CAISA at SemEval-2023 Task 8: Counterfactual Data Augmentation for Mitigating Class Imbalance in Causal Claim Identification

Akbar Karimi[1,2] and **Lucie Flek**[1,2]

[1]Conversational AI and Social Analytics (CAISA) Lab, University of Bonn, Germany

[2]Conversational AI and Social Analytics (CAISA) Lab, University of Marburg, Germany

{akkarimi, lucie.flek}@bit.uni-bonn.de

https://caisa-lab.github.io

Abstract

The `class imbalance` problem can cause machine learning models to produce an undesirable performance on the minority class as well as the whole dataset. Using data augmentation techniques to increase the number of samples is one way to tackle this problem. We introduce a novel counterfactual data augmentation by verb replacement for the identification of medical claims. In addition, we investigate the impact of this method and compare it with 3 other data augmentation techniques, showing that the proposed method can result in a significant (relative) improvement in the minority class.

1 Introduction

Automatic identification of medical claims (Khetan et al., 2023; Wadhwa et al., 2023) is a task with various real-life applications in industries such as healthcare (Herland et al., 2017) and insurance (Wang and Xu, 2018) as well as content moderation (Schlicht et al., 2023). However, it can be a difficult task due to the lack of data for all or some categories. One solution for such an issue is increasing the number of data points in each category, especially the one that has significantly fewer samples. We can do this using data augmentation techniques (Temraz and Keane, 2022), which modify certain characteristics of an input sequence (or its representation in the embedding space) in order to create different versions of it. One example is entity replacement (Zeng et al., 2020), where entities in one sequence can be swapped with equivalent ones from another sequence. The advantage of this type of augmentation is that it provides more real context to the target entities.

Given that the task at hand is claim detection, we hypothesize that the verb in a sentence can be determinant in its category. Therefore, we address the problem of class imbalance using a novel data augmentation technique where we replace a verb in a sentence with other verbs from the training data. Our experiments show that verb replacement can improve the performance of a model on the target category. In addition, for more comparison, we experiment with several other data augmentation techniques, namely noise insertion (Karimi et al., 2021b), entity replacement (Zeng et al., 2020), augmentation with `YouChat`[1], and augmentation in the embedding space (Karimi et al., 2021a).

2 Background

Class Imbalance Problem. This problem frequently comes up in many domains and applications. As a result, it has been tackled by a variety of methods such as oversampling (Ling and Li, 1998) and undersampling (He and Garcia, 2009). The former method randomly selects some of the samples in the minority class and uses them multiple times for training in addition to the original samples. Contrarily, the latter randomly ignores some of the training examples from the majority class. However, the issue with them is that one (oversampling) might not always add new information to the training data, and the other (undersampling) might lose valuable information by not using some of the data points.

Data Augmentation. Another solution to tackling the class imbalance problem is to create synthetic instances from the existing ones (Chawla et al., 2002). With this method, the resulting samples can be more diverse which can help avoid overfitting. However, the trade-off is that it can also introduce noise to the system although introducing noise is not always harmful (Karimi et al., 2021b). In counterfactual data augmentation, words (or phrases) in a sentence are replaced with opposite (or different) ones from other sentences. This way, the focus parts of sentences are combined with different contexts, helping models in a better generalization to

[1]`https://you.com/chat`

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 2118–2123
July 13-14, 2023 ©2023 Association for Computational Linguistics

Data	Texts	Unique words	Max length
Train	5016	39685	1777
Dev	694	11287	1040
Test	1424	16948	7876

Table 1: Dataset statistics in Subtask 1

Data	Texts	Unique words	Max length
Train	501	10834	1802
Dev	96	3570	887
Test	150	4418	657

Table 2: Dataset statistics in Subtask 2

Data	CLA	EXP	O	PER	QUE
Train	8183	33358	316676	138359	51707
Dev	1045	3995	44172	18298	6803

Table 3: Label distribution of train and dev data in Subtask 1.

Data	INT	O	OUT	POP
Train	800	69594	768	444
Dev	185	12883	143	78

Table 4: Label distribution of train and dev data in Subtask 2

unseen sentences and combinations in the original data. For example, Zeng et al. (2020) replace named entities from one sentence with another one to create new samples for the task of named entity recognition. We use the same approach for tackling `PIO` extraction (Subtask 2). To do so, we first create a dictionary of all the `PIO`s. Then, to augment a sentence in the training set, we replace its `PIO`s randomly with other ones from the dictionary.

We compare the performance of the entity replacement with two other augmentation techniques. One is our counterfactual verb replacement method that we also used for Subtask 1 and the second one is an augmentation technique called `BAT` (Karimi et al., 2021a) that takes place in the embedding space instead of the input space.

2.1 Data Exploration

The dataset (Wadhwa et al., 2023) for the first task consists of `5710` texts that we split into two sets of training and development. Tables 1 and 2 show the number of samples for each set as well as the test set. As we can see from the tables, some texts can be longer than `1000` tokens. However, due to their low frequency (Figure 1), we train our baseline (`DistilBERT`) with `512` tokens.

The categories in the dataset in Subtask 1 include `Claims (CLA)`, `Claim per experience (EXP)`, `Per experience (PER)`, and `Questions (QUE)`. Additionally, the dataset for Subtask 2 includes the categories of `Population (POP)`, `Intervention (INT)`, and `Outcome (OUT)`. Tables 3 and 4 show the number of tokens for each category in Subtasks 1 and 2, respectively. In the former, the claims (`CLA`) category has significantly fewer samples than other categories. In the latter, on the other hand, all the entities are vastly outnumbered by the outside (`O`) class.

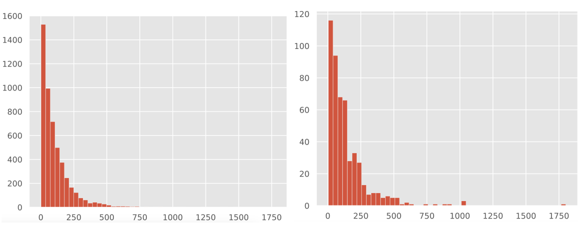

Figure 1: Sentence length distributions in the training data for Subtasks 1 and 2.

2.2 Data Augmentation Methods

We experiment with four data augmentation methods for the identification of causal claims (Subtask 1) and three methods for `PIO` extraction (Subtask 2). Table 5 shows an example for each method.

AEDA (Karimi et al., 2021b). This approach is based on inserting punctuation marks into the input sentences. This can help improve the generalization capability of the model by changing the position of the words in the input sentence.

Entity Replacement (Zeng et al., 2020). This method replaces the existing named entities with similar ones in the training dataset. In order to implement it, we first need to extract the entities from the sequences. We can do this using an off-the-shelf model for named entity recognition such as the `FLAIR` model (Akbik et al., 2018). Then, we create a dictionary of the named entities and in augmentation, we randomly choose one from the dictionary to replace with the original entity. One problem with this approach is that some sentences might not contain any entities. This will result in some sentences being repeated or discarded from the augmented batch.

Verb Replacement. Some verbs might be more indicative of the category to which they belong.

Original sentence	80% of people diagnosed with IBS have Sibo.
ER	100 percent of people diagnosed with IBS have Sibo.
VR (random)	80 % of people diagnosed with IBS cause Sibo .
VR (antonym)	80 % of people diagnosed with IBS abstain Sibo .
AEDA	80% of people diagnosed with IBS ! have Sibo.
YouChat	Only a small fraction of those diagnosed with IBS actually have Small Intestinal Bacterial Overgrowth (SIBO).

Table 5: An example augmentation from the `claims` class by the four augmentation methods, `ER` (entity replacement), `VR` (verb replacement), `AEDA` (an easier data augmentation).

To take advantage of this, we can create more diverse sequences by just replacing the verbs that are present in them. When replacing the verbs, we keep their tense intact. In addition, we experiment with two different ways for verb replacement. In one case, we first create a dictionary of verbs in the training data and select randomly from them when replacing a verb in a sentence. In the second case, we replace the verb with an antonym using WordNet (Miller, 1995). The reason for not using the training data is that antonyms are rare and they might not be found in the data.

Augmentation with YouChat. `YouChat` is a chatbot that can perform various guided actions such as augmenting sentences by producing contradictory ones. To do that we come up with a framing for our prompts that encourages diversity in the output as well as a contradiction. The reason for producing contradictory sentences is that, for the categories that, both the original and its contradiction can belong to the same category. For instance, for the `claims` category, if one sentence is considered to be a claim, then its contradiction can also be seen as a different claim.

We use two prompts for pushing `YouChat` to produce diverse and counterfactual sentences: 1) `Contradict this sentence with colorful words "original sentence"`, and 2) `Without using despite, while, and although, contradict this sentence with colorful words "original sentence"`. We use the first prompt to augment half of the sentences in the `claims` category. However, one problem that we notice with the outcome of this prompt is that after a couple of outputs, `YouChat` begins all sentences with expressions such as `although`, `despite`, and `while`. In order to change this, we augment the second half using the second prompt. This results in augmentations with different sentence structures.

Augmentation with Adversarial Examples. The BAT model (Karimi et al., 2021a) trains the pre-trained language model in an adversarial manner where adversarial examples are created during the training in the embedding space.

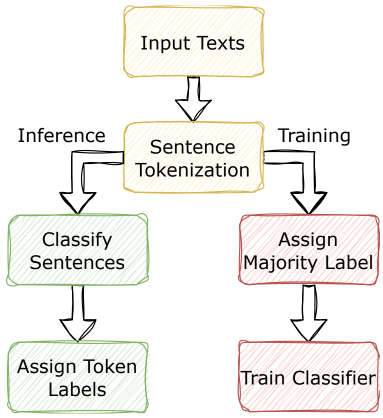

Figure 2: Workflow of our system[2]

Data	**CLA**	**EXP**	**O**	**PER**	**QUE**
Train	401	1917	19826	7824	5064
Dev	49	235	2666	995	633

Table 6: Number of sentences for each category after sentence tokenization in Subtask 1.

3 System Overview

The annotated data gives us the span of each category. The spans can be complete sentences or part of a sentence. One approach to address the task is to frame it as a token classification task, similar to named entity recognition tasks. However, named entities seem to be easier to spot because of their locality. On the contrary, given that a longer sequence of words could belong to a category, we take a broader look to recognize them. As a result, we formulate the problem as sentence classification. The dataset statistics indicate that in more than 87 percent of the resulting sentences, all the words

[2]Figure drawn using https://draw.io.

	Method	CLA	EXP	O	PER	QUE	Precision	Recall	F1
Baseline	CRF	11.1	12.7	76.5	49.5	71.0	51.1	42.1	44.1
	DistilBERT	25.8	32.0	77.6	56.8	76.4	57.9	52.0	53.7
400	CRF + ER	13.7	12.1	77.4	48.2	70.3	56.1	41.7	44.3
	CRF + VR (random)	11.3	15.1	76.4	50.2	69.5	50.5	42.4	44.5
	CRF + VR (antonym)	11.7	8.9	75.9	49.6	69.2	47.2	41.6	43.1
	CRF + AEDA	5.4	4.8	77.0	47.8	67.7	44.5	39.5	40.6
	CRF + YouChat	9.2	12.0	75.3	49.1	68.7	46.1	41.4	42.9
100	DB + ER	27.9	29.3	77.5	56.5	76.8	56.7	52.3	53.6
	DB + VR (random)	27.9	34.2	77.5	56.2	76.6	58.3	52.8	54.5
	DB + VR (antonym)	27.5	29.1	77.6	56.3	77.2	57.7	52.0	53.5
	DB + AEDA	25.5	28.5	77.8	56.7	76.9	56.3	51.9	53.1
	DB + YouChat	26.2	35.8	77.6	56.9	76.1	59.0	52.9	54.5
400	DB + ER	29.7	29.8	77.1	57.1	77.2	56.7	53.1	54.2
	DB + VR (random)	22.3	29.9	77.7	56.5	77.2	56.5	51.3	52.7
	DB + VR (antonym)	18.6	32.1	77.4	56.6	76.3	57.0	50.7	52.2
	DB + AEDA	28.9	35.2	77.6	56.7	76.7	59.1	53.1	55.0
	DB + YouChat	21.0	30.1	77.0	57.0	76.3	54.8	51.1	52.3

Table 7: Subtask 1. Experiments with 100 and 400 augmented samples for the claims class with DistilBERT (DB) and CRF models using Verb Replacement (VR), Entity Replacement (ER), AEDA, and YouChat augmentations. Green shows the best performer, blue is the second best, and red is the worst.

have the same label.

3.1 Workflow

As can be seen in Figure 2, we first split the texts into sentences using a sentence tokenizer toolkit from the NLTK library (Loper and Bird, 2002). Then, for training, we assign the label of the majority of the tokens to the sentence and train the model with the resulting data. For the inference part, after sentence tokenization, we classify each sentence using the trained model and assign the sentence label to the individual tokens.

3.2 Sentence-Tokenized Dataset Statistics

Separating the input texts into sentences results in just over 35K sentences which are distributed heavily in favor of the outside (O) class. Table 6 shows the statistics of the resulting data. As we can see, only a small proportion of the sentences belong to the claims category. We augmented the samples in this category to mitigate the class imbalance.

4 Baseline Models

We compare the performance of augmentation methods with two baseline models, namely a conditional random fields (Lafferty et al., 2001) model and the DistilBERT pre-trained language model for Subtask 1 (Sanh et al., 2019), and the BioBERT model (Lee et al., 2020) for Subtask 2.

Conditional Random Fields (CRF). This model is particularly suited for sequence labeling tasks. It considers a set of manually defined feature functions to predict the label of a token. In our case, we only consider some simple features such as the word itself, the word endings, whether the word is uppercase or lowercase, and if it is a number or not. With the same features, we also consider bigrams.

DistilBERT. This model is a lighter and more robust version of the BERT model (Devlin et al., 2019). We also consider the performance of this model without any augmentation as one of the baselines.

BioBERT. This is another variant of the BERT model that has been trained on medical texts in addition to the general text used for training BERT.

5 Results and Analysis

We perform augmentation on the claims class for Subtask 1 and the whole dataset for Subtask 2.

5.1 Subtask 1: Causal Claim Identification

For this task, the CRF model provides a relatively well-performing baseline despite its simplicity. Notably, from Table 7, we can see that it does well on the QUE class with 71 percent. This can be attributed to the fact that the CLA class is easier to

Method	CLA	EXP	O	PER	QUE	Precision	Recall	F1
DistilBERT	25.8	32.0	77.6	56.8	76.4	57.9	52.0	53.7
DB + ER	23.8	27.8	76.6	55.1	77.1	54.7	51.1	52.1
DB + VR (random)	15.4	36.5	77.7	56.3	77.4	57.7	51.2	52.7
DB + AEDA	24.4	32.1	76.9	55.6	76.9	57.7	51.6	53.2

Table 8: Subtask 1. Results with 4 augmentations for all `400` samples in `claims` class with `DistilBERT` (`DB`) using Verb Replacement (`VR`), Entity Replacement (`ER`), and `AEDA` methods.

detect although the number of samples in this category is a lot lower than the `PER` class that has a performance of `49.5` percent. Quite understandably, the lowest performing classes were the `CLA` and `EXP` classes, which can be due to having only a small number of samples in addition to their difficulty.

`DistilBERT`, on the other hand, shows an almost `10` percent overall improvement as well as on individual classes over the `CRF` model, which is expected given the large number of parameters it has compared to `CRF` (≈`68M` vs. ≈`10M`).

Impact of augmentation on CRF. The impact of augmentations with the `CRF` model is somewhat mixed. While `ER` helps the `claims` class improve by two percent, others show no improvement or negative improvement. It is possible that the `CRF` model is more vulnerable to out-of-distribution changes.

Impact of augmentation on DistilBERT. Considering the effect of the augmentation methods on the overall performance of `DistilBERT`, we experiment with two scenarios: first with `100` augmented sentences and then with `400` augmentations. Table 7 shows that, in the first case, `VR (random)` and `YouChat` have helped the model improve by almost one point while the `ER` method has had a slightly negative effect. The effect on the minority class, however, was positive for all the augmentation methods except for `AEDA`, with `ER` and `VR (random)` showing more than two percent improvement and `YouChat` `0.4` percent. With `400` augmentations, we see that only `ER` and `AEDA` have improved the class performance. This can be attributed to the increase in the amount of noise as we include more augmentations.

Impact of multiple augmentations. In this experiment, we investigate how multiple augmentations can impact the `DistilBERT` model on the studied dataset. Therefore, for each tokenized sentence, we produce four augmentations. We do this only for `ER`, `VR (random)`, and `AEDA` since for `YouChat` the manual work is time-consuming and for `VR (antonym)`, there is only one antonym for a verb. As we can see from Table 8, more augmentations of the `claims` class have a negative effect on the class itself while improving the results on the `EXP` (claim per experience) class. Given that this class also contains claims, it seems that more data for the `claims` class could also help `claims per experience` class.

Method	Precision	Recall	F1
BioBERT	47.2	11.7	18.8
BioBERT + ER	32.5	11.7	17.2
BioBERT + BAT	20.8	17.7	19.1
BioBERT + VR	25.7	16.4	20.1

Table 9: Subtask 2. Results with one augmentation using `ER`, `BAT`, and `VR` with `BioBERT`. We augment the entire dataset.

5.2 Subtask 2: PIO Frame Extraction

For this experiment, we utilized the `BioBERT` model (Lee et al., 2020) as the baseline. Table 9 shows the effect of three augmentation methods on this model with `100` examples augmented from the `claims` category. As we can see, overall, verb replacement is more effective than other methods although entity replacement makes more sense for this task since in `ER`, we increase the number of sentences using similar entities. This should provide a more diverse context for the existing entities in the training data.

6 Conclusion

We proposed verb replacement as a novel counterfactual data augmentation technique to increase the number of samples in the minority class for causal claim identification. Then, we showed that this method can significantly improve the performance of the machine learning model in the minority class. Comparing it with three other augmentation methods, we also found out that the proposed method can outperform them in some cases.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th international conference on computational linguistics*, pages 1638–1649.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

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.

Haibo He and Edwardo A Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284.

Matthew Herland, Richard A Bauder, and Taghi M Khoshgoftaar. 2017. Medical provider specialty predictions for the detection of anomalous medicare insurance claims. In *2017 IEEE international conference on information reuse and integration (IRI)*, pages 579–588. IEEE.

Akbar Karimi, Leonardo Rossi, and Andrea Prati. 2021a. Adversarial training for aspect-based sentiment analysis with bert. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 8797–8803. IEEE.

Akbar Karimi, Leonardo Rossi, and Andrea Prati. 2021b. AEDA: An easier data augmentation technique for text classification. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2748–2754, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vivek Khetan, Somin Wadhwa, Byron Wallace, and Silvio Amir. 2023. Semeval-2023 task 8: Causal medical claim identification and related pio frame extraction from social media posts. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Charles X Ling and Chenghui Li. 1998. Data mining for direct marketing: Problems and solutions. In *Kdd*, volume 98, pages 73–79.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pages 63–70.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Ipek Baris Schlicht, Lucie Flek, and Paolo Rosso. 2023. Multilingual detection of check-worthy claims using world languages and adapter fusion. *arXiv preprint arXiv:2301.05494*.

Mohammed Temraz and Mark T Keane. 2022. Solving the class imbalance problem using a counterfactual method for data augmentation. *Machine Learning with Applications*, 9:100375.

Somin Wadhwa, Vivek Khetan, Silvio Amir, and Byron Wallace. 2023. Redhot: A corpus of annotated medical questions, experiences, and claims on social media. In *European Association of Computational Linguistics (EACL)*.

Yibo Wang and Wei Xu. 2018. Leveraging deep learning with lda-based text analytics to detect automobile insurance fraud. *Decision Support Systems*, 105:87–95.

Xiangji Zeng, Yunliang Li, Yuchen Zhai, and Yin Zhang. 2020. Counterfactual generator: A weakly-supervised method for named entity recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7270–7280, Online. Association for Computational Linguistics.