Spurious Correlations in Cross-Topic Argument Mining

Terne Sasha Thorn Jakobsen[1] **Maria Barrett**[2] **Anders Søgaard** [3]

[1]Copenhagen Center for Social Data Science,
University of Copenhagen terne.thorn@sodas.ku.dk
[2]Computer Science Department, IT University of Copenhagen mbarrett@itu.dk
[3]Department of Computer Science, University of Copenhagen soegaard@di.ku.dk

Abstract

Recent work in *cross-topic* argument mining attempts to learn models that generalise across topics rather than merely relying on within-topic spurious correlations. We examine the effectiveness of this approach by analysing the output of single-task and multi-task models for cross-topic argument mining through a combination of linear approximations of their decision boundaries, manual feature grouping, challenge examples, and ablations across the input vocabulary. Surprisingly, we show that cross-topic models *still* rely mostly on spurious correlations and only generalise within closely related topics, e.g., a model trained only on closed-class words and a few common open-class words outperforms a state-of-the-art cross-topic model on distant target topics.

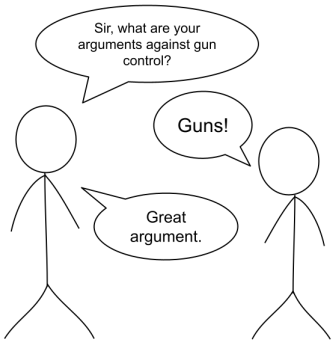


Figure 1: In human interaction, it is evident that relying on topic words for recognizing an argument is nonsensical. It is, nevertheless, what a BERT-based cross-topic argument mining model does.

1 Introduction

When a sentiment analysis model associates the word *Shrek* with positive sentiment (Sindhwani and Melville, 2008), it relies on a spurious correlation. While the movie Shrek was popular at the time the training data was sampled, this is unlikely to transfer across demographics, platforms and years. While there exists a continuum from sentiment words such as *fantastic* to spurious correlations such as *Shrek*, with words such as *Hollywood* or *anticipation* being perhaps in a grey zone, demoting spurious correlations is key to learning robust NLP models (Sutton et al., 2006; Søgaard, 2013; Tu et al., 2020).

This paper studies a similar problem in state-of-the-art cross-topic argument mining systems. The task of argument mining is to recognise the existence of claims and premises in a text span. The standard evaluation protocol is to evaluate argument mining systems *across topics*, i.e., on held-out topics, precisely to avoid over-fitting to a single topic (Daxenberger et al., 2017; Stab et al., 2018; Reimers et al., 2019). This study shows that despite this sensible cross-topic evaluation protocol, state-of-the-art systems nevertheless rely primarily on spurious correlations, e.g., *guns* (Figure 1). These spurious correlations transfer across some topics in popular benchmarks, but only because the topics are closely related.

Contributions We present experiments with an out-of-the-box learning architecture for argument mining, yet with state-of-the-art performance, based on Microsoft's MT-DNN library (Liu et al., 2019). We train models on the UKP Sentential Argument Mining Corpus (Stab et al., 2018), the IBM Debater Argument Search Engine Dataset (Levy et al., 2018), the Argument Extraction corpus (Swanson et al., 2015), and the Vaccination Corpus (Morante et al., 2020). We analyse the models with

All code will be publicly available at https://github.com/terne/spurious_correlations_in_argmin

Proceedings of the 10th Conference on Lexical and Computational Semantics, pages 263–277
August 5–6, 2021, Bangkok, Thailand (online)

respect to spurious correlations using the post-hoc interpretability tool LIME (Ribeiro et al., 2016) and we find that the models rely heavily on these. This analysis is the paper's main contribution: In §5, we: a) evaluate our best-performing model on a small set of challenge examples, which we make available, and which motivate our subsequent analyses; b) manually analyse how many of the words our models rely the most on are spurious correlations; c) evaluate how much weight our models attribute to open class words and whether multi-task training effectively moves emphasis to closed-class items that likely transfer better across topics; d) evaluate how much weight our models attribute to words in a manually constructed claim indicator list (Morante et al., 2020; Stab and Gurevych, 2017), and whether multi-task training effectively moves emphasis to such claim indicators that likely transfer better across topics; and lastly e) evaluate the performance of models trained only on closed-class words or closed class and open class words that are shared across topics. Surprisingly, we find that models with access to only closed-class words, and a few common (topic-independent) open-class words, perform better across distant topics than our baseline, state-of-the-art models (Table 5).

2 Argument mining

We first describe the task of argument mining, focusing, in particular, on the subtle difference between argument mining ('this is an argument for or against x') and stance detection ('this is an expression of opinion for or against x'). Both tasks are very relevant for social scientists, monitoring the dynamics of public opinion. Still, whereas stance detection can be used to see what fractions of demographic subgroups are in favor of or against some topic, argument mining can be used to identify the arguments made for and against policies in political discussions.

What is an argument? An argument is made up of propositions (claims), which are statements that are either true or false. Traditionally, an argument must consist of at least two claims, with one being the conclusion (major claim) and at least one reason (premise) backing up that claim. Some argument annotation schemes ask annotators to label premises and major claims separately (Lindahl et al., 2019). Others simplify the task to identifying claim or claim-like sentences (Morante et al., 2020) or to whether sentences are claims supporting or opposing a particular idea or topic (Levy et al., 2018; Stab et al., 2018). The resources used in our experiments below are of the latter type: Sentences are labeled as arguments if they present evidence or reasoning in relation to a claim or topic and are refutable.

The resources used in our experiments are annotated with arguments in the context of a particular topic, as well as the argument's polarity, i.e., what is annotated relates to *stance*. The key difference between the current task and stance detection is that arguments require the author to present evidence or reasoning for or against the topic.

Spurious correlations of arguments Arguments for or against a policy typically refer to different concepts. Take, for example, discussions of minimum wage and the terms *living wages* and *jobs*. Since these terms are frequent in arguments for and against minimum wage, they will be predictive of arguments (in discussions of minimum wage). Still, mentions of the terms are not themselves markers of arguments, but simply *spurious correlations* of arguments. We use the same definition of spurious correlations as Wang and Culotta (2020), mainly that a relationship between a term and a label is spurious if one cannot expect the term to be a *determining factor* for assigning the label.[1] Examples of the contrary are terms such as *if* and *because* (and to some degree stance terms), which one can reasonably expect to be determining factors for an argument to exist (and therefore to be stable across topics and time).

3 Datasets

The UKP Sentential Argument Mining Corpus (UKP) (Stab et al., 2018) contains 25,492 sentences spanning eight controversial topics (abortion, cloning, death penalty, gun control, marijuana legalization, school uniforms, minimum wage and nuclear energy), each annotated at the sentence level as one of three classes; NO ARGUMENT, ARGUMENT AGAINST, and ARGUMENT FOR. For example, a sentence about *death penalty* may not be arguing for or against death penalty (NO ARGUMENT), may present an argument against having death penalty as a punishment for a severe crime

[1] Arjovsky et al. (2019) provides the example of a classifier trained to distinguish between images of cows and camels; if prone to spurious correlations, the classifier may be challenged by a picture of a cow on a sandy beach. Bommasani and Cardie (2020) also refer to spurious correlations as *reasoning shortcuts.*

(ARGUMENT AGAINST), or may present an argument in favor of the same (ARGUMENT FOR). The data is annotated such that the evaluation of a sentence (being an argument or not) is not strictly dependent on the topic. However, it should still be unambiguously supportive of or against a topic. Claims will not be annotated as an argument unless they include some evidence or reasoning behind the claim; however, Lin et al. (2019) do find a few wrongly annotated sentences in this regard. The corpus comes with a fixed 70-10-20 split.

The IBM Debater Argument Search Engine Dataset (IBM) is from a larger dataset of argumentative sentences defined through query patterns by Levy et al. (2017, 2018). We use only the 2,500 sentences that are gold labelled — with binary labels, where positive labels were given to statements that directly support or contest a topic. The sentences are from Wikipedia articles and span 50 topics. Since the authors used queries to mine the examples, the data is imbalanced (70% positive). We introduce a random 70-30 split.

The Argument Extraction Corpus (AQ) (Swanson et al., 2015) contains 5,374 sentences annotated with argument quality on a continuous scale between 0 (hard to interpret the argument) and 1 (easy to interpret the argument). Of the corpora included in our study, this differs most from the others; however, the topics included are controversial topics (gun control, gay marriage, evolution, and death penalty), similar to the UKP Corpus. The sentences are partly from the Internet Argument Corpus (Walker et al., 2012) and partly from createdebate.com. We introduce a random 70-30 split.

The Vaccination Corpus (VacC) was presented in Morante et al. (2020) and consists of 294 documents from online debates on vaccination with marked claims. A claim is defined as opinionated statements wrt. vaccination. For our purpose, we split the documents into sentences (23,467). We use binary labels (claim or not) and introduce a random 70-10-20 split.

4 Experimental setup

We now describe our learning architecture, an almost out-of-the-box application of the MT-DNN architecture in Liu et al. (2019). It is a strong model that achieves a better performance than previously reported across the benchmarks.

The MT-DNN model of Liu et al. (2019) combines the pre-trained BERT architecture with multi-task learning. The model can be broken up into *shared layers* and *task-specific layers*. The shared layers are initialised with the pre-trained BERT base model (Devlin et al., 2019). We add a task-specific output layer for each task and update all model parameters during training with AdaMax. The task-specific layers are logistic regression classifiers with softmax activation, minimising cross-entropy loss functions for classification tasks or mean squared error for regression tasks. If we only have a single output layer, we refer to the architecture as single-task DNN (ST-DNN) rather than MT-DNN. We train all models over 10 epochs with a batch size of 5 for feasibility and otherwise use default hyperparameters.

Following Stab et al. (2018), we iteratively combine the training and validation data from seven of the eight topics of the UKP Corpus for training and parameter tuning and use the test data of the held-out topic for testing. We firstly treat the task as a single-sentence classification task and train an ST-DNN with the BERT-base model as shared layers. Since Tu et al. (2020) argues multi-task learning effectively reduces sensitivity to spurious correlations, we experiment with MT-DNN models based on different data and task combinations: For each auxiliary dataset (IBM, AQ, and VAcC), we train an MT-DNN model with the UKP Corpus as one task and the auxiliary data as another task. We denote the MT-DNN models as follows: **MT-DNN+IBM** refers to a model trained with the IBM data as an auxiliary claim classification task; **MT-DNN+AQ** is trained with AQ as an auxiliary regression task; **MT-DNN+VacC** is trained with VAcC data as an auxiliary claim classification task; **MT-DNN+AQ+IBM+VacC** is our largest model trained with all auxiliary tasks. **Topic-MT-DNN** provides us with an upper bound: In this setting, all topics are used in training and tuning, including the target topic, as eight separate tasks.

5 Analysis

We evaluate the models on the UKP Corpus using the cross-topic evaluation protocol of (Stab et al., 2018) – training with seven topics and testing on a held-out topic. We report the average macro F_1 across five random seeds. Table 1 shows the average cross-topic results as well as results for each held-out topic for all models. With single-task mod-

Model	Average	abortion	cloning	death penalty	gun control	marijuana legal	school uniforms	minimum wage	nuclear energy
IN-TOPIC MODELS *(upper bounds)*									
Topic-MT-DNN†	.665	.571	.733	.595	.611	.724	.707	.716	.662
CROSS-TOPIC MODELS									
ST-DNN	.642±.011	.473±.012	.715±.012	.595±.009	.593±.011	.703±.010	.698±.015	.710±.013	.650±.002
MT-DNN+IBM	.643±.009	.466±.019	**.726**±.010	.595±.006	.582±.004	.704±.010	.703±.010	**.718**±.009	.655±.006
MT-DNN+AQ	.643±.011	**.479**±.015	.716±.006	**.600**±.012	.590±.010	.699±.011	**.710**±.010	.698±.008	.649±.015
MT-DNN+VacC	.641±.010	.472±.016	.716±.008	.589±.009	**.601**±.009	.701±.011	.690±.010	.699±.013	**.660**±.006
MT-DNN+VacC+IBM+AQ	**.644**±.011	.476±.009	.720±.021	.587±.011	.598±.005	**.716**±.011	.696±.003	.701±.018	.655±.006
CONSTRAINED CROSS-TOPIC MODELS *(lower bounds)*									
CLOSED	.481±.014	.472±.016	.492±.006	.467±.013	.452±.015	.515±.021	.478±.012	.520±.012	.519±.008
CLOSED+SHARED	.501±.010	.426±.012	.508±.016	.475±.009	.469±.006	.552±.004	.490±.005	.565±.017	.519±.008

Table 1: Macro F_1 scores across topics of the three-class UKP data. IN-TOPIC models are (also) trained on the training data of the target topic. CONSTRAINED models only rely on closed-class words and open class words shared across *all* topics. In-topic, cross-topic and constrained models cannot be directly compared. Still, in-topic and constrained models provide upper and lower bounds in the sense that they represent scenarios where models are encouraged, respectively prohibited, to rely on spurious features. We report averages across 5 random seeds except †, which is only one run. The best performances per column within cross-topic models are boldfaced.

els, we achieve an average macro F_1 of .642, which is a big improvement from the .429 reported by Stab et al. (2018). Our ST-DNN model also outperforms the best-reported score in the literature, which, as far as we know, is .633 by Reimers et al. (2019). Reimers et al. (2019) used BERT Large and, unlike us, integrated topic information in the model. Multi-task learning can improve the performance to .644, a 35% error reduction relative to the upper bound of training a model on all eight topics, i.e., including in-topic training data. We see a large variation in the performance across topics for all models, with the abortion topic being hardest to classify and cloning being easiest. With two classes – argument or not – the average macro F_1 is .776, again with large differences across topics; abortion being hardest to classify (.656) and minimum wage being easiest (.828). To analyze our models, we use the popular post-hoc interpretability tool LIME (Ribeiro et al., 2016). By training linear (logistic regression) models on perturbations of each instance, LIME learn interpretable models that locally approximate our models' decision boundaries. The weights of the LIME models tell us which features are locally important.[2]

a) Challenge examples For an initial qualitative error analysis, 19 short text pieces are taken from exercises made by Jon M. Young for his Critical Thinking course at Fayetteville State University.[3][4] Of these, the first six are examples of sentences that comprise an argument or not, and if they do, the conclusions and premises have been annotated by Young. The last 13 examples are from exercises where we annotated the correct answers. We contrast the LIME analyses of the predictions of our best performing model, i.e. MT-DNN+VacC+IBM+AQ, as well as our ST-DNN baseline.[5] An example of the LIME explanations can be seen in Figure 2. The remaining LIME explanations are in the appendix in Figures 4-7.

Out of the 19 examples, seven were incorrectly classified by our best model. Common to these misclassified examples is either a rather uncontroversial, everyday topic (4c, 4g, 5e) or a very informative language (4h, 5g, 5h). Since the model was mainly trained on controversial topics, it is not surprising that these uncontroversial cases make the model misstep. While this is a tiny sample, these incorrect classifications do suggest that our models do not transfer well to *any* topic, possibly indicating they rely more on topic words than on

[2]LIME has several weaknesses: LIME is linear (Bramhall et al., 2020), unstable (Elshawi et al., 2019) and very sensitive to the width of the kernel used to assign weights to input example perturbations (Vlassopoulos, 2019; Kopper, 2019), an increasing number of features also increases weight instability (Gruber, 2019), and Vlassopoulos (2019) argues that with sparse data, sampling is insufficient. Laugel et al. (2018) argues the specific sampling technique is suboptimal. Since we use aggregate LIME statistics across hundreds of data points, these weaknesses should have limited impact on our results; LIME remains a *de facto* standard, and most alternatives suffer from similar weaknesses or are prohibitively costly to run.

[3]https://tinyurl.com/y6ldjtvh

[4]https://tinyurl.com/yyw5uhtm

[5]For LIME, we use a neighbourhood of size 500 both here and in the following experiments. We use models trained with random seed 2018 for the current and following LIME experiments, and for the current analysis, we use models trained with the cloning topic as our held-out topic.

Topic	Argument words	Topic words	Stance words	Other
abortion	if, that, for	abortion, life, women, woman, human, pregnancy, unborn	right, legal, hates	the , is, to, in, it, be
cloning	would, will, if, could, potential	cloning, clone, cloned, genetic	not, no, abnormalities	the , to, is, it, have, be, do
death penalty	would, if	death, penalty, punishment, killing, crime	not, murder, murderers	the , to, in, is, of, are, people, it
gun control		gun, guns, criminals, background, checks, disarm, arms, armed	no, safer, right, more, not, abiding	a , the, are, and, in, is
marijuana legalization	would	marijuana, use, effects, legalizing, legalization, drug, prohibition, drugs	no, not, more, abuse, costs	is , the, are, it
school uniforms	if, but	uniforms, uniform, school, students, clothing, wears	not, less, improve, decreased, uncomfortable, costs	to , can, it, without
minimum wage	would, that, if	wage, minimum, workers, wages, living, jobs, hour	cost, more, no, many	the , it, is, are, can
nuclear energy	that, if, for	nuclear, power, energy, reactors, plants, waste, chernobyl, fuel	safety, less	is , the, to, has, can, it

Table 2: Top 20 words for each topic based on accumulated LIME weights towards the predicted label of each sentence. Divided into word categories.

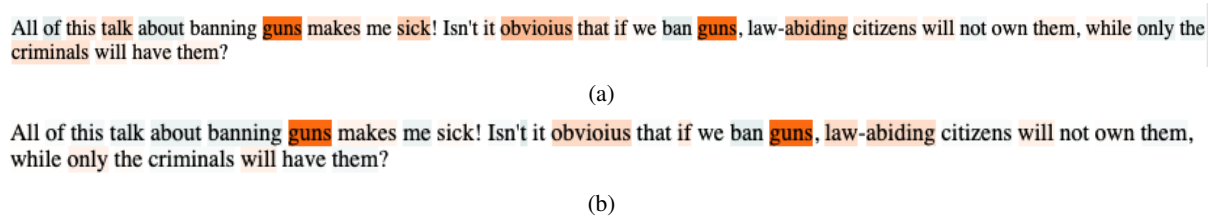


Figure 2: Non-argumentative example sentence (because it is question rather than argument) explained with LIME. The orange highlights indicate words weighted positively towards the ARGUMENT AGAINST class. The darker the colour, the larger the weight. a) using MT-DNN+AQ+IBM+VacC as the predictor. b) using ST-DNN as the predictor. Both models used were trained with the cloning topic held out.

argument markers. This is supported by the observation that open-class words – rather than argumentative language patterns – are given most of the weight towards the argument classes. Open-class words are defined as nouns, verbs and adjectives, and closed-class words are the remains. For example, we see "guns" as an argument indicator rather than "if" in 2a and 2b; we see "people" and "needs" emphasized more than "if" in 5f; and in 5i, the stance indicator "disastrous" and the open-class word "television" have large weights, while "seems" and "caused" are not emphasized at all. Overall, this suggests our models learn what *arguments are about* but not what *constitutes an argument*. The single-task model exhibits similar patterns. In fact, there seems to be little difference between what the two models attend to.

This initial evaluation raises two questions: To what extent do our models rely on topic-specific spurious correlations with limited ability to transfer across (distant) topics instead of relying on more generic argument markers? And to what extent do simple regularization techniques like multi-task learning, as suggested in Tu et al. (2020), prevent our models from over-fitting in this way?

b) How many of the words we rely on are spurious? We generate and accumulate LIME explanations for our single-task models over the corresponding held-out topics' development sets to evaluate how much our models rely on spurious correlations. We accumulate LIME weights for words towards the predicted class. Words are sorted by accumulated weights, and we manually annotate the top k words for whether they are spurious.

Specifically, and to better understand the distribution of word types, we divide the top 20 words into four categories: *argument words*, *topic words*, *stance words*, and *other*. We define argument words as words that likely appear when present-

ing claims, independent on the topic, including markers of evidence and reasons such as "if", "that" and "because" and similar lexical indicators based on (Stab and Gurevych, 2017). Contrary to argument words, we define topic words as words that have no relation to the act of presenting an argument but are clearly related to the specific topic, e.g., nouns or verbs frequently used when debating or merely describing the topic. Lastly, we define stance words as opinionated words that express a stance toward a topic (but is not only used in the context of arguments, i.e., presenting evidence). Examples include describing death penalty as "murder" or school uniforms as "uncomfortable". Three annotators agreed on the classification. Words that did not fit our scheme were categorised as *other*. Table 2 shows the top 20 words, categorised, for all development sets.[6]

Our first observation is that 62.5% of the top 20 words are topic words, and for the GUN CONTROL topic, none of the words are argument words. Instead, topic words such as "criminals", "background" and "checks" receive high weights. These words are neither indicative of an argument or stance – hence, they are spurious correlations. Interestingly, the only topic where *argument words* is the majority category is cloning – the held-out topic where all our models perform best. This suggests reducing our models' reliance on topic words can improve the cross-topic performance of argument mining models, which we will investigate in the following experiments. Of course, our models, nevertheless, show relatively good performance across topics, suggesting that some topic words transfer across topics in the UKP corpus. We will discuss recommendations for experimental protocols and the importance of evaluating across *distant* topics below.

Note that we do not normalize the accumulated LIME weights by word frequency, which favors frequent words. When normalising the weights, our models also rely heavily on low-frequency stance words and for all topics, except cloning, there are many topic words among the top 20. High-frequency words (as well as most argument words) are naturally ranked much lower after normalisation. Stance words are, of course, not spurious for our three-way classification problem, but a near disappearance of argument words in the normalized top 20 suggests our models are unlikely to capture low-frequency argument markers.

c) How much weight do our models attribute to open class words, and does multi-task learning move emphasis to closed-class items? Multi-task learning is a regularization technique (Søgaard and Goldberg, 2016; Liu et al., 2019) and may, as suggested by Tu et al. (2020), reduce the extent to which our models rely on spurious correlations, which tend to be open class words. To compare the weight attributed to open-class words, across single-task and multi-task models, we define a score reflecting the weight put on open class words in a sentence: For each word in the sentence, we consider the maximum LIME weight of the two weights towards the argument classes ARGUMENT AGAINST and ARGUMENT FOR. We then take the sum of LIME weights put on open class words, normalised by the total sum of weights, and divide the normalised weight by the sentence fraction of open-class words. Table 3 shows the average sentence scores for each topic and model. We observe that the weights are very similar across single-task and multi-task models (and topics), and a Wilcoxon signed-rank test confirms that there is no significant difference between single-task and multi-task open class sentence scores. We also performed the test with sentence scores defined for each class separately (rather than taking the maximum weight) and again found no significant differences.

Topic	ST	MT
abortion	1.447	1.408
cloning	1.404	1.416
death penalty	1.441	1.421
gun control	1.436	1.381
marijuana legalization	1.387	1.414
school uniforms	1.461	1.402
minmum wage	1.398	1.412
nuclear energy	1.379	1.366
mean	1.419	1.402

Table 3: The sentence scores reflecting the weight put on open class words across domains and model types. There is no significant difference between mean sentence scores of ST and MT models.

d) How much weight do our models attribute to claim indicators, and does multi-task learning move emphasis to such indicators? As a set of

[6]Top 20 words along with their frequency and LIME weights are provided at github.com/terne/spurious_correlations_in_argmin/top_words

Claim indicators	*indicates, because, proves, however, shows, result, opinion, conclusion, given, accordingly, since, clearly, mean, truth, consequently, must, would, points, therefore, whereas, obvious, demonstrates, thus, fact, if, that, hence, i, could, should, for, contrary, potential, may, believe, suggests, probable, conclude, clear, point, sum, entails, think, implies, explanation, follows, reason*
Shared open	*political, single, debate, had, asked, made, policy, last, legal, cause, long, few, said, want, person, issue, say, group, possible, use, people, believe, good, have, fact, point, society, time, such, going, put, used, come, based, question, think, example, part, other, are, year, including, argument, only, way, effects, go, many, support, more, several, end, has, day, see, need, make, get, means, public, is, high, help, money, find, found, same*

Table 4: Claim indicators (see text) and shared open class words across the UKP topics.

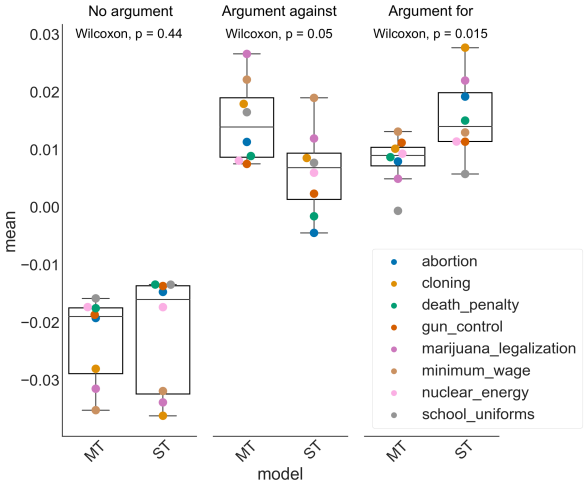


Figure 3: Boxplot of argument word LIME weights with each point representing the topic mean of the argument word weights. We find significant differences between the weights resulting from a single-task and multi-task model towards the two argument classes ARGUMENT AGAINST and ARGUMENT FOR at the 5 and 1 percent significance level, respectively. Furthermore, argument words are weighted significantly higher in the two argument classes than in the NO ARGUMENT class, at the 0.01 significance level.

words indicative of arguments, we use the claim indicator list provided in the appendix for the Vaccination Corpus' annotation guideline (Morante et al., 2020), which is in turn based on (Stab and Gurevych, 2017). We simplify the indicators to unigrams and combine the set with a few additions from Young's Critical Thinking course website; see Table 4. For each held-out topic, we compute the average LIME weight of each claim indicator. Figure 3 shows a boxplot with these averages across single-task and multi-task models. We test for significance using the Wilcoxon signed-rank test. Argument words are weighted significantly higher in the two argument classes compared to NO ARGUMENT, at the 0.01 significance level, as would be expected. With ARGUMENT AGAINST, we find significantly higher weights attributed to argument words by the multi-task models. However, with ARGUMENT FOR, the opposite scenario is observed. Hence, multi-task learning does not robustly move emphasis to claim indicators. Moreover, when normalising the weights by frequency before averaging, the significant difference between single-task and multi-task in ARGUMENT FOR disappears.

e) Removing spurious features We have seen how our models rely on spurious features such as *gun* and *marijuana*. What happens if we remove this? Obviously, removing *only* such words would require expensive manual annotation (like we did for the top-20 LIME words), but we can do something more aggressive (with high recall), namely to remove all open class words. If a model that relies only on closed-class words exhibits better performance across distant topics than state-of-the-art models, this is strong evidence that this model overfits to spurious features.

To this end, we train single-task models (ST-DNN) with all open class words replaced by unknown tokens. We call this model CLOSED. We report macro F_1 on UKP for each held-out topic, as well as an average across topics, in Table 1. We also train a model with closed-class words *and* the open class words that are shared across *all* eight topics. This amounts to 67 open class words, in total; see Table 4.[7] We include these 67 open class words in CLOSED+SHARED (in Table 1) – and find that this small set of words increase the average macro F_1 with 2 percentage points over CLOSED. Another effect of training CLOSED and CLOSED+SHARED models is that the large *variance in performance across topics largely disappears*.

To explore whether removing open class words may improve generalization to more distant topics, we test the constrained models on the test sets of VacC and IBM. While the UKP dataset has three classes, the evaluation datasets have two. We, there-

[7]It is worth noting that the set of 67 common open class words above reflects that *some* words common across topics are in fact of an argumentative nature, with verbs such as "said", "find" and "found" that are often used for referencing sources when providing reasons for claims. We inspected common words among the highest-ranking open class words. We found that very few highly weighted words transfer across more than a few topics, e.g. even at the top 200 level, only one word, namely *cost*, transfer across four, i.e. half, of the topics.

Model	IBM	VacC
ST-DNN	.656	.504
CLOSED+SHARED	**.670**	**.569**
Supervised *(upper bound)*		
MT-DNN+VacC+IBM+AQ	.813	.856

Table 5: ST-DNN and CLOSED+SHARED models are trained solely on the UKP corpus, and we here report these model's performance (macro F1) on the binary, out-of-domain corpora (IBM and VacC). The supervised upper bound is (multi-task) trained on the training data of all four datasets.

fore, merge the two argument classes in UKP when evaluating test performance on VacC and IBM. We report the average test score of the eight models (holding out different UKP topics). Results are found in Table 5 along with a single-task model baseline, i.e., the standard ST-DNN model trained on the UKP corpus, as well as the upper bound on performance provided by an MT-DNN model trained on all four datasets, including the two target datasets. The CLOSED+SHARED model – somewhat surprisingly and very encouragingly – performs *better* than the unconstrained ST-DNN for both test sets (by some margin). This indicates that state-of-the-art argument mining systems overfit to spurious correlations, as well as the need for evaluation on more distant topics.

6 Related Work

Feature analysis in argument mining Daxenberger et al. (2017) underline, like us, the challenge of cross-domain generalization in argument mining, finding that models performing best in-domain may not be the ones performing best out-of-domain, which they argue may in part be due to different notions of claims in the dataset development. Through experiments with different feature groups, such as embeddings, syntax or lexical features, they find lexical clues to be the "essence" of claims and that simple rules are important for cross-domain performance. Simple lexical clues are also found to be effective for argument mining in Levy et al. (2018), who create a claim lexicon, as well as in Lin et al. (2019) who investigate the effectiveness of integrating lexica (a claim lexicon, a sentiment lexicon, an emotion lexicon and the Princeton WordNet[8]) in the attention mechanism of a BiLSTM, but evaluate this only in the context of in-domain argument mining.

Feature analysis in deep neural networks Feature analysis in deep neural networks is not straightforward but, by now, several approaches to attribute importance in deep neural networks to features or input tokens are available. One advantage of LIME is that it can be applied to any model post-hoc. Other approaches for interpreting transformers, specifically, focus on inspections of the attention weights (Abnar and Zuidema, 2020; Vig, 2019) and vector norms (Kobayashi et al., 2020).

Spurious correlations in text classification Landeiro and Culotta (2018) provide a thorough description of spurious correlations deriving from confounding factors in text classification and outline methods from social science of controlling for confounds. However, these methods require the confounding factors to be known, which is often not the case. This problem is tackled by Wang and Culotta (2020) who, in contrast, develop a computational method for distinguishing spurious from genuine correlations in text classification to adjust for the identified spurious features to improve model robustness. They consider spurious correlations in sentiment classification and toxicity detection. McHardy et al. (2019) identified similar problems in sarcasm detection and suggested adversarial training to reduce sensitivity to spurious correlations. Kumar et al. (2019) present a similar method to avoid "topical confounds" in native language identification.

MTL to regularize spurious correlations Tu et al. (2020) suggest multi-task learning increase robustness to spurious correlations. Multi-task learning has previously been shown to be an effective regularizer (Søgaard and Goldberg, 2016; Sener and Koltun, 2018), leading to better generalization to new domains (Cheng et al., 2015; Peng and Dredze, 2017). Jabbour et al. (2020), though, presents experiments in automated diagnosis of disease based on chest X-rays suggesting that multi-task learning is not always robust to spurious correlations. In our study, we expected multi-task learning to move emphasis to closed-class items and claim indicators and away from the spurious correlations that do not hold as general markers of claims and arguments across topics and domains. Still, our analysis of feature weights does not indicate that multi-task learning is effective to this end.

[8] https://wordnet.princeton.edu/

7 Conclusion

We have shown that cross-topic evaluation of argument mining is insufficient to prevent models from relying on spurious features. Many of the spurious correlations that our models rely on are shared across some pairs of UKP topics but fail to generalise to distant topics (IBM and VacC). This shows cross-topic evaluation *can* encourage learning from signals, rather than spurious features; the problem with the protocol in Stab et al. (2018) is using *multiple* source topics. When using multiple source topics for training (and if the annotation relies on arguments being related to these topics), the models may overly rely on features that are frequent in debates of these topics but are not related to the forming of an argument and hence do not generalise well to unseen topics. The variance in cross-topic performance may be explained by some topic words transferring across a few topics, since the large variance disappears when removing open-class words. We propose evaluating on more distant held-out topics or simply considering the worst-case performance across all pairs of topics to estimate real-world out-of-topic performance.[9]

Acknowledgements

Maria Barrett is supported by a research grant (34437) from VILLUM FONDEN.

References

Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, Online. Association for Computational Linguistics.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. Cite arxiv:1907.02893.

Rishi Bommasani and Claire Cardie. 2020. Intrinsic evaluation of summarization datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8075–8096, Online. Association for Computational Linguistics.

Steven Bramhall, Hayley Horn, Michael Tieu, and Nibhrat Lohia. 2020. Qlime-a quadratic local interpretable model-agnostic explanation approach. *SMU Data Science Review*, 3.

[9] See Rüd et al. (2011) or Sultan et al. (2016), for example, for similar arguments in cross-domain NLP.

Hao Cheng, Hao Fang, and Mari Ostendorf. 2015. Open-domain name error detection using a multi-task RNN. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 737–746, Lisbon, Portugal. Association for Computational Linguistics.

Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. 2017. What is the essence of a claim? cross-domain claim identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2055–2066, Copenhagen, Denmark. 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.

Radwa Elshawi, Mouaz Al-Mallah, and Sherif Sakr. 2019. On the interpretability of machine learning-based model for predicting hypertension. *BMC Med Inform Decis Mak.*, 19.

Sebastian Gruber. 2019. LIME and sampling. In Christoph Molnar, editor, *Limitations of Interpretable Machine Learning Methods*, chapter 13.

Sarah Jabbour, David Fouhey, Ella Kazerooni, Michael W. Sjoding, and Jenna Wiens. 2020. Deep learning applied to chest x-rays: Exploiting and preventing shortcuts.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7057–7075, Online. Association for Computational Linguistics.

Philipp Kopper. 2019. Lime and neighborhood. In Christoph Molnar, editor, *Limitations of Interpretable Machine Learning Methods*, chapter 13.

Sachin Kumar, Shuly Wintner, Noah A. Smith, and Yulia Tsvetkov. 2019. Topics to avoid: Demoting latent confounds in text classification. 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 4153–4163, Hong Kong, China. Association for Computational Linguistics.

Virgile Landeiro and Aron Culotta. 2018. Robust text classification under confounding shift. *Journal of Artificial Intelligence Research*, 63:391–419.

Thibault Laugel, Xavier Renard, Marie-Jeanne Lesot, Christophe Marsala, and Marcin Detyniecki. 2018. Defining locality for surrogates in post-hoc interpretablity. *arXiv preprint arXiv:1806.07498*.

Ran Levy, Ben Bogin, Shai Gretz, Ranit Aharonov, and Noam Slonim. 2018. Towards an argumentative content search engine using weak supervision. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2066–2081, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ran Levy, Shai Gretz, Benjamin Sznajder, Shay Hummel, Ranit Aharonov, and Noam Slonim. 2017. Unsupervised corpus–wide claim detection. In *Proceedings of the 4th Workshop on Argument Mining*, pages 79–84, Copenhagen, Denmark. Association for Computational Linguistics.

Jian-Fu Lin, Kuo Yu Huang, Hen-Hsen Huang, and Hsin-Hsi Chen. 2019. Lexicon guided attentive neural network model for argument mining. In *Proceedings of the 6th Workshop on Argument Mining*, pages 67–73, Florence, Italy. Association for Computational Linguistics.

Anna Lindahl, Lars Borin, and Jacobo Rouces. 2019. Towards assessing argumentation annotation - a first step. In *Proceedings of the 6th Workshop on Argument Mining*, pages 177–186, Florence, Italy. Association for Computational Linguistics.

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.

Robert McHardy, Heike Adel, and Roman Klinger. 2019. Adversarial training for satire detection: Controlling for confounding variables. 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 660–665, Minneapolis, Minnesota. Association for Computational Linguistics.

Roser Morante, Chantal van Son, Isa Maks, and Piek Vossen. 2020. Annotating perspectives on vaccination. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4964–4973, Marseille, France. European Language Resources Association.

Nanyun Peng and Mark Dredze. 2017. Multi-task domain adaptation for sequence tagging. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 91–100, Vancouver, Canada. Association for Computational Linguistics.

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and clustering of arguments with contextualized word embeddings. *arXiv*, page 1906.09821v1.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144.

Stefan Rüd, Massimiliano Ciaramita, Jens Müller, and Hinrich Schütze. 2011. Piggyback: Using search engines for robust cross-domain named entity recognition. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 965–975, Portland, Oregon, USA. Association for Computational Linguistics.

Ozan Sener and Vladlen Koltun. 2018. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems*, volume 31, pages 527–538. Curran Associates, Inc.

Vikas Sindhwani and Prem Melville. 2008. Document-word co-regularization for semi-supervised sentiment analysis. In *ICDM*.

Anders Søgaard. 2013. Part-of-speech tagging with antagonistic adversaries. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 640–644, Sofia, Bulgaria. Association for Computational Linguistics.

Anders Søgaard and Yoav Goldberg. 2016. Deep multi-task learning with low level tasks supervised at lower layers. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 231–235, Berlin, Germany. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43:619–659.

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674.

Md Arafat Sultan, Jordan Boyd-Graber, and Tamara Sumner. 2016. Bayesian supervised domain adaptation for short text similarity. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 927–936, San Diego, California. Association for Computational Linguistics.

Charles Sutton, Michael Sindelar, and Andrew McCallum. 2006. Reducing weight undertraining in structured discriminative learning. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 89–95, New York City, USA. Association for Computational Linguistics.

Reid Swanson, Brian Ecker, and Marilyn Walker. 2015. Argument mining: Extracting arguments from online dialogue. In *Proceedings of the 16th annual meeting of the special interest group on discourse and dialogue*, pages 217–226.

Lifu Tu, Garima Lalwani, Spandana Gella, and He He. 2020. An empirical study on robustness to spurious correlations using pre-trained language models. *Transactions of the Association for Computational Linguistics*, 8:621–633.

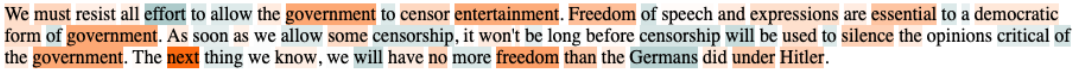
Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy. Association for Computational Linguistics.

Georgios Vlassopoulos. 2019. Decision boundary approximation: A new method for locally explaining predictions of complex classification models. Technical report, University of Leiden.

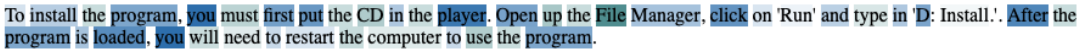
Marilyn A Walker, Jean E Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *LREC*, volume 12, pages 812–817. Istanbul.

Zhao Wang and Aron Culotta. 2020. Identifying spurious correlations for robust text classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3431–3440, Online. Association for Computational Linguistics.

Appendix

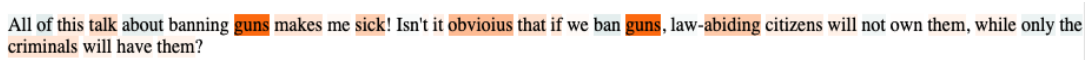


(a) This is an argument with the claim as the first sentence. The model has predicted ARGUMENT AGAINST. This makes sense because it is an argument against censorship, with this being the focus of the conclusion.

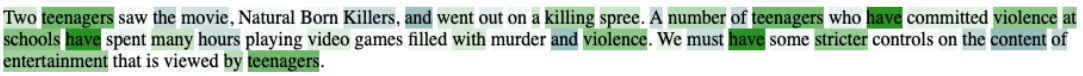


(b) The model has rightly predicted the example as not being an argument.

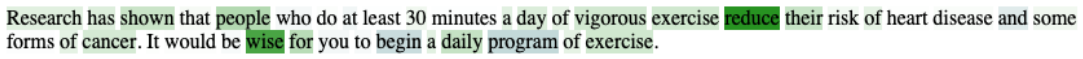
(c) This is an argument with the last sentence as the conclusion. The model incorrectly predicts it as not being an argument.



(d) This is not an argument. The model incorrectly predicts it as being an argument against something. This example is not formally an argument because it is formulated as a question. We note that Stab et al. (2018) likewise found questions among false positives in their error analysis.

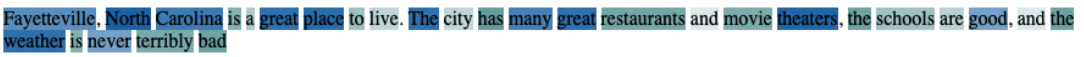


(e) This is an argument with the conclusion as the last sentence. The model correctly predicts it as an argument for something (for stricter controls on the content of entertainment).



(f) This is an argument with the conclusion as the last sentence. The model correctly predicts it as an argument for something (for exercise).

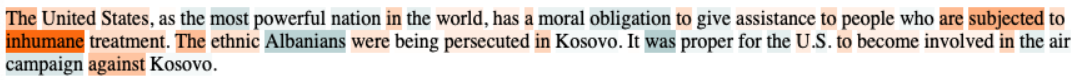
(g) This is an example is an argument with the confusion as the first sentence. The model incorrectly predicts it as not being an argument.



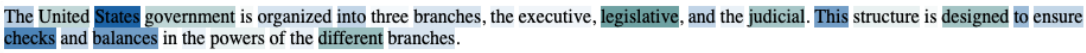
(h) This is an argument with the conclusion as the first sentence. The model incorrectly predicts it as not being an argument.

Figure 4: LIME explanations of the first eight challenge examples predicted by the best MT model, MT-DNN+AQ+IBM+VacC. Highlight colours represents weight towards a class; blue: NO ARGUMENT; orange: ARGUMENT AGAINST; green: ARGUMENT FOR. Darker colours mean larger weights.

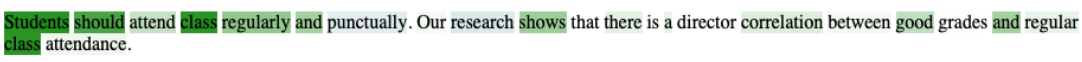
(a) This example is not an argument. The model correctly predicts it so.



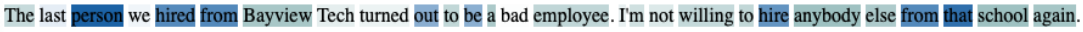
(b) This example is an argument with the conclusion as the last sentence. The model correctly predicts it as an argument, although as an argument against something rather than for (U.S. involvement).



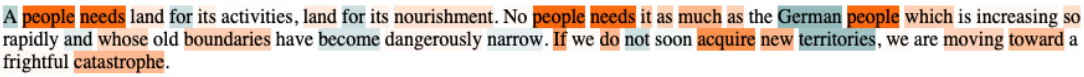
(c) This example is not an argument, and the model correctly predicts it so.



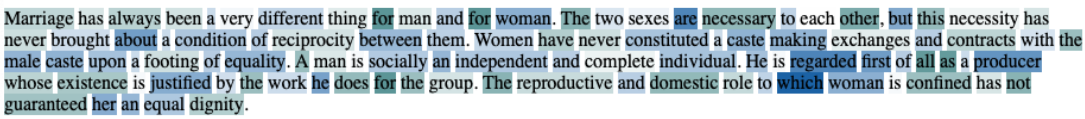
(d) This example is an argument with the conclusion as the first sentence. The model correctly predicts it as an argument for something (for student class attendance).



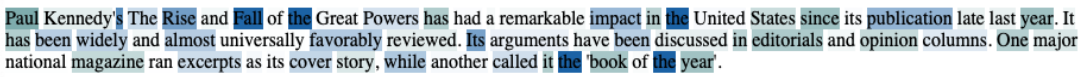
(e) This is an argument with the conclusion as the last sentence. The model incorrectly predicts is as not being an argument.



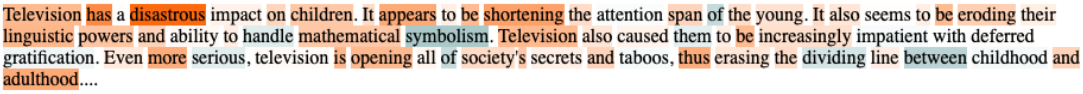
(f) This example is an argument with the conclusion as the last sentence. The model correctly predicts it as an argument against something (against an upcoming catastrophe caused by not acquiring new territories).



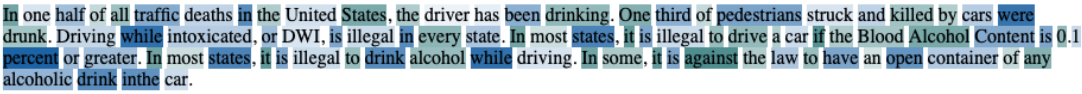
(g) This example is an argument with the conclusion as the first sentence. The model incorrectly predicts it as not being an argument.



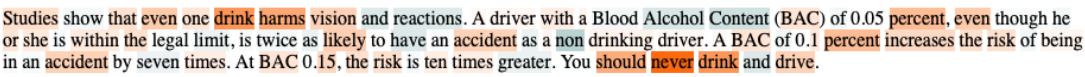
(h) This example is an argument with the conclusion as the first sentence. The model incorrectly predicts it as not being an argument.



(i) This example is an argument with the conclusion as the first sentence. The model correctly predicts it as an argument against something (against children watching television).



(j) This example is not an argument and the model correctly predicts it so.



(k) This example is an argument with the conclusion as last sentence. The model correctly predicts it as an argument against something (against drinking and driving).

Figure 5: LIME explanations of the last 11 challenge examples predicted by the best model MT model, MT-DNN+AQ+IBM+VacC. Highlight colours represents weight towards a class; blue: NO ARGUMENT; orange: ARGUMENT AGAINST; green: ARGUMENT FOR. Darker colours mean larger weights.

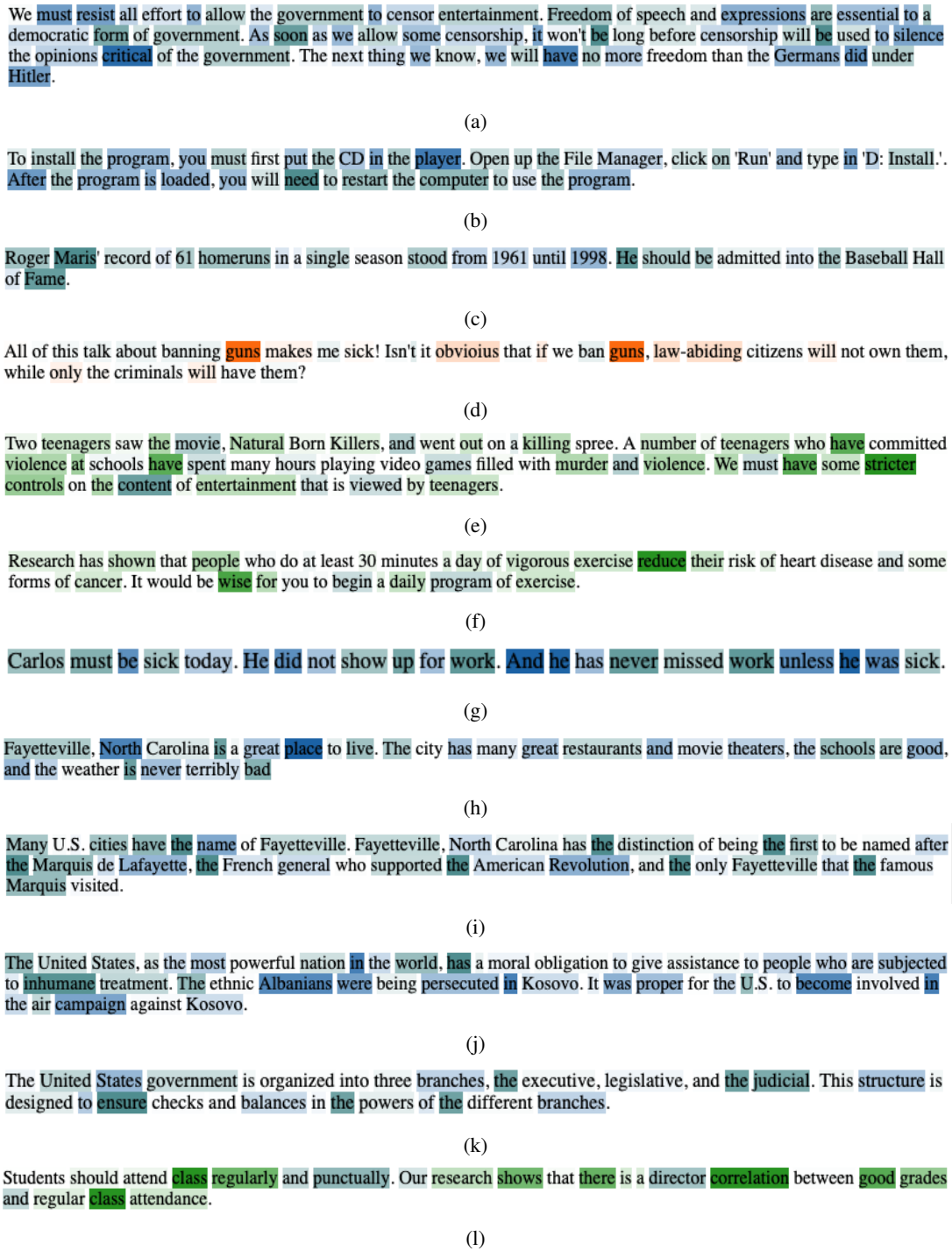


Figure 6: LIME explanations of the first 12 challenge examples predicted the single-task model. Highlight colours represents weight towards a class; blue: NO ARGUMENT; orange: ARGUMENT AGAINST; green: ARGUMENT FOR. Darker colours means larger weights.

Figure 7: LIME explanations of the last seven challenge examples predicted by the single-task model. Highlight colours represents weight towards a class; blue: NO ARGUMENT; orange: ARGUMENT AGAINST; green: ARGUMENT FOR. Darker colours means larger weights.