Improving Dutch Vaccine Hesitancy Monitoring via Multi-Label Data Augmentation with GPT-3.5

Jens Van Nooten, Walter Daelemans
CLiPS (University of Antwerp)
Lange Winkelstraat 40, 2000 Antwerp (Belgium)
firstname.lastname@uantwerpen.be

Abstract

In this paper, we leverage the GPT-3.5 language model both using the Chat-GPT API interface and the GPT-3.5 API interface to generate realistic examples of anti-vaccination tweets in Dutch with the aim of augmenting an imbalanced multi-label vaccine hesitancy argumentation classification dataset. In line with previous research, we devise a prompt that, on the one hand, instructs the model to generate realistic examples based on the human dataset (gold standard) and, on the other hand, to assign one or multiple labels to the generated instances. We then augment our gold standard data with the generated examples and evaluate the impact thereof in a cross-validation setting with several state-of-the-art Dutch BERT models. This augmentation technique predominantly shows improvements in F1 for classifying underrepresented classes while increasing the overall recall, paired with a slight decrease in precision for more common classes. Furthermore, we examine how well the synthetic data generalises to human data in the classification task. To our knowledge, we are the first to utilise Chat-GPT and GPT-3.5 for augmenting a Dutch multi-label dataset classification task.

1 Introduction

In many text classification settings, the label distribution in datasets is imbalanced, which impacts the learning process of machine learning models and often leads to a degraded performance (Tarekegn et al., 2021). This phenomenon is especially inherent to multi-label datasets and therefore complicates multi-label classification further. Multiple approaches for tackling classification tasks where the data is imbalanced have been proposed. These approaches, which aim to alleviate the issue of class imbalance, comprise multiple categories and can be divided into data re-sampling, classifier adaptation and ensemble methods (Tarekegn et al., 2021). Bayer et al. (2022) describe multiple Data Augmentation (DA) methods, which encompass DA of the text itself at character, word, sentence/paragraph, or document level. These methods include noise induction by replacing and swapping characters or words, embedding replacement, replacing words with synonyms, translation, or even a combination of multiple methods.

With the popularisation of large generative models such as the GPT models (Radford et al., 2019), researchers have leveraged them for DA purposes, ranging from augmenting text translation datasets (Sawai et al., 2021), to generating examples for an Event Detection dataset (Pouran Ben Veyseh et al., 2021) and augmenting a COVID-19 Q&A dataset (Zhou and Zhang, 2021). However, most research is restricted to binary or multi-class classification, thus leaving much uncovered ground to research the potential benefits of augmenting multi-label data. Additionally, the cross-lingual capabilities of the State-of-the-Art (SotA) generative models for DA purposes remain under-explored. In this paper, we augment a Dutch dataset for vaccine hesitancy argumentation monitoring, as introduced in Lemmens et al. (2021) using SotA language models, including Chat-GPT, and build upon previous research by introducing these generative models in a realistic multi-label text classification setting. Additionally, we contextualise the performance with a strong back-translation baseline and extensively compare different GPT-based DA methods for multi-label text classification.

Our contributions are the following:

- We build upon previous research that utilises large language models to augment datasets by extending the SotA techniques to a multi-label setting and showing their effectiveness.
- We bring new insights into applying the SotA DA techniques for a real-world application of multi-label text classification.
- We extensively compare data generated by the two instances of the GPT-3.5 model.

Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 251–270

The present work is organised as follows: Section 2 provides an overview of research into data augmentation and the most recent generative models. Section 3 describes the datasets that are used for the experiments, including how the generative models were prompted, introduces the classification models used and outlines the evaluation methods. In Section 4, the results from the experiments are summarised and additional extensive analyses are conducted. Lastly, Section 5 briefly summarises the presented methodology and findings, along with outlining steps for future work.

2 Related Work

2.1 Generative Models

The GPT models are generative language models developed by OpenAI that have showcased an immense versatility in tasks that they can perform, ranging from classification and translation to summarisation (Radford et al., 2019). The models are auto-regressive, meaning that the models take previous outputs that they have generated into account for future outputs.

GPT-3.5 (or InstructGPT) is one of the most recent additions to OpenAI's roster of models, which is a fine-tuned version of GPT-3 using Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022). Chat-GPT[1] is the latest installment of GPT-3-based models and was released in December 2022. The web interface and API have already attracted many users and researchers for various use cases, such as text summarisation, creative and academic text generation, translation, and so on. Chat-GPT is quite similar to GPT-3.5, in that the model generates an answer, given an instruction as a prompt. Analogous to GPT-3.5's training method, Chat-GPT was trained using RLHF (Ouyang et al., 2022).

2.2 GPT Models as In-Context Learners

GPT models have already displayed impressive in-context learning and few-shot learning capabilities (Radford et al., 2019; Brown et al., 2020): based on a handful of gold standard examples that are included in the prompt, the models have shown SoTA performance on a variety of tasks. However, GPT-3's performance is heavily dependent on the selection of prompt examples, as mentioned in Liu et al. (2022) and Min et al. (2022). The former authors denote the in-context learning scenario with GPT-3 as a conditional text generation problem, where given context (which can include the task description and gold standard text examples), several target samples are generated. To find the optimal gold standard samples to include in the prompt as the context, the researchers leveraged RoBERTa (and variations thereof fine-tuned on NLI tasks) to embed training texts and retrieve the most similar ones to the test source. Their approach was applied to the SST-2 benchmark and showed an improvement of 6% accuracy compared to randomly selecting examples from the training data, in addition to superior performance compared to T5 on multiple Question-Answering datasets.

With these in-context learning capabilities in mind and taking inspiration from the MixUp learning technique in computer vision (Zhang et al., 2017), Yoo et al. (2021) introduced a multi-task prompt for GPT-3 that lets the model generate data and simultaneously labels the generated data. With the inclusion of examples from the training data in the prompt, the researchers aimed to generate examples that are close to the training distribution. With their multi-task approach, they reached superior results to other DA techniques on GLUE benchmark datasets.

In this paper, we implement a similar technique for augmenting a Dutch multi-label dataset, thus leveraging the in-context learning and cross-lingual abilities of Chat-GPT and GPT-3.5.

2.3 Data Augmentation

Data scarcity is a common problem in machine learning, and more so in commercial settings. This can manifest itself in data imbalance, leading to a degradation in performance on downstream classification tasks. DA aims to alleviate this problem by artificially enhancing the dataset through transformations or synthetic additions, while still preserving class information (Bayer et al., 2022). Consequently, DA also circumvents the high costs of human annotations (Bayer et al., 2022). More generally speaking, DA can be understood as "a strategy to prevent overfitting via regularization" (Shorten et al., 2021).

Bayer et al. (2022) describe three main categories of DA. First, augmentation in the data space, which is concerned with transformations of the data in its textual form on either character level, word level, phrase/sentence level, or document level. A wide array of different techniques and approaches

[1] https://openai.com/blog/chatgpt

have been proposed at each of these levels, such as noise induction in the shape of random swap and deletion of words (Wei and Zou, 2019). Other approaches include synonym replacement based on WordNet or thesauri (replacement by synonyms) (Kolomiyets et al., 2011; Li et al., 2017), embedding replacement (replacement by words with a similar latent representation) (Rizos et al., 2019) and back-translation (Xie et al., 2020). The latter describes the method of translating a text entry to another language and then translating it back to the source language, thus yielding a paraphrase or a slightly different variation of the original text and keeping the same gold-standard label(s).

A great deal of recent work on DA has explored the usage of generative models. Multiple researchers have explored SotA encoder/decoder models, such as BART (Kumar et al., 2020; Abonizio et al., 2022), while others focused on utilizing the popular GPT models for augmenting datasets, though few have leveraged them for augmenting multi-label datasets. For example, Zhang et al. (2020) have utilised GPT-2 for augmenting their dataset for extreme multilabel classification (XMC), a setting where a large pool of labels is available and to which data imbalance is inherent as a result. By comparing rule-based augmentation, Wordnet-based augmentation and augmentation based on text pairs with the same labels as a prompt in GPT-2, the authors found GPT-2 to be the most effective approach for improving the model's performance. Similarly, Dirting et al. (2022) generated synthetic Facebook comments that contain hate speech using GPT-2 to balance out a multi-label hate speech detection dataset, after which pseudo-labels were added to these comments by using the prediction of a fine-tuned BERT classifier.

3 Methodology

In this work, we aim to build upon existing research by augmenting a Dutch multi-label vaccine hesitancy detection dataset using novel generative models. Given the capabilities of the most recent models, we aim to generate realistic examples of vaccine-hesitant tweets to alleviate the data scarcity issue of the present application. We hypothesise that introducing new synthetic examples in the training data will improve the performance of large language models on the mentioned down-stream classification task.

3.1 Datasets

3.1.1 Vaccinpraat

For the experiments, we used the Vaccinpraat dataset for vaccine hesitancy argumentation classification, which was introduced in Lemmens et al. (2021). The dataset is used to classify the reason(s) why the writer of a tweet or Facebook comment is vaccine-hesitant. In total, there are nine labels: alternative medicine, conspiracy, criticism of vaccination strategy, development, efficacy, institutional motives, liberty, morality and safety. A description of these labels can be found in Lemmens et al. (2021). This dataset consists of 8,244 tweets and 3,917 Facebook comments in Dutch that are annotated with a binary label for vaccine stance (hesitant or not) and if the entry expresses vaccine hesitancy, the entries are annotated with one or multiple of the previously mentioned classes ("arguments"). For the present experiments, we only use the Twitter portion of the data [2].

The label distribution of this dataset and the generated datasets is visualised in Figure 1 and shows that the dataset is imbalanced, where the 'morality' and 'alternative medicine' are the least frequent classes. The dataset also contains 1,453 instances where no label is assigned.

3.1.2 Generated Datasets

Prompt Construction As mentioned previously, GPT-3.5 and Chat-GPT[3] were used as generative models. Taking inspiration from the method described in Yoo et al. (2021), an English prompt was devised that instructs the models to generate items in Dutch based on the description of labels in addition to examples from the gold standard data and assigns one or multiple labels to these items. The following paragraphs describe how the prompt was constructed.

The prompt itself consists of three main parts, namely the labels and their descriptions, the description of the task and the examples from the gold standard data.

- **Label Descriptions.** The label descriptions were retrieved from the Vaccinpraat website, which conform to the annotation guidelines for the gold standard dataset. These descriptions were then manually translated to En-

[2] This way we only have to generate synthetic tweets and do not have to mix different sources of text in training.

[3] We used the model provided in the API released on March 1 2023.

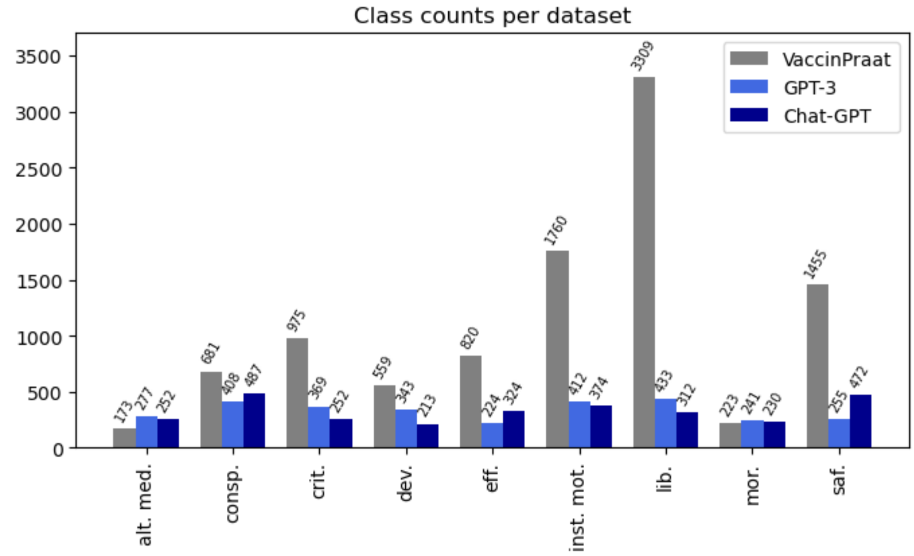


Figure 1: Distribution of the classes in the tweets portion of the Vaccinpraat dataset, in addition to the datasets generated with GPT-3.5 and Chat-GPT. The total number of instances does not correspond to the number of entries in the dataset, because the data is multi-label.

glish[4], after which clarifications of how the label is manifested in the data were added to the descriptions. In order to retrieve such clarifications, named entities, nouns and general phenomena for a certain label were selected manually from texts that belong to a certain label. For example, tweets with the label 'alternative medicine' often referred to ivermectine or hydroxochloroquine, so this was added to the label description as follows:

> "alternative_medicine. This label refers to alternatives to the COVID vaccine, such as natural remedies or other medicine, such as *ivermectine* or *hydroxychloroquine*."

- **Task Description.** The models were instructed (in English) to generate Dutch anti-vaccination tweets and label the data it generates at the same time. Additionally, we also prompt the model to generate instances that are similar to some provided examples.

- **Example Selection.** In order to select text-label pairs from the dataset, one multi-label and one single-label example were manually selected that represent the label(s) well to guide the model to produce text in the correct output format.

[4]Experiments were also conducted where the label descriptions were in Dutch, though this did not yield a significant difference in performance or class distribution.

Generative Model Parameters For GPT-3, we used the *text-davinci-003* completion model, for which the temperature parameter was set to .6 [5], while the 'top p', 'frequency penalty' and 'presence penalty' were all set to 1. The model generated 500 tokens maximum.

For Chat-GPT, the standard implementation of the *gpt-3.5-turbo* model was used. No additional contextual messages besides the prompt were added to the conversation as history. Details about the generated datasets can be found in Table 1. With Chat-GPT and GPT-3.5, 1,700 instances were generated using the same prompt as described in Section 3.1.2. GPT-3.5 generated shorter and more lexically diverse tweets, as indicated by the higher Type-Token Ratio (TTR). Examples of the generated datasets can be found in Appendix 6 and 7. These examples also showcase some differences between two versions of the generative model: though GPT-3.5 generates slightly more lexically diverse data, the data is sometimes ungrammatical.

Generated Data The distributions of the two synthetic datasets differ slightly, in that GPT-3.5 generated fewer instances for 'efficacy' and 'safety' than Chat-GPT, but more for the 'liberty' and 'safety' classes. Moreover, GPT-3.5 generated 80 unique label combinations in total, while Chat-GPT generated 72 unique combinations. In comparison, the

[5]We aimed for a balance between consistency and creativity. Any higher value for temperature yielded ungrammatical text after the model had generated a tweet.

Dataset	Length	TTR
GPT-3.5	27 (6)	.10
Chat-GPT	32 (7)	.07
Vaccinpraat	32 (14)	.07

Table 1: Average length of the generated data in tokens (with standard deviation) and average type-token ratio.

gold-standard data contains 104 unique label combinations. Additionally, GPT-3.5 generated slightly more examples with one assigned label than Chat-GPT (cf. Figure 2). The label co-occurrence matrices for each dataset can be found in Appendix 6 and 7.

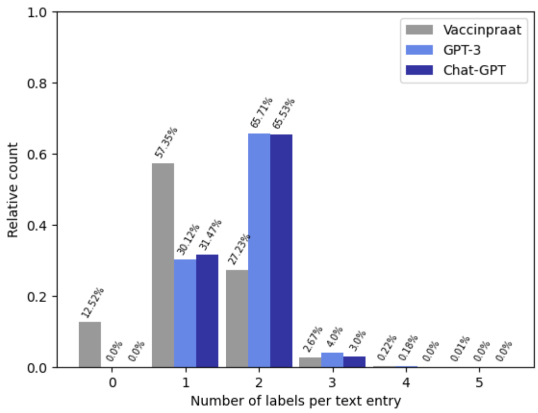


Figure 2: The number of labels per generated instance for both models.

3.2 Back-translation Baseline

As a baseline for the experiments, we opted for back-translation. The Dutch tweets were translated to English and then back to Dutch by employing an ALIGN transformers model (Garg et al., 2019) [6].

3.3 Classification Models

Several large language models that were trained on Dutch were utilised for the experiments:

1. **BERTje**: A Dutch version of BERT (de Vries et al., 2019).
2. **RobBERT-v2**: A Dutch version of RoBERTa (Delobelle et al., 2020).
3. **CoNTACT**: A domain-adapted version of RobBERT, specialised in COVID-related tweets (Lemmens et al., 2022).
4. **RobBERTje**: A distilled version of RobBERT (Delobelle et al., 2021).

[6]We used the Helsinki-NLP/opus-mt-en-nl and Helsinki-NLP/opus-mt-nl-en model on the HuggingFace hub.

All models were trained for 5 epochs and the best performing model based in terms of the loss on the validation set was chosen to make predictions on the test fold. For all models, a learning rate of 5e-5, a batch size of 16 and a maximum sequence length of 256 were used. As the input for the classification head of the models, we used the hidden state of the CLS token at the last layer of the model, thereby following the original implementation of BERT (Devlin et al., 2019). We used Binary Cross Entropy Loss (with a built-in sigmoid activation applied to the model logits), the AdamW optimizer and a learning rate scheduler with linear decay. These hyper-parameters remained consistent across all models.

3.4 Model Evaluation

The gold standard dataset was split into five stratified parts of equal length, out of which one fold was used as a test set for each repetition. The synthetic data was then added to the remaining training data. In order to prevent data leakage with the back-translation baseline, we selected 1,700 random samples from each training fold to be back-translated[7]. We repeated each experiment on each fold five times, each time with a different random seed. The performance of the models is reported in terms of the average precision, recall and F1 (all macro- and micro-averaged).

4 Results and Discussion

In this section, the results for the experiments are summarised. First, the effect of data augmentation will be discussed. Then, in order to gain more insight into the nature of the synthetic data and the generative models, additional sets of experiments were conducted. First, we perform a typicality study on the synthetic and human data. Then, we directly compare the data generated by Chat-GPT and the GPT-3.5 API by balancing the data and utilizing the balanced sets for the classification tasks in order to determine if there is a clear difference in quality between the two synthetic datasets. Additionally, we measure how consistent the generative models are in labeling the generated instances by training and testing RobBERT on the synthetic data only. Furthermore, we investigate how well the synthetic data generalises to human-written data by training models on the synthetic data only and

[7]Experiments were also conducted where all training data was back-translated.

testing on the cross-validation test sets. Finally, we evaluate the impact of the synthetic data in a cross-platform setting where the models are trained on Twitter data and tested on Facebook comments.

4.1 Effect of Data Augmentation

The results for each model are visualised in Figure 3[8]. The results per class for RobBERT can be found in Table 2 (more detailed results and the results for the other models can be found in Appendix 9 - 12). Overall, DA leads to a significant increase in performance across all models. Examining the results more closely, it can be observed that the performance increases are the most significant for under-represented classes in the dataset, such as 'alternative medicine', 'development' and 'criticism of vaccination strategy' (cf. Table 2). Most notably, the 'alternative medicine' class sees an increase of more than 40 F1 points across all models. Additionally, we observe a small, though statistically significant[9] increase for the 'morality' class for some models. However, it remains the most difficult class to classify for all models. The effect of DA is also less notable on classes that occur more frequently in the gold standard dataset, such as 'institutional motives' and 'liberty', where some models show no improvement or only a minor improvement[10]. In general, we observe a reduction in false positives and an increase in true positives across all classes. However, this is paired with an increase in false negatives as well for most classes (cf. Figure 12 - 16).

Even though the augmentation with generated data yields a significant performance increase for all models, back-translation yields a roughly equal performance to the GPT-based augmentation, except for RobBERT, where the GPT-based augmentation significantly outperforms back-translation. However, both methods seem to complement each other when they are combined, since this leads to the highest performance for almost every class across all models[11] (cf. Figure 9-12.).

[8]The complete experimental matrix can be found in Appendix 8.

[9]The statistical significance was calculated with the McNemar test.

[10]Future work could explore adjusting the prompt to generate more diverse examples for these classes.

[11]An even higher performance for all models is yielded when the entire training set is back-translated and the GPT-generated data is added. The results for these experiments can be found in Table 8.

4.1.1 Data Typicality

We also calculated the typicality for the synthetic and gold-standard datasets (Zhang, 1992). Typicality measures how well an instance represents instances from the same label in a dataset. This concept is especially meaningful for the synthetic datasets to measure how well the generative models can capture the patterns in the labels. Therefore, it serves as an indication for how prototypical the synthetic data is compared to itself and/or the human data. For this, we group the instances per label[12] and embed all instances with Sentence Transformers[13] (Reimers and Gurevych, 2019). Then, for each generated instance from dataset a, we average the cosine similarity (sim) between the generated instance and each entry from dataset a (or gold standard dataset b) belonging to the same label ($N_l^{(a)}$):

$$\frac{1}{N_l^{(a)}} \sum_{j=1}^{N_l^{(a)}} sim(\mathbf{a}_i, \mathbf{a}_j^{(l)})$$

This is then divided by the average cosine similarity between the generated instance and the cosine similarity with entries from the same dataset (or gold standard dataset b) from all other labels ($N_k^{(a)}$):

$$\frac{1}{(L-1)\sum_{k\neq l}^{L} N_k^{(b)}} \sum_{k\neq l}^{L} \sum_{j=1}^{N_k^{(a)}} sim(\mathbf{a}_i, \mathbf{a}_j^{(k)})$$

The results of this analysis are summarised in Table 3. On average, the instances from both synthetic datasets are more prototypical than the gold standard data (intra-dataset typicality), as can be derived from Table 1. Comparing the synthetic instances to the human data (inter-dataset typicality), the same observations are made. This could be explained by Chat-GPT and GPT-3.5 possibly not accessing knowledge about very specific events or political figures, thus generating more generic or prototypical texts compared to the gold standard. Ultimately, the generated data is a distillation of what the GPT models have been shown during training.

[12]Since this is a multi-label dataset, we copy a text n times for n number of labels assigned to that text.

[13]For this, the 'textgain/allnli-GroNLP-bert-base-dutch-cased' model on the Hugging Face hub was used.

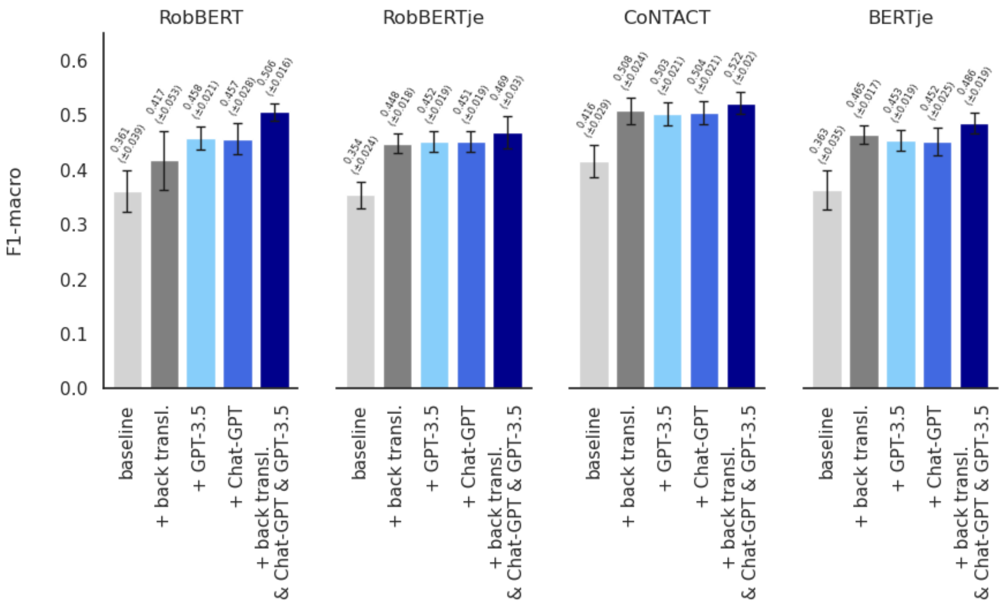


Figure 3: Performance (macro-averaged F1 scores) of the models, averaged across all folds and random seeds.

	Baseline	+ Back-transl.	+ GPT-3.5	+ Chat-GPT	+ Back-transl + GPT-3.5 + Chat-GPT
alt med	.01 (.02)	**.14** (.13)	**.5** (.09)	**.44** (.11)	**.49** (.08)
consp	.44 (.05)	**.48** (.04)	**.48** (.05)	**.49** (.06)	**.52** (.04)
crit	.15 (.1)	**.29** (.09)	**.26** (.06	**.28** (.06)	**.34** (.05)
dev	.2 (.14)	**.32** (.12)	**.36** (.1)	**.34** (.11)	**.42** (.08)
eff	.48 (.07)	**.51** (.09)	**.49** (.06)	**.51** (.06)	**.56** (.02)
inst	.58 (.05)	**.58** (.08)	.57 (.04)	**.58** (.02)	**.6** (.02)
lib	.76 (.02)	.75 (.02)	**.77** (.02)	**.76** (.02)	**.76** (.02)
mor	.0 (.0)	**.03** (.04)	**.06** (.07)	**.07** (.05)	**.19** (.05)
saf	.64 (.02)	**.65** (.02)	**.64** (.02)	**.64** (.02)	**.67** (.01)
micro avg.	.59 (.02)	**.6** (.02)	**.6** (.01)	**.6** (.01)	**.62** (.01)
macro avg.	.36 (.04)	**.42** (.05)	**.46** (.02)	**.46** (.02)	**.51** (.02)

Table 2: Classification results (F1) per class from RobBERT (averaged across folds and random seeds).

4.2 Comparing Chat-GPT and GPT-3.5

In order to provide a fair comparison between the two generated datasets to measure the text quality for the classification task, both the Chat-GPT and GPT-3 datasets were balanced by performing random undersampling so that each label occurs n times, where n is equal to the occurrence count of the least frequent class across both datasets. The undersampling in this case was necessary, as both generative models did not generate an equal distribution of data. In this case, both datasets were balanced to have 213 instances per label. The results of these supplemental experiments are summarised in the experimental matrix in Appendix 8. For all models, a statistically significant difference [14] in performance is observed between the two balanced setups, though the performance difference in terms of F1 between Chat-GPT and GPT-3 is negligible.

[14] $P < 0.001$ for RobBERT, RobBERTje and BERTje, $p = 0.03$ for ConTACT

4.3 Labeling Consistency

In order to validate whether the GPT models label the generated instances consistently, we conducted an extra set of cross-validation experiments on the synthetic data only. If there are noticeable label inconsistencies in the generated datasets, the standard deviation of the performance scores on the folds should be high. The results of these experiments are summarised in Table 4. Comparing Chat-GPT, GPT-3 and the gold standard data with each other, RobBERT shows a smaller standard deviation between folds when it is trained on the Chat-GPT dataset, thus suggesting that the labels are more consistent in this dataset than the other datasets.

Dataset	Intra-dataset typicality	Inter-dataset typicality	Cosine
GPT-3.5	1.23 (.19)	1.17 (.57)	.35 (.17)
Chat-GPT	1.22 (.22)	1.14 (.52)	.37 (.16)
Vaccinpraat	.34 (.17)	/	.49 (.19)

Table 3: Average inter-dataset/intra-dataset typicality and average cosine similarity to gold standard data. The latter is calculated the same way as typicality, though only the cosine similarity to instances with the same label is taken into account. Inter-typicality compares the text entries from a synthetic dataset to the human data, while intra-typicality compares instances from a dataset to instances from the same dataset.

Dataset	Mean F1	Std.
Vaccinpraat	.35	.034
GPT-3	.721	.018
Chat-GPT	.747	**.015**

Table 4: Mean performance (F1-macro) of RobBERT with standard deviations across folds.

Train data	Train samples	Mean F1
Vaccinpraat (sample)	1,700	.21 (.044)
GPT-3	1,700	.28 (.019)
Chat-GPT	1,700	.29 (.013)
Vaccinpraat (sample)	3,400	.324 (.05)
GPT-3 + Chat-GPT	3,400	**.349** **(.015)**

Table 5: Mean F1-macro of RobBERT trained on the synthetic datasets and tested on gold standard data.

4.4 Generalisability of Synthetic Data

Additional cross-validation experiments were conducted to measure how well the synthetic data generalises to the gold-standard data. For this, we trained the models on the synthetic datasets separately and a combination of the two, after which they were tested on the test folds of the gold standard data. In order to compare the datasets properly, a sample of 1,700 and 3,400 was taken from each gold standard training fold, which is equal to the number of entries in the synthetic dataset. The results of these experiments, as summarised in Table 5, indicate that with an equal number of samples, the synthetic data yield a better downstream performance than the gold standard data with an equal number of samples, with statistical significance for all experiments (p<0.001).

The performance difference could be attributed to the prototypical nature of the synthetic data, as mentioned in Section 3. As opposed to the synthetic examples, the human data is "atypical" in nature, suggesting that a great deal of the data is a very specific manifestation of the label(s) that the data is assigned. This is crystallized in the training data as implicit language or references to very specific events, people or other entities related to the COVID-19 pandemic. Using only prototypical data during training in this specific setting seems to guide the model to understand the test examples more quickly.

4.5 Effect of Prompt Examples

Two additional datasets were also generated where the prompt included ten examples from the gold standard data. Comparing this data to the previously discussed synthetic datasets, the distribution is quite different, as is visualised in Figure 4. Moreover, the results from the experiments conducted on these new datasets, which are summarised in Table 8, demonstrate that the dataset with more gold standard examples in the prompt yields worse results. This is most likely caused by difference in the class distribution between the datasets. For instance, Chat-GPT and GPT-3 generated more examples for the 'morality' and 'development' classes when less examples were provided in the prompt, thus leading to a performance increase on those classes.

4.6 Cross-platform Performance

Experiments were also conducted in a cross-platform setting where the model was trained on Twitter data (with or without augmentation) and tested on the Facebook portion of the dataset, which consists of 3,917 comments. For each augmentation method, we used 1,700 synthetic examples in addition to the gold-standard data, as described previously. The results (cf. Figure 5) indicate that all augmentation methods contribute to higher performance in a cross-platform setting, though a combination of all methods yields the best performance.

5 Conclusion

In this paper, we leveraged two instances of the GPT-3.5 model for augmenting a Dutch multi-label anti-vaccination dataset. Using these models for generating data and adding pseudo-labels, we evaluated the impact of this augmentation method by

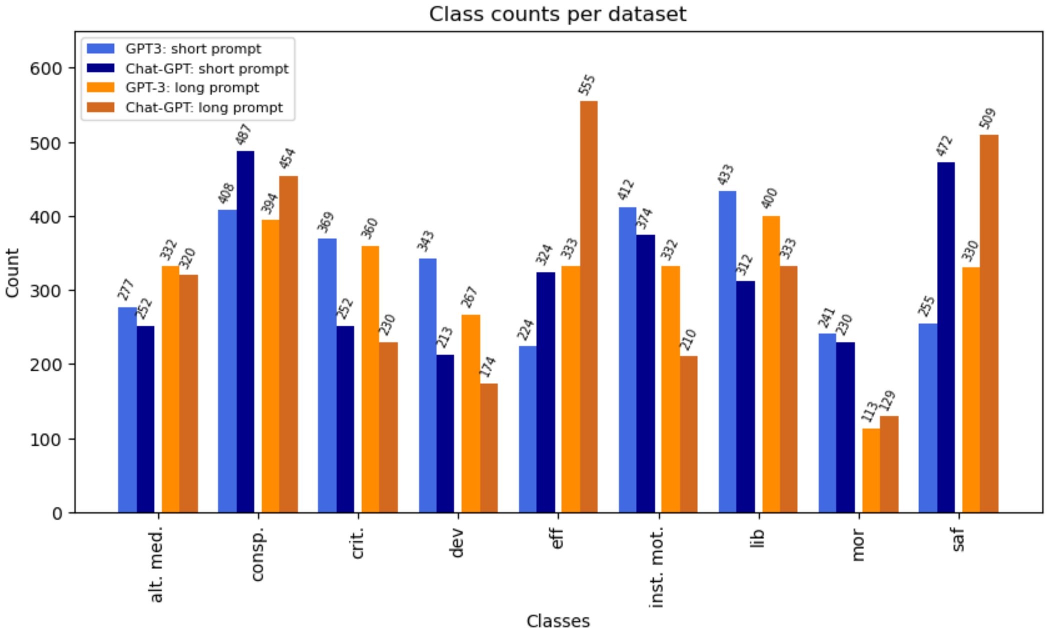


Figure 4: Label distribution of the synthetic datasets when generated with two gold-standard examples or ten in the prompt.

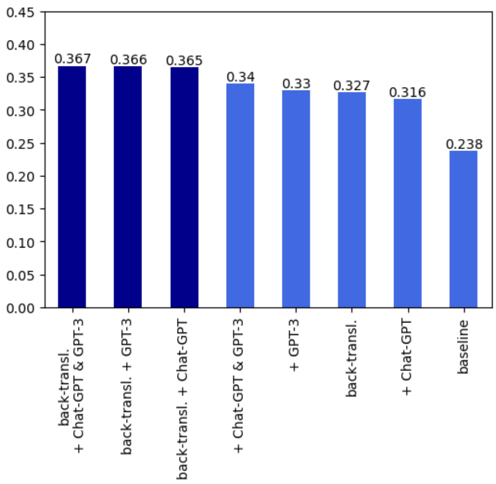


Figure 5: Results (macro-averaged F1) from the cross-platform experiments with RobBERT. The dark blue bars indicate that these results are from models trained on more data compared to the other results. The higher results are therefore attributed to a higher number of training samples.

comparing the performance of multiple language models on a multi-label downstream classification task. We show that the present method can lead to a significant performance increase, especially for underrepresented classes, thus highlighting the potential of the present approach for other applications. Moreover, we demonstrate that the class distribution of the generated datasets depends on the model instance that is used and the number of gold standard examples in the prompt itself. We compared the performance of the proposed method to back-translation, a strong baseline method for performing DA that achieves a roughly equal performance to GPT-based augmentation. Furthermore, we found that a combination of back-translation and GPT-based augmentation was the most effective for our application. With research into GPT models still continuing, future work should extend this method to other multi-label tasks, as data scarcity and data imbalance are inherent to this classification problem. Additionally, while the GPT models have the potential to majorly accelerate progress in the field of AI, the inner workings of the GPT models lack transparency, thus undermining attempts at versioning and replicating results. We therefore believe that providing powerful, open-access models are vital to progress in the field.

6 Limitations

Some limitations of the present study should be considered. First and foremost, the inherent statistical biases of Chat-GPT and GPT-3.5 might skew the data distribution, which is difficult to control without knowing what data the models are trained on. It is therefore also certainly possible that the generative models have already been trained on the Vaccinpraat dataset. Additionally, we found that the text examples in the prompts also affect the distribution. Moreover, these statistical biases might lead to repetitive sentence structures in the data.

Second, the generated data contains false information about vaccines and COVID. One should therefore act with caution when interpreting the synthetic data and should only consult fact-checked sources for information about COVID vaccines. Moreover, while the messages are believable enough as a reader to be vaccine-hesitant, the messages are more "neutral" in nature than the gold-standard data. This is especially apparent in the Chat-GPT dataset, which was to be expected because of the guardrails imposed on the model. This distribution shift could explain the degraded performance compared to the back-translated data in the in-platform. Future work could explore tuning the prompts further to minimize this distribution shift.

Third, since we only focused on the vaccine hesitancy monitoring task, more research should be conducted with the presented method for more multi-label tasks. However, this method could only work effectively for datasets with a relatively small number of labels, as the descriptions need to fit in the prompt. However, the promising results from the conducted experiments and analyses should stimulate further exploration for other multi-label text classification tasks.

Acknowledgements

This research was carried out in the framework of the EOSC Future project (Grant agreement ID: 101017536), funded under EXCELLENT SCIENCE - Research Infrastructures as part of the Horizon Europe 2020 Research Program.

References

Hugo Queiroz Abonizio, Emerson Cabrera Paraiso, and Sylvio Barbon. 2022. Toward text data augmentation for sentiment analysis. *IEEE Transactions on Artificial Intelligence*, 3(5):657–668.

Markus Bayer, Marc-André Kaufhold, and Christian Reuter. 2022. A survey on data augmentation for text classification. *ACM Comput. Surv.*, 55(7).

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A dutch BERT model.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265, Online. Association for Computational Linguistics.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2021. Robbertje: A distilled dutch bert model. *Computational Linguistics in the Netherlands Journal*, 11:125–140.

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.

Bakwa Dunka Dirting, Gloria A. Chukwudebe, Euphemia Chioma Nwokorie, and Ikechukwu Ignatius Ayogu. 2022. Multi-label classification of hate speech severity on social media using bert model. In *2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development (NIGERCON)*, pages 1–5.

Sarthak Garg, Stephan Peitz, Udhyakumar Nallasamy, and Matthias Paulik. 2019. Jointly learning to align and translate with transformer models. 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 4453–4462, Hong Kong, China. Association for Computational Linguistics.

Oleksandr Kolomiyets, Steven Bethard, and Marie-Francine Moens. 2011. Model-portability experiments for textual temporal analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 271–276, Portland, Oregon, USA. Association for Computational Linguistics.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26, Suzhou, China. Association for Computational Linguistics.

Jens Lemmens, Tess Dejaeghere, Tim Kreutz, Jens Van Nooten, Ilia Markov, and Walter Daelemans. 2021. Vaccinpraat: Monitoring vaccine skepticism in dutch twitter and facebook comments. *Computational Linguistics in the Netherlands Journal*, 11:173–188.

Jens Lemmens, Jens Van Nooten, Tim Kreutz, and Walter Daelemans. 2022. CoNTACT: A Dutch COVID-19 adapted BERT for vaccine hesitancy and argumentation detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6837–6845, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Yitong Li, Trevor Cohn, and Timothy Baldwin. 2017. Robust training under linguistic adversity. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 21–27, Valencia, Spain. Association for Computational Linguistics.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work?

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Amir Pouran Ben Veyseh, Viet Lai, Franck Dernoncourt, and Thien Huu Nguyen. 2021. Unleash GPT-2 power for event detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6271–6282, Online. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084.

Georgios Rizos, Konstantin Hemker, and Björn Schuller. 2019. Augment to prevent: Short-text data augmentation in deep learning for hate-speech classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 991–1000, New York, NY, USA. Association for Computing Machinery.

Ranto Sawai, Incheon Paik, and Ayato Kuwana. 2021. Sentence augmentation for language translation using gpt-2. *Electronics*, 10(24).

Connor Shorten, Taghi M Khoshgoftaar, and Borko Furht. 2021. Text data augmentation for deep learning. *Journal of big Data*, 8:1–34.

Adane Nega Tarekegn, Mario Giacobini, and Krzysztof Michalak. 2021. A review of methods for imbalanced multi-label classification. *Pattern Recognition*, 118:107965.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. 2020. Unsupervised data augmentation for consistency training. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. 2021. GPT3Mix: Leveraging large-scale language models for text augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2225–2239, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Danqing Zhang, Tao Li, Haiyang Zhang, and Bing Yin. 2020. On data augmentation for extreme multi-label classification.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2017. Mixup: Beyond empirical risk minimization.

Jianping Zhang. 1992. Selecting typical instances in instance-based learning. In Derek Sleeman and Peter Edwards, editors, *Machine Learning Proceedings 1992*, pages 470–479. Morgan Kaufmann, San Francisco (CA).

Shuohua Zhou and Yanping Zhang. 2021. Datlmedqa: A data augmentation and transfer learning based solution for medical question answering. *Applied Sciences*, 11(23).

A Appendix

Given are the following labels:
1. 'safety'. This label is concerned with the safety of the vaccines, side effects and presumable vaccine deaths.
2. 'liberty'. This label refers to the infringement of individual rights, freedoms and liberty with regards to mandatory vaccinations, tracking apps and vaccination certificates.
3. 'efficacy'. This label refers to the efficacy of vaccines, whether they are really needed and if they actually help combatting the coronavirus.
4. 'development'. This label details concerns about the rapid development, research, (lack of) transparency into the development of the vaccines and the supply of vaccines.
5. 'alternative_medicine'. This label refers to alternatives to the COVID vaccine, such as natural remedies or other medicine, such as ivermectine or hydroxychloroquine.
6. 'institutional_motives'. This label refers to hidden motives from companies, politicians, billionaires, or others to gain political or economic power.
7. 'conspiracy'. This label details conspiracy theories surrounding the vaccine, such as the Great Reset, control, genocide, chips in the vaccines, fake infection numbers, propaganda, covid as a fake virus, corruption, and so on.
8. 'criticism_of_vaccination_strategy'. This label is concerned with the waiting times for getting vaccinated and other groups of people getting vaccinated earlier or later than others.
9. 'morality'. Ethics, morals, discrimination of people who refuse the vaccine, references to God, references to religious texts and religion, people as lab rats, for example.

Generate 15 stylistically vastly different, realistic examples of anti-vaccination tweets in Dutch and Flemish that could be assigned one or multiple of those labels. **Label** the tweets with one or multiple of the labels that suit the tweets the best. The tweets should be similar to the examples below:

1. ['efficacy'] Vaccinatie help niet tegen covid. Niet voor de gevaccineerde en niet voor de niet gevaccineerde. Hij als viroloog zou daarvan op de hoogte moeten zijn!
2. ['institutional_motives', 'safety', 'liberty'] gogo denmark!! weg met de coronamaatregelen!! #vrijheid #vaccinatieplicht #mondkapjesplicht #vaccin #coronamaatregelen #demonstratie #kickoutrutte #kickoutdejonge #dictatuur #bigpharma #pcrtest #pcrgate
4. ['liberty', 'morality'] ik ben hier echt sprakeloos van. compleet tegen het zelfbeschikkingsrecht in. als mens bepaal jij wat en of er iets in je lichaam gespoten wordt. hoe komen we hier ooit uit mensen? #vaccinatiepaspoort #vaccinatieplicht #testenvoortoegang
5. ['criticism_of_vaccination_strategy', 'development'] als denemarken helemaal stopt met het astrazeneca vaccin, kunnen wij die doses dan overnemen? dan kunnen wij het priktempo flink opvoeren. #astrazeneca #ikwildieprik
6. ['conspiracy'] dat komt omdat testen stap 1 is. stap 2 is het vaccinatiepaspoort. het einddoel is stap 3, een chip bij iedereen onder de huid. noem mij maar gek, de tijd zal het leren. #testmaatschappij #vaccinatiepaspoort
7. ['institutional_motives'] derde shot voor de gelovige tv kijkertjes in het najaar. gaat lekker met die mega biljoenen #vaccinatie handel. scheelt toch als je commerciëel medewerkers op tv mogen als 'experts' en 'virologen'. gaat helemaal goed komen. #pfizer #astrazeneca #vaccin #vaccinatiepaspoort
8. ['development'] beschaamd in jullie plaats. door jullie laks aankoopbeleid lopen de vaccinaties achter met verstrekkende gevolgen en veel menselijk leed. de afrekening volgt bij de volgende verkiezingen.
9. ['morality', 'safety'] #vaccine #vaccinatiebewijs #vaccinatieplicht #vaccination #covid #coronadebat het is een unethical experiment in strijd met de nuremberg code en dus een oorlogsmisdaad
10. ['morality'] je bent een valse profeet en god zal iedereen straffen die zich vaccineert. #ikvaccineerniet

Figure 6: The complete prompt that was used for generating data with the label descriptions (blue), the task (green) and gold standard examples (orange). The examples marked in bold are the examples that are used for the main experiments.

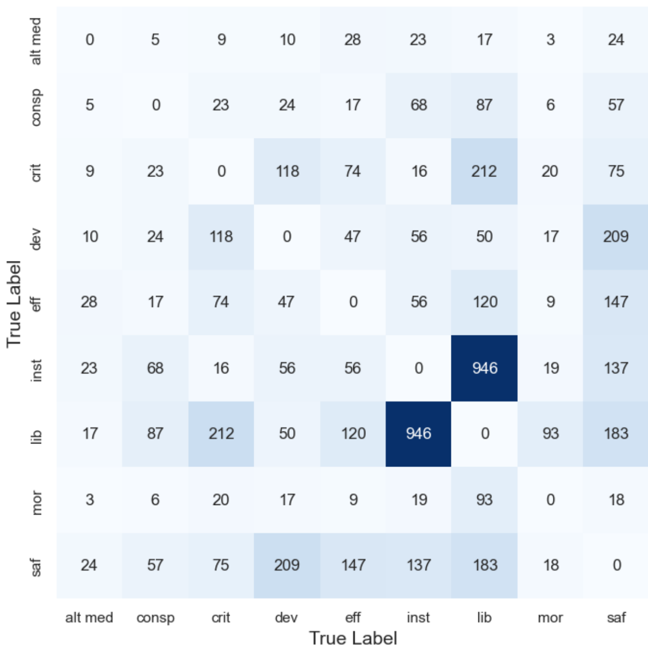


Figure 7: Label co-occurrence matrix from the Vaccinpraat dataset.

GPT-3.5

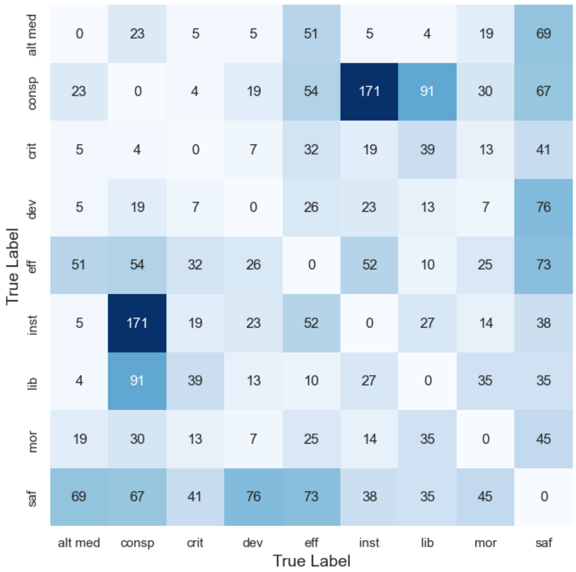
Text	Classes
Allemaal naar buiten met hun verhalende propaganda dat we allemaal eerlijk moeten vaccineren terwille van 'de gemeenschap'... Maar wat als ze onze kinderen letterlijk dodelijk experimenteren? #moraal	morality, safety
#vaccinatie #antivaccinatie Er zijn natuurlijke remedies die veel effectiever en veiliger zijn dan de snel ontwikkelde COVID19-vaccins!	development, alternative_medicine
Er zijn genoeg reden waarom sommige mensen twee keer denken voordat ze besluit om een covid-19 vaccin te krijgen omdat de effectiviteit hiervan tot dusverre betwistbaar is.#effectiefdebat	efficacy
Het coronavaccin wordt gebruikt om een wereldwijd controlesysteem op te zetten voor de elite! We moeten hiertegen in verzet komen!	conspiracy, institutional_motives
Waarom wordt er gezwegen over alternatieve therapieën die misschien veel beter werken dan vaccins? Terwijl andere landen al minder drastische maatregelen treffen?! #vaccinatiemachtiging	criticism_of_vaccination_strategy, alternative_medicine
Hoe kunnen ze zeggen dat we dit covid-19 vaccin moeten gebruiken wanneer er zo snel is ontwikkeld en er gezondheidsrisico's aan verbonden zijn?	development
Hydroxychloroquine of HCQ staat erom bekend effectief te zijn in het behandelen/voorkomen van Covid-19 symptomen - vergeleken met Vaccines die meestal alleen symptomen maskeren!	alternative_medicine
Covid-certificaten?? Wat volgt er nog meer?? We worden allemaal geruisloos ingeperkt door Big Brother!!! Stop met die complottheorieën!!!!	conspiracy, liberty
Is het wel moreel verantwoord dat we experimenteren met het menselijk lichaam in strijd met Gods woord? #Godverbiedhet	morality, safety
Er was gezegd dat na 6 weken na 1e prik je bent ingeënt, maar nu blijkt dat je ook na 12 weken nog niet 100% immuun bent.. De informatie die door overheidsinstanties verspreiden is vaak (te) summier & valse hoop creërend !	development, criticism_of_vaccination_strategy, safety
Niet alle landgenoten zulln bereid zijn om hun persoonlijke rechten op te offeren tengunste van eenzelfde vaccinatieplicht voor alles wat vel theorisch mogelijk covid-19 verspreiding kan stopppen. #vrijheidsontneming	liberty, morality
De échte agenda achter de Covid-19 pandemie: politieke machtsgreep, technologische censuur en economische transformatie. #greatreset #covidsamenzwering	conspiracy, institutional_motives
Er werd in 6 maand tijd eenzelfde vaccin uiteindelijk op 8 verschillende manieren toegepast... Wat was hun doel daarin?? Alleen winstbejag?? Kritisch denken vergt durf!!	development, institutional_motives
Wordt covid gebruikt als een tool om mensen te controleren via chips in het vaccin? Of is dit slechts een theorie zonder bewijs? #covidchipsconspiracy	conspiracy
Vaccins bevatten giftige stoffen zoals aluminium waardoor ze schadelijk kunnen zijn voor je gezondheid.. Kies liever andere manier om je immuunsysteem te boosteren dan vaccineren!	safety

Table 6: Text examples (with assigned pseudo-labels) generated by GPT-3.5.

Chat-GPT

Text	**Classes**
Vaccinaties zijn een inbreuk op mijn vrijheid. Het proces van ontwikkeling van vaccins is te snel gegaan. Ik vertrouw de ontwikkelaars niet en ik vertrouw de overheid al helemaal niet.	development, liberty
Dit hele #COVID19 verhaal is opgezet om ons te onderdrukken en te beheersen. Het vaccin is een manier om ons te tracker, onze vrijheid te ontnemen en wereldwijde controle te verwerven. #GreatReset #complot	conspiracy, institutional_motives
Waarom worden risicogroepen niet eerst gevaccineerd? Die hebben het vaccin toch het meest nodig? En is het vaccin eigenlijk wel echt nodig?	efficacy, criticism_of_vaccination_strategy
De farmaceutische industrie en de overheid hebben samengewerkt om de covid-pandemie te creëren zodat ze geld kunnen verdienen aan de verkoop van het vaccin. Dit kan niet anders dan een complot zijn!	conspiracy, development
Waarom hebben we zo lang gewacht op vaccinatie? Waarom is de media zo stil over de echte effecten van vaccins? Wie probeert ons te manipuleren? #vaccins #coronavirus #institutional_motives	institutional_motives
Het is onze keuze om ons wel of niet te laten vaccineren en niemand kan ons dat afnemen! #mybodymychoice	liberty
Het is duidelijk dat deze hele pandemie is opgezet om de bevolking te onderwerpen en vaccinatie is slechts een middel om dat te bereiken. Het is tijd om wakker te worden en te vechten terug! #COVID19 #vaccinatie #conspiracy	conspiracy, efficacy
Al die bijwerkingen en doden door de vaccins, dat kan toch geen toeval zijn? Ik vertrouw het niet.	efficacy, safety
Niemand weet wat er precies in dat vaccin zit en wat voor effecten het op lange termijn heeft. #transparantie #vaccinatie	development
Waarom zou ik een vaccin nemen als er zoveel natuurlijke alternatieven zijn om mijn immuunsysteem te versterken? Ik kies voor natuurlijke remedies zoals vitamines en kruiden om gezond te blijven.	alternative_medicine
Hoe weten we zeker dat het vaccin veilig is als het zo snel is ontwikkeld? We weten niet wat de lange termijn effecten kunnen zijn. Het is belangrijk dat er meer transparantie is over de ontwikkeling en onderzoek.	development, safety
De vaccins worden gepromoot als een 'oplossing' voor covid-19, maar er zijn talloze gevallen waarin mensen die al gevaccineerd zijn, nog steeds hebben getest op het virus. Dit werkt niet en we moeten een beter plan hebben.	efficacy, criticism_of_vaccination_strategy
De bijwerkingen van het vaccin zijn enorm. Het is niet veilig om jezelf te laten vaccineren. #vaccine #covid19 #sideeffects #safetyconcerns	efficacy, safety
De vaccins zijn niet alleen nutteloos, maar schaden ook je gezondheid! Het gaat om vertrouwen in je lichaam! #eigenkeuze #zelfhelendvermogen	efficacy, morality
Het is een schande dat onze fundamentele vrijheden worden afgenomen door verplichte vaccinaties en tracking apps. We moeten ons verzetten tegen deze inbreuk op onze burgerrechten. #vaccine #coronavirus #liberty #freedom	liberty, morality

Table 7: Text examples (with assigned classes) generated by Chat-GPT.



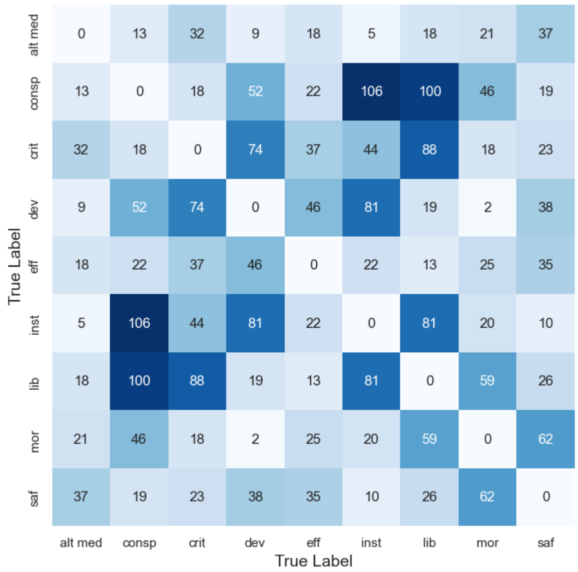
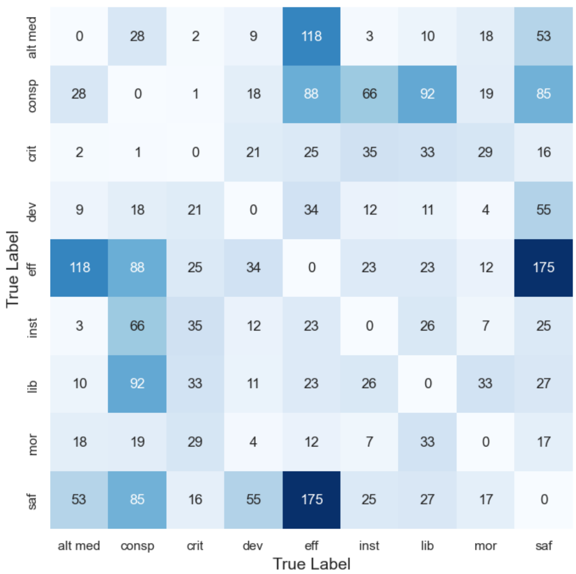


Figure 8: Label co-occurrence matrix from the ChatGPT dataset.

Figure 9: Label co-occurrence matrix from the GPT-3.5 dataset.



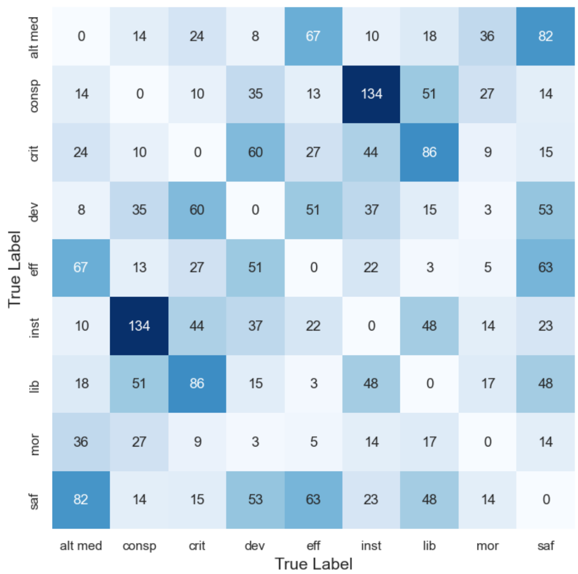


Figure 10: Label co-occurrence matrix from the ChatGPT dataset generated with more text examples in the prompt.

Figure 11: Label co-occurrence matrix from the GPT-3.5 dataset generated with more text examples in the prompt.

	RobBERT			RobBERTje			CoNTACT			BERTje		
setup	**prec.**	**rec.**	**F1**	**prec.**	**rec.**	**F1**	**prec.**	**rec.**	**F1**	**prec.**	**rec.**	**F1**
baseline	.49 (.05)	.32 (.04)	.36 (.04)	.5 (.05)	.31 (.03)	.35 (.02)	.53 (.05)	.37 (.03)	.42 (.03)	.51 (.03)	.32 (.04)	.36 (.04)
back-translation	.54 (.08)	.37 (.05)	.42 (.05)	.55 (.05)	.4 (.02)	.45 (.02)	.58 (.04)	.47 (.02)	.51 (.02)	.57 (.04)	.42 (.02)	.46 (.02)
GPT-3.5 (short prompt)	.58 (.03)	.4 (.03)	.46 (.02)	.58 (.03)	.4 (.02)	.45 (.02)	.61 (.03)	.46 (.03)	.5 (.02)	.6 (.04)	.4 (.03)	.45 (.02)
GPT-3.5 (short prompt; balanced)	.58 (.05)	.39 (.05)	.44 (.05)	.59 (.03)	.4 (.02)	.45 (.02)	.6 (.03)	.45 (.03)	.5 (.02)	.59 (.05)	.4 (.02)	.45 (.02)
GPT-3.5 (long prompt)	.55 (.04)	.38 (.05)	.43 (.05)	.56 (.02)	.39 (.02)	.44 (.01)	.6 (.04)	.45 (.03)	.49 (.02)	.56 (.02)	.39 (.02)	.44 (.02)
GPT-3.5 (long prompt; balanced)	.55 (.07)	.36 (.06)	.41 (.06)	.57 (.03)	.37 (.03)	.42 (.03)	.61 (.04)	.44 (.03)	.49 (.02)	.58 (.04)	.37 (.03)	.42 (.03)
Chat-GPT (short prompt)	.58 (.03)	.41 (.03)	.46 (.03)	.59 (.04)	.39 (.02)	.45 (.02)	.61 (.03)	.46 (.03)	.5 (.02)	.61 (.04)	.4 (.03)	.45 (.03)
Chat-GPT (short prompt; balanced)	.56 (.03)	.38 (.05)	.44 (.04)	.59 (.03)	.38 (.03)	.44 (.03)	.59 (.03)	.44 (.03)	.49 (.02)	.57 (.04)	.37 (.03)	.42 (.03)
Chat-GPT (long prompt)	.56 (.03)	.4 (.03)	.44 (.02)	.56 (.03)	.39 (.02)	.44 (.02)	.59 (.03)	.44 (.03)	.49 (.02)	.56 (.02)	.39 (.02)	.44 (.02)
Chat-GPT (long prompt; balanced)	.56 (.08)	.38 (.07)	.44 (.07)	.58 (.05)	.37 (.02)	.42 (.02)	.62 (.03)	.44 (.03)	.49 (.03)	.59 (.03)	.38 (.03)	.44 (.02)
GPT-3.5 + Chat-GPT	.59 (.02)	.42 (.04)	.47 (.03	.59 (.02)	.42 (.03)	.47 (.03)	.61 (.02)	.46 (.04)	.51 (.03)	.6 (.03)	.43 (.03)	.48 (.02)
back-translation + GPT-3.5 (short prompt)	.57 (.04)	.45 (.04)	.49 (.04)	.59 (.03)	.41 (.03)	.46 (.02)	.6 (.03)	.47 (.03)	.51 (.02)	.58 (.04)	.42 (.02)	.47 (.02)
back-translation + Chat-GPT (short prompt)	.57 (.03)	.44 (.04)	.49 (.04)	.59 (.03)	.42 (.02)	.47 (.02)	.61 (.03)	.46 (.02)	.51 (.02)	.58 (.03)	.42 (.03)	.47 (.02)
back-translation + Chat-GPT(short prompt) + GPT-3.5 (short prompt)	.58 (.02)	.47 (.02)	.51 (.02)	.6 (.04)	.42 (.03)	.47 (.03)	.62 (.02)	.48 (.03)	.52 (.02)	.58 (.02)	.44 (.03)	.49 (.02)
back-translation (all)	.56 (.03)	.44 (.02)	.48 (.02)	.56 (.02)	.44 (.01)	.49 (.01)	.58 (.02)	.5 (.01)	.53 (.01)	.57 (.02)	.46 (.01)	.5 (.01)
back-translation (all) + Chat-GPT(short prompt) + GPT-3.5 (short prompt)	**.57** (.02)	**.49** (.02)	**.52** (.01)	**.56** (.02)	**.48** (.01)	**.51** (.01)	**.59** (.02)	**.51** (.01)	**.54** (.01)	**.57** (.02)	**.48** (.02)	**.52** (.01)

Table 8: Experimental matrix with macro-averaged results per model. The results are averaged across all 25 experiments per setting per model. 'Backtranslated (all)' refers to experiments where the entire training set was back-translated, as opposed to the other back-translation setup where 1,700 random instances from the training set were back-translated.

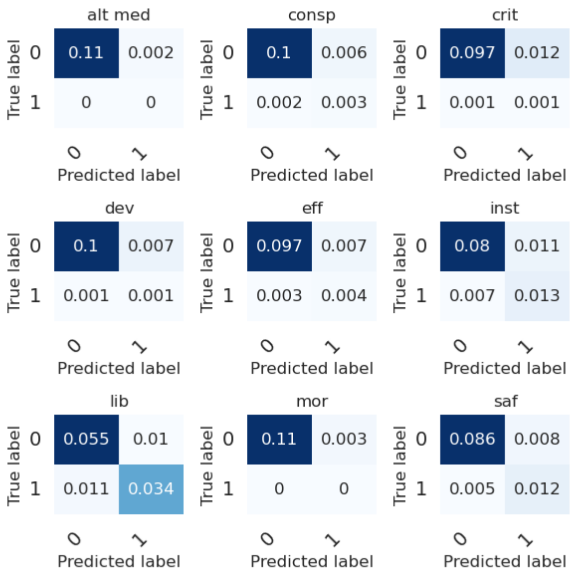


Figure 12: Confusion matrix from the baseline RobBERT model.

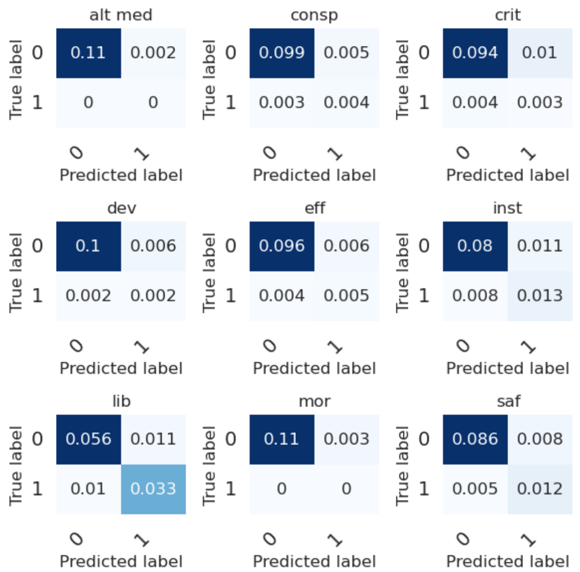


Figure 13: Confusion matrix from RobBERT trained on augmented data (back-translation).

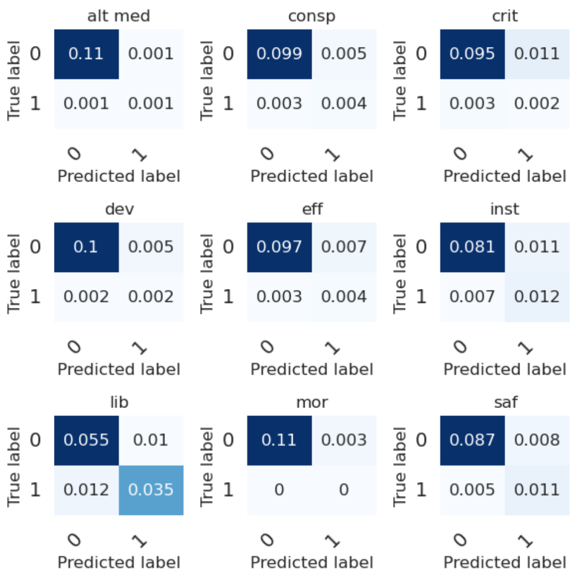


Figure 14: Confusion matrix from RobBERT trained on augmented data (GPT-3.5)

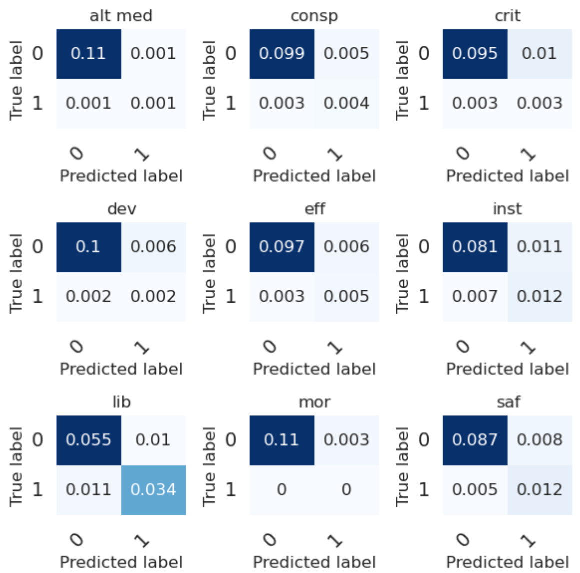


Figure 15: Confusion matrix from RobBERT trained on augmented data (Chat-GPT).

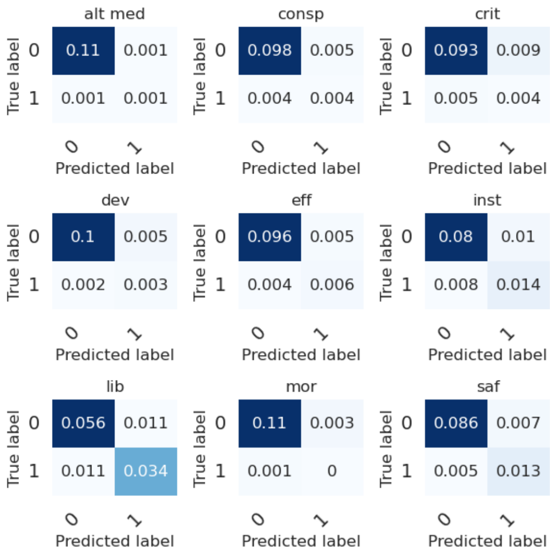


Figure 16: Confusion matrix from RobBERT trained on augmented data (back-translation + Chat-GPT + GPT-3.5)

	Baseline			+ Back-transl.			+ GPT-3.5			+ Chat-GPT			+ Back-transl + Chat-GPT +GPT-3.5		
	prec	**rec**	**F1**	**prec**	**rec**	**F1**	**prec**	**rec**	**F1**	**rec**	**prec**	**F1**	**prec**	**rec**	**F1**
alt med	.1 (.29)	.0 (.01)	.01 (.02)	.39 (.29)	.09 (.1)	**.14** (.13)	.61 (.09)	.45 (.13)	**.5** (.09)	.62 (.11)	.36 (.13)	**.44** (.11)	.64 (.11)	.41 (.11)	**.49** (.08)
consp	.69 (.1)	.33 (.08)	.44 (.05)	.59 (.05)	.41 (.06)	**.48** (.04)	.62 (.09)	.4 (.08)	**.48** (.05)	.59 (.09)	.44 (.09)	**.49** (.06)	.56 (.04)	.48 (.05)	**.52** (.04)
crit	.48 (.19)	.1 (.08)	.15 (.1)	.41 (.1)	.23 (.08)	**.29** (.09)	.5 (.08)	.18 (.06)	**.26** (.06)	.48 (.06)	.2 (.06)	**.28** (.06)	.45 (.07)	.28 (.06)	**.34** (.05)
dev	.46 (.22)	.13 (.11)	.2 (.14)	.48 (.13)	.26 (.11)	**.32** (.13)	.54 (.07)	.28 (.11)	**.36** (.1)	.56 (.09)	.26 (.11)	**.34** (.11)	.53 (.06)	.36 (.09)	**.42** (.08)
eff	.61 (.05)	.4 (.09)	.48 (.07)	.57 (.07)	.46 (.11)	**.51** (.09)	.63 (.07)	.41 (.09)	**.49** (.06)	.61 (.08)	.45 (.1)	**.51** (.06)	.58 (.05)	.54 (.05)	**.56** (.03)
inst	.65 (.05)	.53 (.09)	.58 (.05)	.62 (.05)	.55 (.09)	**.58** (.08)	.66 (.05)	.52 (.07)	.57 (.04)	.65 (.05)	.52 (.07)	**.58** (.03)	.64 (.04)	.58 (.05)	**.6** (.02)
lib	.76 (.03)	.77 (.05)	.76 (.02)	.76 (.02)	.75 (.04)	.75 (.02)	.75 (.03)	.79 (.06)	**.77** (.02)	.76 (.03)	.77 (.06)	**.76** (.02)	.76 (.02)	.76 (.05)	**.76** (.02)
mor	.0 (.0)	.0 (.0)	.0 (.0)	.32 (.43)	.01 (.02)	**.03** (.04)	.23 (.26)	.04 (.05)	**.06** (.07)	.21 (.14)	.04 (.04)	**.07** (.05)	.32 (.07)	.15 (.05)	**.19** (.05)
saf	.7 (.06)	.6 (.07)	.64 (.03)	.7 (.02)	.61 (.04)	**.65** (.02)	.72 (.05)	.58 (.05)	**.64** (.02)	.71 (.05)	.6 (.06)	**.64** (.03)	.7 (.04)	.64 (.03)	**.67** (.01)
micro avg	.7 (.02)	.51 (.04)	.59 (.02)	.67 (.01)	.54 (.04)	**.6** (.03)	.68 (.02)	.54 (.03)	**.6** (.01)	.68 (.02)	.55 (.03)	**.6** (.01)	.66 (.01)	.58 (.02)	**.62** (.01)
macro avg	.49 (.05)	.32 (.04)	.36 (.04)	.54 (.08)	.37 (.05)	**.42** (.05)	.58 (.03)	.4 (.03)	**.46** (.02)	.58 (.03)	.41 (.03)	**.46** (.03)	.58 (.02)	.47 (.02)	**.51** (.02)

Table 9: Classification results per class from RobBERT (averaged across folds and random seeds).

	Baseline			+ Back-transl.			+ GPT-3.5			+ Chat-GPT			+ Back-transl + Chat-GPT +GPT-3.5		
	prec	**rec**	**F1**	**prec**	**rec**	**F1**	**prec**	**rec**	**F1**	**prec**	**rec**	**F1**	**prec**	**rec**	**F1**
alt med	.08 (.28)	.0 (.01)	.01 (.02)	.56 (.17)	.22 (.09)	**.31** (.11)	.56 (.07)	.42 (.12)	**.47** (.09)	.61 (.07)	.34 (.11)	**.43** (.09)	.64 (.11)	.42 (.12)	**.49** (.09)
consp	.71 (.1)	.32 (.07)	.43 (.06)	.6 (.05)	.42 (.05)	**.49** (.04)	.62 (.07)	.4 (.07)	**.48** (.05)	.62 (.09)	.4 (.08)	**.48** (.04)	.61 (.06)	.43 (.06)	**.5** (.04)
crit	.48 (.13)	.08 (.05)	.14 (.07)	.41 (.05)	.25 (.04)	**.3** (.04)	.49 (.09)	.18 (.07)	**.25** (.07)	.45 (.07)	.21 (.07)	**.28** (.07)	.5 (.08)	.2 (.07)	**.27** (.07)
dev	.48 (.21)	.11 (.09)	.16 (.12)	.51 (.06)	.33 (.07)	**.4** (.06)	.53 (.06)	.26 (.1)	**.34** (.08)	.53 (.08)	.23 (.08)	**.31** (.08)	.53 (.06)	.29 (.1)	**.37** (.09)
eff	.63 (.07)	.42 (.08)	.49 (.06)	.6 (.05)	.5 (.05)	**.54** (.04)	.65 (.06)	.41 (.07)	**.49** (.05)	.62 (.08)	.46 (.07)	**.52** (.05)	.64 (.05)	.43 (.1)	**.5** (.07)
inst	.67 (.06)	.5 (.07)	.57 (.04)	.63 (.03)	.55 (.04)	**.59** (.01)	.67 (.04)	.49 (.07)	**.56** (.04)	.68 (.06)	.5 (.09)	**.56** (.05)	.65 (.06)	.53 (.08)	**.57** (.03)
lib	.75 (.04)	.77 (.06)	.76 (.02)	.75 (.02)	.74 (.04)	.75 (.02)	.75 (.03)	.78 (.06)	**.76** (.02)	.76 (.04)	.76 (.08)	**.76** (.02)	.75 (.02)	.78 (.05)	**.76** (.02)
mor	.0 (.0)	.0 (.0)	.0 (.0)	.16 (.33)	.01 (.01)	**.01** (.03)	.27 (.24)	.04 (.04)	**.07** (.06)	.29 (.21)	.05 (.04)	**.08** (.06)	.34 (.26)	.07 (.06)	**.11** (.09)
saf	.71 (.05)	.58 (.07)	.63 (.03)	.69 (.03)	.6 (.05)	**.64** (.03)	.71 (.04)	.58 (.06)	**.64** (.03)	.71 (.05)	.6 (.07)	**.64** (.03)	.7 (.04)	.6 (.06)	**.64** (.02)
micro avg	.7 (.02)	.5 (.03)	.58 (.02)	.66 (.01)	.55 (.01)	**.6** (.01)	.69 (.02)	.53 (.02)	**.6** (.01)	.68 (.02)	.53 (.03)	**.6** (.02)	.68 (.02)	.55 (.03)	**.61** (.01)
macro avg	.5 (.05)	.31 (.03)	.35 (.02)	.55 (.05)	.4 (.02)	**.45** (.02)	.58 (.03)	.4 (.02)	**.45** (.02)	.59 (.04)	.39 (.02)	**.45** (.02)	.6 (.04)	.42 (.03)	**.47** (.03)

Table 10: Classification results per class from RobBERTje (averaged across folds and random seeds).

	Baseline			+ Back-transl.			+ GPT-3.5			+ Chat-GPT			+ Back-transl + Chat-GPT +GPT-3.5		
	prec	rec	F1	prec	rec	F1	prec	rec	F1	rec	prec	F1	prec	rec	F1
alt med	.21 (.35)	.03 (.05)	.05 (.09)	.66 (.1)	.4 (.13)	**.48** (.13)	.67 (.12)	.41 (.1)	**.5** (.09)	.67 (.12)	.42 (.09)	**.51** (.08)	.66 (.09)	.43 (.1)	**.51** (.09)
consp	.67 (.09)	.38 (.07)	.48 (.05)	.58 (.05)	.47 (.04)	**.52** (.03)	.6 (.08)	.47 (.09)	**.52** (.05)	.6 (.08)	.47 (.09)	**.51** (.05)	.6 (.08)	.46 (.08)	**.51** (.04)
crit	.59 (.07)	.2 (.06)	.29 (.06)	.49 (.03)	.33 (.05)	**.39** (.04)	.55 (.05)	.28 (.06)	**.36** (.05)	.55 (.05)	.28 (.06)	**.36** (.05)	.51 (.07)	.3 (.08)	**.37** (.06)
dev	.53 (.18)	.27 (.15)	.33 (.14)	.57 (.05)	.43 (.08)	**.49** (.06)	.58 (.07)	.34 (.09)	**.42** (.08)	.58 (.07)	.33 (.09)	**.41** (.08)	.57 (.07)	.4 (.11)	**.45** (.08)
eff	.63 (.07)	.52 (.11)	.56 (.06)	.62 (.03)	.58 (.06)	**.6** (.04)	.63 (.05)	.57 (.08)	**.6** (.04)	.63 (.05)	.57 (.08)	**.59** (.04)	.64 (.06)	.57 (.07)	**.6** (.04)
inst	.69 (.06)	.53 (.09)	.59 (.05)	.65 (.04)	.59 (.03)	**.62** (.02)	.68 (.05)	.54 (.09)	**.6** (.04)	.68 (.05)	.54 (.08)	**.6** (.04)	.66 (.05)	.58 (.07)	**.62** (.03)
lib	.77 (.03)	.79 (.05)	.78 (.02)	.77 (.02)	.76 (.03)	.76 (.02)	.77 (.04)	.78 (.06)	.77 (.02)	.77 (.04)	.79 (.06)	.77 (.02)	.77 (.03)	.78 (.05)	.77 (.02)
mor	.0 (.0)	.0 (.0)	.0 (.0)	.19 (.27)	.02 (.03)	**.03** (.04)	.29 (.19)	.05 (.03)	**.08** (.06)	.29 (.18)	.05 (.03)	**.09** (.05)	.39 (.12)	.13 (.05)	**.18** (.06)
saf	.73 (.05)	.64 (.07)	.68 (.03)	.72 (.03)	.66 (.03)	**.69** (.02)	.72 (.05)	.66 (.06)	**.68** (.02)	.72 (.05)	.66 (.06)	**.68** (.02)	.75 (.06)	.63 (.07)	**.68** (.02)
micro avg	.71 (.02)	.55 (.03)	.62 (.02)	.68 (.01)	.6 (.01)	**.64** (.01)	.69 (.02)	.59 (.03)	**.63** (.01)	.69 (.02)	.59 (.03)	**.63** (.01)	.69 (.02)	.59 (.02)	.64 (.01)
macro avg	.53 (.05)	.37 (.03)	.42 (.03)	.58 (.04)	.47 (.02)	**.51** (.02)	.61 (.03)	.46 (.03)	**.5** (.02)	.61 (.03)	.46 (.03)	**.5** (.02)	.62 (.02)	.48 (.03)	.52 (.02)

Table 11: Classification results per class from ConTACT (averaged across folds and random seeds).

	Baseline			+ Back-transl.			+ GPT-3.5			+ Chat-GPT			+ Back-transl + Chat-GPT +GPT-3.5		
	prec	rec	F1	prec	rec	F1	prec	rec	F1	rec	prec	F1	prec	rec	F1
alt med	.16 (.28)	.03 (.07)	.05 (.11)	.64 (.09)	.25 (.07)	**.35** (.07)	.63 (.11)	.38 (.11)	**.46** (.08)	.65 (.11)	.34 (.14)	**.42** (.12)	.61 (.08)	.39 (.12)	**.46** (.09)
consp	.71 (.09)	.26 (.07)	.37 (.05)	.57 (.06)	.38 (.04)	**.46** (.03)	.6 (.08)	.36 (.07)	**.44** (.04)	.61 (.11)	.32 (.07)	**.41** (.05)	.56 (.09)	.42 (.07)	**.47** (.03)
crit	.53 (.12)	.1 (.06)	.16 (.07)	.43 (.04)	.28 (.04)	**.33** (.04)	.53 (.08)	.19 (.07)	**.26** (.08)	.47 (.11)	.23 (.08)	**.29** (.07)	.48 (.08)	.24 (.08)	**.31** (.06)
dev	.5 (.18)	.16 (.1)	.23 (.13)	.52 (.06)	.34 (.07)	**.41** (.07)	.56 (.09)	.29 (.08)	**.37** (.07)	.57 (.09)	.27 (.11)	**.35** (.1)	.55 (.1)	.34 (.1)	**.41** (.09)
eff	.61 (.06)	.45 (.09)	.51 (.05)	.59 (.05)	.52 (.05)	**.55** (.05)	.62 (.07)	.43 (.09)	**.5** (.05)	.62 (.07)	.46 (.08)	**.52** (.05)	.6 (.06)	.5 (.09)	**.54** (.05)
inst	.64 (.05)	.51 (.09)	.56 (.05)	.63 (.03)	.57 (.04)	**.6** (.02)	.65 (.05)	.51 (.07)	**.57** (.03)	.67 (.06)	.51 (.09)	**.57** (.05)	.63 (.05)	.58 (.06)	**.6** (.02)
lib	.75 (.03)	.78 (.06)	.76 (.02)	.76 (.02)	.76 (.03)	.75 (.01)	.75 (.04)	.78 (.07)	**.76** (.02)	.75 (.04)	.78 (.06)	**.76** (.02)	.75 (.03)	.77 (.07)	**.76** (.02)
mor	.01 (.04)	.0 (.0)	.0 (.01)	.29 (.32)	.04 (.04)	**.07** (.07)	.33 (.32)	.04 (.05)	**.07** (.08)	.42 (.26)	.06 (.05)	**.1** (.07)	.33 (.12)	.13 (.09)	**.17** (.1)
saf	.7 (.06)	.59 (.07)	.63 (.03)	.7 (.03)	.64 (.03)	**.66** (.02)	.71 (.05)	.6 (.06)	**.65** (.03)	.69 (.06)	.62 (.07)	**.65** (.02)	.72 (.06)	.61 (.06)	**.66** (.03)
micro avg	.69 (.02)	.51 (.03)	.58 (.02)	.66 (.01)	.57 (.01)	**.61** (.01)	.68 (.02)	.54 (.03)	**.6** (.01)	.67 (.02)	.55 (.04)	**.6** (.02)	.66 (.01)	.57 (.03)	**.61** (.01)
macro avg	.51 (.03)	.32 (.04)	.36 (.04)	.57 (.04)	.42 (.02)	**.46** (.02)	.6 (.04)	.4 (.03)	**.45** (.02)	.61 (.04)	.4 (.03)	**.45** (.03)	.58 (.02)	.44 (.03)	**.49** (.02)

Table 12: Classification results per class from BERTje (averaged across folds and random seeds).