A Robustness Evaluation Framework for Argument Mining

Mehmet Sofi[*], **Matteo Fortier**[*], and **Oana Cocarascu**[†]
Department of Informatics, King's College London
{mehmet.sofi, matteo.fortier, oana.cocarascu}@kcl.ac.uk

Abstract

Standard practice for evaluating the performance of machine learning models for argument mining is to report different metrics such as accuracy or F_1. However, little is usually known about the model's stability and consistency when deployed in real-world settings. In this paper, we propose a robustness evaluation framework to guide the design of rigorous argument mining models. As part of the framework, we introduce several novel robustness tests tailored specifically to argument mining tasks. Additionally, we integrate existing robustness tests designed for other natural language processing tasks and re-purpose them for argument mining. Finally, we illustrate the utility of our framework on two widely used argument mining corpora, UKP topic-sentences and IBM Debater Evidence Sentence. We argue that our framework should be used in conjunction with standard performance evaluation techniques as a measure of model stability.

1 Introduction

Deep learning models have obtained state-of-the-art results on a wide range of Natural Language Processing (NLP) tasks and have even achieved superhuman performance on benchmark tasks (Wang et al., 2019). The standard approach for evaluating machine learning models is to use held-out data and report various performance metrics such as accuracy and F_1.

However, reporting an aggregate statistic on benchmarks does not reflect the model's performance and robustness when applied to real-world texts. Indeed, recent works have shown that NLP models are not robust to perturbations. For instance, natural language inference (NLI) models classify a permuted example where word positions are randomly changed, as they would classify the original input (Sinha et al., 2021), and sentiment analysis models give a lower sentiment score when a positive phrase is added to the original example (Ribeiro et al., 2020). Koch et al. (2021) argue for rigorous evaluation to avoid poor generalisability, whereas Raji et al. (2021) propose systematic development of test suites. Several frameworks have been developed for evaluating the robustness of NLP models, for example CheckList (Ribeiro et al., 2020), TextAttack (Morris et al., 2020), Robustness Gym (Goel et al., 2021), and TextFlint (Wang et al., 2021). There is limited work on evaluating the robustness of argument mining models (Mayer et al., 2020; Schiller et al., 2021), and the linguistic and logical reasoning required in argument mining tasks have so far been ignored.

In this paper we propose a robustness evaluation framework for machine learning-based argument mining models. In particular, we propose a variety of *simulation functions* that, given a *seed dataset*, automatically create *simulated datasets*. The simulated datasets are designed to mimic realistic settings which can be used to test the model's robustness.

Our framework is model-agnostic and only requires access to the data. We propose several novel robustness tests tailored to the argument mining task (e.g. argument removal, motion syntax inversion, motion negation, motion synonym/antonym verb replacement, etc.) as well as re-purpose robustness tests previously applied to other NLP tasks (e.g. contract/expand contraction, verb tense change, back-translation, etc.). We focus on two major corpora available for argument mining: the UKP topic-based sentential argument mining corpus (Stab et al., 2018) where the task is to determine whether a sentence is an argument for a topic and whether it supports or opposes the topic, and the IBM Debater Evidence Sentences corpus (Ein-Dor et al., 2020) where the task is to determine whether a sentence includes evidence for a given

[*] Equal contribution.

[†] Corresponding author.

Proceedings of the 9th Workshop on Argument Mining, co-located with the 29th International Conference on Computational Linguistics, pages 171–180
October 17, 2022.

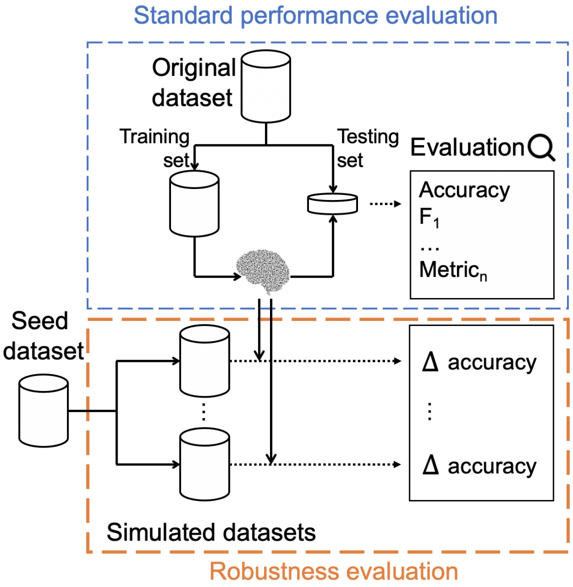

Figure 1: An overview of our proposed robustness evaluation framework for argument mining and how it complements standard performance evaluation.

motion. While other works on robustness focus on adversarial training (e.g. Morris et al. (2020)), our contributions are a range of *functions* that generate *simulated datasets* that reflect real-world examples. We believe our robustness evaluation framework can be used to enhance the standard performance evaluation in order to create better models for argument mining. Figure 1 gives an overview of our proposed robustness evaluation framework.

2 Related Work

There is a plethora of work in evaluating the robustness of NLP models that cover a variety of tasks: sentiment analysis (Ribeiro et al., 2020; Goel et al., 2021; Kiela et al., 2021; Moradi and Samwald, 2021; Wu et al., 2021; Jin et al., 2020; Wang et al., 2021; Li et al., 2020), machine translation (Sai et al., 2021; Morris et al., 2020; Wang et al., 2021), natural language inference (Tarunesh et al., 2021; Goel et al., 2021; Kiela et al., 2021; Morris et al., 2020; Wu et al., 2021; Jin et al., 2020; Wang et al., 2021; Li et al., 2020), question answering (Goel et al., 2021; Moradi and Samwald, 2021; Kiela et al., 2021), duplicate question detection (Ribeiro et al., 2020; Wu et al., 2021), and fake news classification (Jin et al., 2020; Li et al., 2020).

Robustness is evaluated by perturbing data and checking whether the model responds correctly to these changes. Amongst the most commonly used transformations (note that we use “perturbation" and “transformation" interchangeably in this paper) we find: punctuation errors (Sai et al., 2021), typos (Ribeiro et al., 2020; Sai et al., 2021; Wang et al., 2021), synonym replacement (Ribeiro et al., 2020; Moradi and Samwald, 2021; Sai et al., 2021; Morris et al., 2020; Jin et al., 2020; Wang et al., 2021), contractions (Sai et al., 2021; Wang et al., 2021), verb tense change (Wang et al., 2021; Moradi and Samwald, 2021), entity replacement (Ribeiro et al., 2020), back-translation (Goel et al., 2021; Wang et al., 2021), negation (Ribeiro et al., 2020; Moradi and Samwald, 2021; Wu et al., 2021), and using BERT (Devlin et al., 2019) for word replacement (Li et al., 2020). In this paper, we draw from previous works and apply commonly used data transformations in NLP tasks to argument mining.

Regarding task-specific perturbations, TextFlint includes perturbations for NLI, machine translation, and sentiment analysis amongst others, while Tarunesh et al. (2021) extend CheckList with templates tailored for the NLI task to cover more linguistic and logical reasoning such as causal, spatial, and pragmatic.

To the best of our knowledge, only two works have considered the robustness of argument mining models, for topic-dependent argument classification models (Mayer et al., 2020) and stance detection (Schiller et al., 2021). Schiller et al. (2021) used simple linguistic transformations such as two typos and negation by adding the tautology “and false is not true" after each sentence. Mayer et al. (2020) proposed more transformations such as punctuation errors, entity replacement, replacing a noun with its hyponym, using topic alternatives (e.g. *death penalty* → *capital punishment*), and adding speculative adverbs in the evidence text (e.g. *cannabis leads to other drugs* → *cannabis indeed leads to other drugs*), and used these transformations in adversarial training. In both works, the sentence-level topic information within an argument or motion, which we believe to be a key aspect in argument mining, is ignored. In this paper, we propose a robustness evaluation framework and introduce a variety of novel transformations tailored for the argument mining task as well as use existing transformations for NLP tasks and apply them to argument mining.

3 Robustness Tests for Argument Mining

We first introduce the terminology used in this paper. Given an original dataset with N instances

Topic	Sentence	Label
nuclear energy	It has been determined that the amount of greenhouse gases have decreased by almost half because of the prevalence in the utilization of nuclear power.	supporting arg
minimum wage	A 2014 study [. . .] found that minimum wage workers are more likely to report poor health, suffer from chronic diseases, and be unable to afford balanced meals.	opposing arg
minimum wage	We should abolish all Federal wage standards and allow states and localities to set their own minimums.	non-arg

Table 1: Examples from the UKP dataset.

Motion	Sentence	Label
We should legalize doping in sport	Although the number of cases is low, the Basque regional governments started introducing anti-doping measures in 1997 and created the office of Official Veterinarian in 2005 to help ensure good practice.	arg
We should legalize doping in sport	Contador signed a commitment in which he stated: "I am not involved in the Puerto affair nor in any other doping case".	non-arg
We should lower the drinking age	Alcohol and minors: initiatives seek to discourage underage drinking by providing tools and supporting parents and teachers to engage with minors.	arg
We should lower the drinking age	Some bottles now carry a warning stating that they are not for consumption by people under the legal drinking age (under 18 in the UK and 21 in the United States).	non-arg

Table 2: Examples from the IBM dataset.

$\mathcal{X} = \{X_1, X_2, ..., X_N\}$, where X_i is a pair of texts, and a corresponding set of N labels $\mathcal{Y} = \{Y_1, Y_2, ..., Y_N\}$, we train a model $F : \mathcal{X} \to \mathcal{Y}$ that maps the inputs $\mathcal{X}$ to the label space $\mathcal{Y}$.

We define a *simulation function* sim to be a function that takes a labelled dataset, called *seed dataset*, and creates a new, labelled *simulated dataset* $\mathcal{S}$ with the corresponding set of labels $\mathcal{Y}'$. For example, we may have $(\mathcal{S}, \mathcal{Y}') = sim(\mathcal{X}, \mathcal{Y})$, but other sub-sets of $\mathcal{X}$ could be used, such as the training set or the validation set.

A *robustness test* consists of applying a *simulation function* to obtain a *simulated dataset* and then evaluating a model's robustness on the *simulated dataset* (see Figure 1 for an overview). The model robustness is recorded as the difference between the model's performance on the original dataset and the model's performance on the simulated dataset.

Next, we describe the two argument mining datasets we use as *seed datasets* and the *simulation functions* we propose for obtaining *simulation datasets* that can be used to test the robustness of argument mining models.

3.1 Seed Datasets

There are two major corpora available for argument mining: UKP topic-based sentential argument mining corpus (Stab et al., 2018) and IBM Debater Evidence Sentences corpus (Ein-Dor et al., 2020).

The UKP dataset consists of 25,492 sentences for 8 topics (abortion, cloning, death penalty, gun control, marijuana legalization, minimum wage, nuclear energy, school uniforms), labelled as *supporting*, *opposing*, or *non-argument*. A text is deemed to be an argument if it provides evidence or reasoning that can be used to support or oppose a given topic. Table 1 shows examples from UKP.

The IBM dataset consists of 29,429 sentences for 221 motions that have a “dominant concept" (e.g. higher education, distance education, athletic scholarship, olympic games, alcoholic drink, hydroelectricity). Each sentence in a motion-sentence pair has an acceptance rate between 0 and 1 reflecting whether the sentence can be considered as evidence supporting or opposing the motion. Here, we consider sentences with an acceptance rate above 0.5 as *arguments*, and sentences with an acceptance rate below 0.5 as *non-arguments*. Table 2 shows examples from the IBM dataset.

3.2 Simulation Functions for Robustness Tests

We propose 15 simulation functions for testing the robustness of argument mining models. We define novel robustness tests tailored for the argument classification task which exploit the sentence-level topic information within an argument or motion: topic change, argument removal, motion syntax inversion, motion negation, motion verb replacement, and motion replacement. In addition, we integrate existing robustness tests and apply them to argument mining: motion topic synonym, motion adverbial modifier, punctuation error, typo, contract/expand contraction, synonym replacement, verb tense change, entity replacement, back-translation. Some simulation functions result in a change in the label (i.e. topic change, argument removal, motion replacement), while the rest of the simulation functions keep the label unchanged.

In the following, we describe our simulation functions, and indicate in brackets if a function can be applied to only one of the datasets.

Topic change (UKP): In this simulation function, we randomly change the topic of the argument to one of the other topics in the dataset. As an argument for a topic (e.g. “abortion") cannot be an argument for another topic (e.g. “minimum wage"),[1] the model should classify the new text as *non-argument*. This test is also applied to instances labelled *non-argument* to check whether the model can consistently classify texts that are unrelated or provide no evidence for the topic as *non-argument*.

Argument removal (UKP): An argument expresses evidence for/against a topic, thus a sentence that expresses an opinion for/against a topic but does not provide evidence would be classified as *non-argument*. In this test, we remove the evidence from an argument and expect the model to classify the new text as *non-argument*. We use premise and conclusion indicators to implement this test. In particular, premise indicators can be found before the evidence, thus removing the text after the indicators would remove the evidence; similarly, conclusion indicators can be found after the evidence and removing the text before the indicators would remove the evidence. We remove the evidence based on the occurrence of certain keywords used in discourse that indicate the presence of a premise or conclusion. We use the following *conclusion indicators*: {"therefore", "thus", "hence", "consequently", "ergo", "it proves that", "in conclusion", "suggests that", "so", "it follows that", "implies that", "we can infer that", "we can conclude that"}, and the following *premise indicators*: {"because", "since", "supposing that", "assuming that", "given that", "as indicated by", "the fact that", "it follows from", "for", "as", "follows from", "as shown by", "the reason is that"}. We implement two variations of this test on the instances labelled as *supporting/opposing argument*: *i)* testing whether removing the evidence using indicators will result in the model classifying the text as *non-argument*, and *ii)* confidence testing which uses the model's output for each label and evaluates whether the text with the argument removed has a higher confidence in the *non-argument* label when compared with the text where the evidence is preserved.

Motion topic synonym (IBM): In this simulation function, we replace the topic of a motion with a synonymous topic. The topic can be the passive nominal subject or direct object of the motion sentence. We use spaCy[2] to identify the motion topic and sense2vec (Trask et al., 2015) to obtain topic alternatives and their similarity scores, and select the top-scoring alternative topic with similarity score above 80%.

Motion syntax inversion (IBM): This simulation function recognises and reconstructs motion sentences using a different syntax. We identify four types of motion syntax, defined by the dependency of the topic within the motion: passive nominal subject topic, nominal subject topic, direct object topic, and object of preposition topic. We use spaCy, in particular dependency tags and part-of-speech tags to recognise the motion syntax, and then invert it.

Motion negation (IBM): We negate a motion by adding the word *not*. We expect the model to predict the label of the instance in the seed dataset as negation does not affect whether a sentence is or is not an argument for the motion, distinguishing the argument classification task from a supporting/opposing relation prediction task.

Motion adverbial modifier (IBM): In this simulation function, we add adverbial modifiers (i.e. *absolutely*, *indeed*, *certainly*, and *definitely*) or use them to replace existing adverbial modifiers. We use dependency tags and part-of-speech tags to ensure the adverbial modifier is added in the correct location in the sentence.

Motion verb replacement (IBM): We replace the root verb in a motion with a synonymous or antonymous verb. Similarly to motion negation, using an antonym of the root verb does not affect whether a sentence is or is not an argument for the motion. We use SupWSD (Papandrea et al., 2017), a supervised Word Sense Disambiguation (WSD) model, to obtain the WordNet (Fellbaum, 1998) senses of words in a sentence from which we determine the synonyms and antonyms that we use to replace the root verbs. We also ensure that all verbs replaced are conjugated as in the original sentence.

Motion replacement (IBM): In this test, we replace the motion text of a motion-sentence pair with another motion text from the dataset, and ex-

[1] Note that this is possible due to the non-overlapping topics in the UKP dataset.

[2] https://spacy.io/

pect the model to predict *non-argument*. We implement two variations of this test: *i)* replacing the motion with the most similar motion in the dataset given the motion topic and *ii)* replacing the motion with the most different motion in the dataset given the motion topic. We use sense2vec on the "dominant concept" in the IBM dataset to sort motions based on their similarity score to a given motion. If the concept cannot be found in the sense2vec model, we use spaCy's similarity score computed using the average vector of word embeddings.

The remaining simulation functions are applied to the sentences in the topic/motion sentence pairs.

Punctuation error: Punctuation errors arise from the misuse or absence of punctuation marks. In this simulation function, we use CheckList that adds/removes a single punctuation mark. Given that texts found in online sources often omit several or all punctuation marks, to test the model's robustness we also implement a simulation function where all punctuation marks are removed.

Typo: Typos represent mistakes made when typing. As the datasets were collected from online sources where typos are common, it is important to test the model's robustness against these errors. We use CheckList to implement this simulation function as CheckList has support for adding typos. We introduce different number of typos: 1, 2, and 3 typos, respectively.

Contract/Expand contraction: Contractions represent shortened versions of words. In this simulation function, we expand contractions (e.g. *aren't* → *are not*) or contract the expanded contractions (e.g. *are not* → *aren't*), depending on the form used in the sentence. We use Checklist to contract and to expand contractions in texts.

Synonym replacement: Synonyms are words that are similar or have a related meaning and we use them to increase the language variety. In this simulation function, we replace each word in the text with a context appropriate synonym using CheckList's inbuilt synonym replacement feature.

Verb tense change: Grammar errors occur frequently in online sources. We introduce grammar errors by changing the verb tense. We use spaCy and LemmInflect[3] to identify the verbs in text and to change their tense. We create a new text for each verb inflection; if an argument contains several verbs, we create a new text for each verb.

Entity replacement: We identify entities (e.g. date, event, location, etc.) using spaCy and replace them with 10 words/phrases chosen randomly from their respective categories. We limit the number of replacements to 10 due to the high number of entities in each category.[4]

Back-translation: Back-translation is the process by which a text is translated from one language L_1 to another language L_2 and then back to L_1, resulting in a text with similar meaning, but different structure. We experiment with 3 configurations to capture the linguistic variance between the original sentence and its back-translated counterpart: English → French → English, English → Russian → English, and English → Arabic → English. We use the OPUS-MT (Tiedemann and Thottingal, 2020) model from EasyNMT[5] to translate texts from English to the target languages and back.

Table 3 shows examples from the simulated datasets obtained from UKP and IBM.

4 Experiments

In this section, we apply our proposed *simulation functions* and evaluate the robustness of argument mining models. We use UKP and IBM, respectively, as seed datasets. We adopt the methodology in Wang et al. (2021) and apply each simulation function on the original dataset to generate the corresponding simulated dataset. Depending on the simulation function used, the simulated dataset may be of different size compared to the seed dataset. For example, some functions are not applicable to all instances in the seed dataset (e.g. contraction), while other functions may result in creating one example (e.g. punctuation error) or several examples for each instance in the dataset (e.g. synonym replacement, entity replacement).

We experiment with BERT (Devlin et al., 2019), a pre-trained transformer network (Vaswani et al., 2017) which set state-of-the-art performance on various sentence classification and sentence-pair classification tasks. We use bert-base-cased and fine-tune on each dataset. For UKP, we train using the proposed train-test-validation sets and we obtain 71.7% accuracy and 67.4% macro F_1, using e-3 as learning rate and training for 21 epochs. For IBM, we split the dataset into 70% for training, 15% for testing and 15% for validation, and obtain 83.4% accuracy and 71% F_1, using 2e-5 as

[3] http://github.com/bjascob/LemmInflect

[4] We experimented with higher values, but the results were similar and hence decided not to include them.

[5] https://github.com/UKPLab/EasyNMT

Simulation function	Original text in seed dataset	New text in simulated dataset
Topic change	(**Abortion**, Abortion is wrong because it is taking a human life.)	(**Minimum Wage**, Abortion is wrong because it is taking a human life.)
Argument removal	Abortion is wrong **because it is taking a human life**.	Abortion is wrong.
Motion topic synonym	We should ban **alternative medicine**	We should ban **naturopathy**
Motion syntax inversion	Private universities should be banned	We should ban private universities
Motion negation	We should subsidize cultivation of tobacco	We should **not** subsidize cultivation of tobacco
Motion adverbial modifier	We should ban lotteries	We should **absolutely** ban lotteries
Motion adverbial modifier	We should **further** exploit wind turbines	We should **indeed** exploit wind turbines
Motion syn verb replacement	We should **abolish** the monarchy	We should **get rid of** the monarchy
Motion ant verb replacement	We should **prohibit** flag burning	We should **permit** flag burning
Motion similar replacement	**We should fight global warming**	**Tattoos should be banned**
Motion different replacement	**We should fight global warming**	**We should subsidize renewable energy**
Punctuation (single)	The war on poverty has not had any effect in the 40 + years that it has been going on**.**	The war on poverty has not had any effect in the 40 + years that it has been going on
Punctuation (all)	It is true**,** as conservative commentators often point out**,** that some minimum**-**wage workers are middle**-**class teenagers or secondary earners in fairly well**-**off households**.**	It is true as conservative commentators often point out that some minimumwage workers are middleclass teenagers or secondary earners in fairly welloff households
Typo	Milton Friedman **called them** a form of **discrimination against low**-skilled workers.	Milton Friedman **calledt hem** a form of **discriminatio nagainst lwo**-skilled workers.
Contraction	Not true: The typical minimum wage worker **is not** a high school student earning weekend pocket money.	Not true: The typical minimum wage worker **isn't** a high school student earning weekend pocket money.
Synonym replacement	And those employers, in turn, would be unable to hire as many people – an undesirable **result** when unemployment continues to hover at **about** 8 percent.	And those employers, in turn, would be unable to hire as many people – an undesirable **outcome** when unemployment continues to hover at {**around/nearly**} 8 percent.
Verb tense change	You really **want** your kids on that?	You really **wanting** your kids on that?
Entity replacement	In **2012** the richest 1% of the US population earned 22.83% of the nation 's total pre-tax income resulting in the widest gap between the rich and the poor since the 1920s.	In **1934** the richest 1% of the US population earned 22.83% of the nation's total pre-tax income resulting in the widest gap between the rich and the poor since the 1920s.
Back-translation	A woman can not sincerely be considered to have equal standing in society if she does not at least have the choice to remove the challenges that will come with a pregnancy.	A woman cannot sincerely be considered equal in society if she does not at least have the option to overcome the difficulties of pregnancy.

Table 3: Examples from the simulated datasets. The orange highlights indicate the portions of the original text in the seed dataset on which the function is applied and the green highlights indicate the changes in the new text.

learning rate and training for 3 epochs. Our results are higher than those previously reported on UKP (63.25% macro F_1) and on a smaller, but similar IBM dataset (81.37% accuracy) (Reimers et al., 2019).

Robustness has been evaluated in different ways: Ribeiro et al. (2020) check that the model's output is invariant when certain transformations are applied to the input, while others calculate the accuracy on the transformed set (Wang et al., 2021; Morris et al., 2020). We experiment with both methods and discuss model robustness and model consistency.

4.1 Model Robustness

We evaluate the robustness of the model in predicting the labels $\mathcal{Y}'$ of each simulated dataset $\mathcal{S}$. We report the percentage point change between the accuracy on the seed dataset and the accuracy on the simulated dataset in Table 4. In this paper, we used a single metric per transformation function, however additional metrics can be used. Overall, the results show that the BERT model trained on the UKP dataset is more robust than the model trained on the IBM dataset.

The tests topic change and argument removal are only applicable to UKP as the dataset contains topics rather than motions and three labels, *non-argument*, *supporting* and *opposing* argument, in contrast to the IBM dataset that has motions and two labels only, *argument* and *non-argument*. The results for the topic change test show that the model struggles to draw a distinction between an argument and its relation to the topic input. For example, the model classified the argument "But those predisposed to defending the interests of cor-

	Simulated UKP datasets (3 classes)		**Simulated IBM datasets (2 classes)**	
Simulation function	**Data size**	**Δ**	**Data size (# motions)**	**Δ**
Topic change	25,492	-8.99	n/a	n/a
Argument removal	5,963	-45.77	n/a	n/a
Argument removal (confidence)	5,963	-11.83	n/a	n/a
Motion topic synonym	n/a	n/a	10,455 (63)	-5.39
Motion syntax inversion	n/a	n/a	29,429 (221)	-5.65
Motion negation	n/a	n/a	29,429 (221)	-4.2
Motion adverbial modifier	n/a	n/a	29,429 (221)	-4.34
Motion synonym verb replacement	n/a	n/a	11,834 (205)	-2.64
Motion antonym verb replacement	n/a	n/a	6,781 (86)	-4.46
Motion similar replacement	n/a	n/a	29,429 (221)	-16.91
Motion different replacement	n/a	n/a	29,429 (221)	-0.82
Punctuation (single)	25,492	-0.18	29,429	-8.56
Punctuation (all)	25,492	-0.47	29,429	-8.31
One Typo	25,492	-1.26	29,429	-2.08
Two Typos	25,492	-3.52	29,429	-4.16
Three Typos	25,492	-5.69	29,429	-5.72
(Expand) Contraction	5,226	-1.67	4,182	-2.35
Synonym replacement	53,867	-0.97	53,867	+0.62
Verb tense change	201,786	-0.94	313,121	-2.06
Entity replacement	267,916	-0.44	772,870	+0.24
Back-Translation (French)	25,492	-2.56	29,42	+3.17
Back-Translation (Russian)	25,492	-5.88	29,42	-4.23
Back-Translation (Arabic)	25,492	-11.25	29,42	-3.75

Table 4: The percentage point change between the model's accuracy on the seed dataset and the accuracy on the simulated dataset for each simulation function.

porate America - including retailers and fast-food restaurants - oppose any increase" as an *opposing argument* for topic “school uniforms", when this is in fact an argument against the topic “minimum wage". We run two types of tests when removing the argument, the first in which we expect the model to predict *non-argument* and the second in which we expect an increase in the model's confidence for the class *non-argument*. The results of the confidence test show that the model is not robust, however the absence of premise and conclusion indicators increases the model's confidence that the argument has no reasoning or evidence.

For the IBM tests where we modified the motion, we sample ten instances from the simulated datasets and check their correctness. The results for motion topic synonym show that the model struggles with topics that it has not seen during training. The model failed when the following topic synonyms were used: “alternative medicine" → “naturopathy", “assisted suicide" → “euthanasia". Whilst we generate tests at large and thus improve over existing manual methods for generating similar tests (Mayer et al., 2020), we acknowledge that the automatic generation of synonyms is not a perfect task. For example, we noticed that the topic “fraternities" was replaced with “sororities" and topic “abortions" with “legal abortions"; assuming the concepts overlap sufficiently, the evidence sentences should still be arguments. In addition, we also observed incorrect topic synonyms such “wealth distribution" → “progressive taxation".

Regarding the other simulation functions that modify the motion, we evaluate the generated texts in the simulated datasets and find that they match their intended design. The motion syntax inversion test evaluates the models' ability for predicting motion-evidence relations by identifying the subject or topic in texts with different syntax. The motion negation test checks whether the model is able to identify motion-evidence relation even in the presence of the word *not* in the motion, while the motion adverbial modifier evaluates the model's robustness to adding or replacing adverbs. The motion synonym/antonym verb replacement tests are useful in determining the model's robustness towards the role of the root verb in predicting the motion-evidence relation. Similarly to previous cases when synonyms were used, we noticed one replacement to be incorrect: “We should ban fast food" → “We should censor fast food", otherwise, the simulated dataset matches the intended designs. For motion different replacement, all generated examples appear to be correct and the model classifies the motion-evidence as *non-argument*. However, for motion similar replacement, the motion concepts may overlap significantly, and thus the expected label should not change to *non-argument*. For example, “We should protect endangered species" and “We should increase eco-tourism" may share several arguments, as well as “We should ban lotteries" and “Casinos should be banned". Thus, further work is required to assess the suitability of the *motion similar replacement* test.

Regarding simulation functions applied to the sentences in the topic/motion sentence pairs, while the UKP model’s accuracy is relatively unaffected by the absence or addition of punctuation marks, the IBM model is sensitive to these types of changes. As the number of typos in a single argument increases, the model’s performance in identifying the correct label for the argument decreases. Upon inspection, we noticed that shorter arguments from UKP that were correctly classified under the one typo setup were misclassified under the two typo setup. Regarding the contraction test, the model struggled with the less common contractions such as “that would" → “that’d". The performance for this test was lower than expected; we believe that a BERT model with a large training set should be robust to contractions as they do not change the meaning of the sentence. The verb tense change result shows that the UKP model is able to identify the relation between sentences and the topic regardless of a verb’s tense, highlighting the fact that it can correctly classify the stance of an argument even in the presence of grammar errors. With respect to synonym replacement and entity replacement, the models for both datasets appear to be robust, with the UKP model yielding a small decrease in robustness while the IBM model yielded a small increase. We experimented with three languages for the back-translation tests: French, Russian and Arabic. As expected, French back-translation performed the best as English shares more similarities with French than with the other languages. We observed that the model failed on all three languages in cases where the translation model added noise and resulted in the argument losing its meaning.

4.2 Model Consistency

Beyond model robustness measured as the difference in accuracy between the seed dataset and the simulated dataset, we also evaluate whether the model is consistent in making predictions, i.e. we compare whether the model predicts the same label for an instance in the seed dataset and for its corresponding instance in the simulated dataset. Thus, we evaluate model consistency using the simulation functions that introduce minimal changes to the syntax (i.e. punctuation errors, typos, and contraction/expand contraction).

Simulation function	**UKP (%)**	**IBM (%)**
Punctuation (single)	99.08	95.19
Punctuation (all)	97.76	94.99
One Typo	93.24	95.01
Two Typos	86.73	90.32
Three Typos	81.91	86.71
(Expand) Contraction	97.95	99.47

Table 5: Model consistency results.

Table 5 shows the model consistency results. On the UKP dataset, the model’s prediction for adding/removing a single punctuation mark is consistent, while we see a decrease in consistency when removing all the punctuation marks. In contrast, the model’s prediction on the IBM dataset is less consistent. The consistency of both models’ predictions decreased as the number of typos increased, highlighting that the models are sensitive to small changes in the argument. The model consistency is higher on the IBM dataset than on the UKP dataset for typos and contractions.

5 Conclusion

We proposed a robustness evaluation framework for machine learning-based argument mining models. Our framework is model-agnostic and only requires access to the data. We presented 15 simulation functions, amongst which 6 are novel and tailored for the argument classification task by exploiting sentence-level topic information within an argument or motion, with the rest of the functions re-purposed for argument mining tasks. These can be used to automatically create simulated datasets, designed to mimic realistic settings which can be used to test the model’s robustness. We illustrated the utility of our framework on two widely used argument mining corpora, UKP topic-sentences and IBM Debater Evidence Sentence and showed that, while robust, BERT models can still be vulnerable to new inputs.

Our robustness evaluation framework can be used to enhance the standard performance evaluation in order to create better models for argument mining by measuring model stability. We experimented with the major corpora available for argument mining, however our framework can be applied to datasets for relation prediction in argument mining (Cocarascu et al., 2020).

There are several avenues for future work. First, we plan to apply our framework to other datasets and models used in argument mining. We also plan to use the simulated datasets in adversarial training to evaluate whether model robustness can be improved. Further, it would be useful to explore combining several simulation functions to create simulated datasets. Finally, one interesting line of research is to provide explanations and/or summaries of failures on the simulated datasets that can be used to understand why a model fails and thus work on improving it.

References

Oana Cocarascu, Elena Cabrio, Serena Villata, and Francesca Toni. 2020. Dataset independent baselines for relation prediction in argument mining. In *Computational Models of Argument - Proceedings of COMMA*, volume 326 of *Frontiers in Artificial Intelligence and Applications*, pages 45–52. IOS Press.

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, NAACL-HLT*, pages 4171–4186. Association for Computational Linguistics.

Liat Ein-Dor, Eyal Shnarch, Lena Dankin, Alon Halfon, Benjamin Sznajder, Ariel Gera, Carlos Alzate, Martin Gleize, Leshem Choshen, Yufang Hou, Yonatan Bilu, Ranit Aharonov, and Noam Slonim. 2020. Corpus wide argument mining - A working solution. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI*, pages 7683–7691. AAAI Press.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Karan Goel, Nazneen Fatema Rajani, Jesse Vig, Zachary Taschdjian, Mohit Bansal, and Christopher Ré. 2021. Robustness gym: Unifying the NLP evaluation landscape. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations, NAACL-HLT*, pages 42–55. Association for Computational Linguistics.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is BERT really robust? A strong baseline for natural language attack on text classification and entailment. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI*, pages 8018–8025. AAAI Press.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*, pages 4110–4124. Association for Computational Linguistics.

Bernard Koch, Emily Denton, Alex Hanna, and Jacob G. Foster. 2021. Reduced, reused and recycled: The life of a dataset in machine learning research. In *Proceedings of the Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks 1*.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. BERT-ATTACK: adversarial attack against BERT using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 6193–6202. Association for Computational Linguistics.

Tobias Mayer, Santiago Marro, Elena Cabrio, and Serena Villata. 2020. Generating adversarial examples for topic-dependent argument classification. In *Computational Models of Argument - Proceedings of COMMA*, volume 326 of *Frontiers in Artificial Intelligence and Applications*, pages 33–44. IOS Press.

Milad Moradi and Matthias Samwald. 2021. Evaluating the robustness of neural language models to input perturbations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 1558–1570. Association for Computational Linguistics.

John X. Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP*, pages 119–126. Association for Computational Linguistics.

Simone Papandrea, Alessandro Raganato, and Claudio Delli Bovi. 2017. SupWSD: A flexible toolkit for supervised word sense disambiguation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 103–108. Association for Computational Linguistics.

Inioluwa Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. 2021. AI and the everything in the whole wide world benchmark. In *Proceedings of the Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks 1*.

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and clustering of arguments with contextualized word embeddings. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL*, pages 567–578. Association for Computational Linguistics.

Marco Túlio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with checklist. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL*, pages 4902–4912. Association for Computational Linguistics.

Ananya B. Sai, Tanay Dixit, Dev Yashpal Sheth, Sreyas Mohan, and Mitesh M. Khapra. 2021. Perturbation checklists for evaluating NLG evaluation metrics. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 7219–7234. Association for Computational Linguistics.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Stance detection benchmark: How robust is your stance detection? *Künstliche Intell.*, 35(3):329–341.

Koustuv Sinha, Prasanna Parthasarathi, Joelle Pineau, and Adina Williams. 2021. Unnatural language inference. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP*, pages 7329–7346. Association for Computational Linguistics.

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 EMNLP*, pages 3664–3674. Association for Computational Linguistics.

Ishan Tarunesh, Somak Aditya, and Monojit Choudhury. 2021. Trusting RoBERTa over BERT: insights from checklisting the natural language inference task. *CoRR*, abs/2107.07229.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT - building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, EAMT*, pages 479–480. European Association for Machine Translation.

Andrew Trask, Phil Michalak, and John Liu. 2015. sense2vec - A fast and accurate method for word sense disambiguation in neural word embeddings. *CoRR*, abs/1511.06388.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems*, pages 5998–6008.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems, NeurIPS*, pages 3261–3275.

Xiao Wang, Qin Liu, Tao Gui, Qi Zhang, Yicheng Zou, Xin Zhou, Jiacheng Ye, Yongxin Zhang, Rui Zheng, Zexiong Pang, Qinzhuo Wu, Zhengyan Li, Chong Zhang, Ruotian Ma, Zichu Fei, Ruijian Cai, Jun Zhao, Xingwu Hu, Zhiheng Yan, Yiding Tan, Yuan Hu, Qiyuan Bian, Zhihua Liu, Shan Qin, Bolin Zhu, Xiaoyu Xing, Jinlan Fu, Yue Zhang, Minlong Peng, Xiaoqing Zheng, Yaqian Zhou, Zhongyu Wei, Xipeng Qiu, and Xuanjing Huang. 2021. TextFlint: Unified multilingual robustness evaluation toolkit for natural language processing. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL*, pages 347–355. Association for Computational Linguistics.

Tongshuang Wu, Marco Túlio Ribeiro, Jeffrey Heer, and Daniel S. Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP*, pages 6707–6723. Association for Computational Linguistics.