nona Naderi and Graeme Hirst. 2017. Classifying Frames at the Sentence Level in News Articles. In *Proceedings of RANLP*, pages 536–542.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of EMNLP*.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Daniel Jurafsky. 2013. Linguistic Models for Analyzing and Detecting Biased Language. In *Proceedings of ACL*.

Marek Rei. 2017. Semi-supervised Multitask Learning for Sequence Labeling. In *Proceedings of ACL*.

Sebastian Ruder, Joachim Bingel, Isabelle Augenstein, and Anders Søgaard. 2019. Multi-Task Architecture Learning. In *AAAI*.

Anders Søgaard and Yoav Goldberg. 2016. Deep multi-task learning with low level tasks supervised at lower layers. In *Proceedings of ACL*.

Reid Swanson, Brian Ecker, and Marilyn A. Walker. 2015. Argument Mining: Extracting Arguments from Online Dialogue. In *SIGDIAL Conference*.

Oren Tsur, Dan Calacci, and David Lazer. 2015. A Frame of Mind: Using Statistical Models for Detection of Framing and Agenda Setting Campaigns. In *Proceedings of ACL-IJCNLP*, pages 1629–1638. Association for Computational Linguistics.

Marilyn A. Walker, Jean E. Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *LREC*, pages 812–817. European Language Resources Association (ELRA).

Appendix A   Data Preprocessing

For the Twitter and news articles datasets, we remove all instances that do not correspond to the five target frames. Table 6 shows the class distributions in the filtered datasets. We tokenize all sequences using spaCy [7], which we also use for sentence splitting in the news articles dataset. For the Twitter dataset, we follow Johnson et al. (2017) in removing URLs and @-mentions.

Appendix B   Hyperparameters in Experiments

The hyperparameters for all neural models were tuned on the online disc. dev set. We report test results for the optimal settings found by averaging over 3 training runs, which we determine by the best macro-averaged F-score and smallest variance between the runs. We set the DyNet weight

---

[7] https://spacy.io/

Dataset	# instances	# instances per class					# multi
		1	13	5	6	7	
NEWSPAPER (TRAIN)	10,480	1088	1959	2023	924	890	45
TWITTER (TRAIN)	1,636	73	300	137	27	174	554
ONLINE DISC. (TEST)	692	78	96	234	166	186	67
		0	1				
ARGUMENT QUALITY	3,785	1,350	2,435				0
ONLINE DISC. unlabeled	4731						

Table 6: Dataset statistics and class distributions. The frame labels correspond to the classes *Economic* (`1`), *Political* (`13`), *Legality, Jurisprudence and Constitutionality* (`5`), *Policy prescription and evaluation* (`6`) and *Crime and Punishment* (`7`). *# multi* refers to the number of multi-label instances. For Argument quality, label 1 indicates a score greater or equal 0.5.

decay parameter to 1e-7 for all neural models, batch size is 128, and the word embeddings are not updated during training.

For the multi-task and adversarial model, we do a grid-search over the weight of the coin flip used to decide on sampling from main/aux or main/adversarial task in the range of [0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9]. The optimal weight for sampling the main task is 0.5 for the multi-task model and 0.3 for the adversarial task.

All models are trained using early stopping (after at least 80 epochs of training) with a patience of 5 epochs. The number of iterations (updates) per epoch is a hyperparameter, that we set by default as twice the number of data batches for the main task. For a fair coin flip, the models hence see as much data for the main task as for the auxiliary/adversarial task per epoch.