SUPERVISED VS. ZERO-SHOT LEARNING AUTOMATIC CLASSIFICATION OF COMMENTS ON EDUCATIONAL VIDEOS USING PRE-TRAINED LANGUAGE MODELS

Benedict Kettler, Stefan Hillmann

Technische Universität Berlin
benedict.kettler@gmail.com

Abstract: Despite the potential of AI, only a small percentage of small and medium-sized enterprises (SMEs) are adopting it due to data issues, expertise gaps, and implementation barriers. Zero-shot learning offers a promising approach for SMEs by minimizing these obstacles. This paper explores the use of zero-shot learning in a real-world NLP classification task on online comments (comparable with intent classification tasks) from the e-learning platform Sofatutor. While fine-tuning has achieved high accuracy (82.3–86.5%), zero-shot models have shown lower performance (39.3–61.4%) due to different label selection, grouping of different scenarios in one class and the type of classification task. Even if the current accuracy is not sufficient for practical application, pre-filtering the data using zero-shot learning might be a promising option for SMEs.

1 Introduction

The potential of artificial intelligence (AI) seems immense. According to a study by the McKinsey Global Institute, the potential of generative AI alone is estimated at up to USD 4.4 trillion [1]. Despite this, only 17.6% of small and medium-sized enterprises (SMEs) in Germany do currently utilize AI [2]. The reasons behind SMEs not adopting AI were examined in a study by Deloitte, revealing "Lack of skills" (65%), "barriers to implementation" (52%), and "data problems" (52%) as the top three obstacles [3].

In the domain of language models, these justifications are reasonable, considering the complexity and the required resources needed to train a language model from scratch. But recently, language models have undergone rapid development [4] and therefore also the accessibility for SMEs improved. By the invention of the transformer architecture [5] new opportunities for companies with limited resources came up. Pre-trained Transformer models are accessible to other users, often requiring only fine-tuning to adapt to specific use cases. Consequently, the utilization of language models is no longer exclusive to major technology firms. SMEs now have the chance to access and leverage large pre-trained Transformer models [6]. However, despite fine-tuning being less complex than developing an entire (large) language model, it necessitates fundamental programming skills, a general understanding of (large) language models and available training data, therefore pre-trained language models are still not a suitable solution for all SMEs.

A common strategy for enhancing the performance of language models involves increasing the model size and expanding the training dataset. The resulting large language models (LLMs) exhibit impressive results even with minimal fine-tuning [7]. Brown et al. also investigated, models that did not require any fine-tuning at all, a technique known as zero-shot learning (ZSL). These language models present a significant opportunity for SMEs. A zero-shot model requires minimal skills, almost no implementation effort, and no training data for fine-tuning.

DOI: 10.35096/othr/pub-7092, URN: urn:nbn:de:bvb:898-opus4-70928

The by Deloitte investigated obstacles can therefore be nearly eliminated. But how effective is zero-shot learning compared to the conventional fine-tuning approach?

For this purpose, an exemplary investigation of a text classification task with two zero-shot and one fine-tuned model is implemented using user comments from the E-learning platform Sofatutor. The comments are short and comparable to the length of user utterance directed to chatbots or speech based systems. Thus, our comment classification task is comparable to an intent classification task. Furthermore, the comments from a real world application reflect young peoples' language as used in the messaging applications. This property, allows an evaluation of models for text classification that has high relevance for engineers and researcher, especially in SMEs and higher education institutes with limited resources, building language based systems for such comparable taget groups.

Sofatutor was founded in Berlin in 2008 and currently has around 250 employees. Regarding the European Commission's definition, Sofaturor is classified as a typical medium-sized company [8]. The investigation aims to explore the accuracy of the text classification, as well as to elaborate the potential challenges in zero-shot learning. In the end, a conclusion about the usability of zero-shot learning for SMEs by the example of the Sofatutor case will be discussed.

2 Methodology

The platform specializes in producing educational content for pupils from the first to the twelfth grade. With over 1.5 million registered users, Sofatutor is the largest E-learning platform for pupils in the German-speaking region [9]. User are able to post comments below the educational content. These comments are manually classified by an employee to answer user questions as well as to filter out inappropriate content.

2.1 Dataset

Three classes are used for this purpose. Examples for each class are provided in Table 1. The class *support* includes content related and administrative questions, error reports, as well as feedback on the content. These comments are then forwarded to the editorial team, which can then take appropriate action. The class *hide* contains insults, spam or pointless comments. This encompasses comments that use an excessive number of emojis or the elongation of letters. Comments that fall into the hide class are deleted from the platform. Furthermore, greetings posted without context are also classified as *hide*. Finally, the third class *no action* is the default class. Every comment that does not fall into *support* or *hide* is classified as *no action*. In most cases of *no action*, the content is praised, or it answers questions that were made in the video for a higher engagement of the pupils (e.g., "do you also have pets?"). As the name of the class suggests, no further action is taken.

At the time of the experiment, there were in total 7,840 labeled comments distributed as the following: *support* (1,543), *hide* (1,262) and *no action* (5,035). Around 20% of the data, measured by the number of comments of the smallest class is used for testing, resulting in the following distribution for the training data: *support* (1,290), *hide* (1,009) and *no action* (4,782). Since the comments are mainly written by pupils, most of them are not orthographically correct, and many comments are unconventional in their syntactic structure. This could become a major challenge for the language models.

Table 1 – Examples of comments in the used dataset.

Class	**Examples**
support	I still don’t understand why there are still dinosaurs in the year 3005 (?o?)Ã, …
	I love this video. But I’m silly and can’t remember it :c
	I can’t find any videos for the 3rd grade, can you help me ???????????????????????????
	In the video, it says: ’you write the result starting with the ones below’..... but you unfortunately start with the thousands.... that is confusing!
	Im in sixt grade but I just see videos from fourth grade...
	Good video but the colors don’t match my textbook ?? that makes it a bit complicated :/
	Hello, can you also make a video about the twenties strips. Best regards
	Not helpful
hide	Follow me on Instagram @******
	Suuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuupa
	You stink sandra
	Hello
	shitty
	What bro what should I say brother?
	That’s shit
	Ewg3r&5Dg4
no action	Thank you, it helped me a lot ??
	:)
	Super, great video, it couldn’t be easier to understand. I will definitely use it.
	Thanks to the author!
	the end was very funny
	Super!
	I understood everything and was able to continue straight away thanks
	REALLY COOL

2.2 Fine-tuning for Classification

For fine-tuning, the dataset is used with the pre-trained model *xlm-roberta-base*. Due to class imbalance, two approaches are taken to balance the classes. In the first approach, the *no action* and *support* classes are adjusted to match the smaller *hide* class so that there are in total 3 × 1,009 comments for training. The number of comments will be doubled for the fine-tuning until no further improvements are observed. Missing training data in the classes *hide* and *support* is filled with duplicates. This procedure should also investigate how crucial the amount of training data is, since the limited availability of training data is a common problem for SMEs. Various hyperparameters, such as batch size, number of epochs, and learning rate, are tested to achieve optimal accuracy. Fine-tuning is conducted using the AllenNLP framework.

2.3 Zero-shot classification

Several zero-shot classification runs are performed to test different methods for optimal label selection in relation to the used models. In order to ensure a diverse set of label options, the Sofatutor employee who previously labeled the comments manual as well as ChatGPT (based on GPT-3.5) were asked for three suitable additions to the hypothesis statement "This text is ". Therefore, some example comments were provided. The resulting labels have been combined, resulting in $3^3 = 27$ label combinations for each label method (employee and ChatGPT). Two models MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7 (called mDeBERTa in the following) and joeddav/xlm-roberta-large-xnli (called roberta in the following), both available via Hugging Face, are used. The two models and the two labeling methods are then also combined, resulting in $4*27 = 108$ zero-shot classification runs. This extensive zero-shot learning investigation was undertaken to investigate the variation in classification across labels and models.

3 Results

3.1 Results for fine-tuning classification

This section presents the results of the fine-tuning approach. In order to balance the classes, the two larger classes were adjusted to match the size of the smaller class, resulting in the use of 3 x 1,009 = 3,027 comments for the initial training iteration. The total number of training data was quadrupled by duplicating instances within the smaller classes. As shown in Table 2, after tripling the data, no improvement in the accuracy was observed. Increasing the volume of the training data by multiplying samples from smaller classes shows a higher accuracy. Nevertheless, even without the multiplication of comments, an accuracy of 82.3% has been achieved. For a better understanding, the confusion matrices are shown in Table 3.

Table 2 – Accuracy and amount of training data for the fine-tuning approach.

No. of Training Samples per Class			
support	***hide***	***no action***	**Accuracy**
1.009	1.009	1.009	**0.823**
2.018	2.018	2.018	**0.841**
3.027	3.027	3.027	**0.865**
4.036	4.036	4.036	**0.862**

Table 3 – Confusion matrices of the fine-tuning approaches.

	Prediction (% accuracy)					
	3,027 comments			**6,054 comments**		
Label	*no action*	*support*	*hide*	*no action*	*support*	*hide*
no action	77.8	7.1	15.1	82.7	7.1	10.2
support	7.1	83.3	9.5	17.5	77.0	5.6
hide	7.9	6.3	85.8	3.2	4.0	92.9
	9,081 comments			**12,108 comments**		
Label	*no action*	*support*	*hide*	*no action*	*support*	*hide*
no action	81.1	9.4	9.4	86.6	7.1	6.3
support	3.2	86.5	10.3	7.1	90.5	2.4
hide	5.6	2.4	92.1	11.1	7.1	81.7

When looking at the confusion matrices in Table 3, no major structural weaknesses become apparent. The fine-tuning approach shows (also for the other runs) slight weaknesses in the prediction of the no action class, while the hide class is predicted quite well.

3.2 Results for zero-shot classification

For the zero-shot approach, three labels per class were generated by a Sofatutor employee and by ChatGPT as shown in Table 4. These labels were combined and tested with two models, which resulted in 108 classification runs (3^3 label combinations * 2 label approaches * 2 models).

Table 4 – Label selection for the zero-shot classification.

Class	Sofatutor Employee	ChatGPT
no action	praising helpful positive	Helpful Praiseworthy Enthusiastic
support	a problem case critical questioning	Confusing Improvable Critical
hide	spam offensive meaningless	Not acceptable Inappropriate Offensive

Table 5 – Average accuracy of the zero-shot classification per approach

	Prediction (% average accuracy)	
Model	**Employee labels**	**ChatGPT labels**
mDeBERTa	0.545	0.530
roberta	0.530	0.462

As seen in Table 5 the mDeBERTa model beats the roberta model in both studies. It can also be observed that the determination of the labels by the employee led to better results than the determination of the labels with the help of ChatGPT. The roberta model in particular shows strong differences in accuracy depending on the type of label creation. The ranges and variances are similar for all experimental approaches.

The confusion matrices are also examined in order to determine any structurally incorrect predictions. For this purpose, the predictions of the 27 label combinations of an approach are added together and shown as a percentage in Table 6. It can be observed that the *hide* class performed very poorly in all experiments. The *no action* class, on the other hand, was predicted relatively well across the board and the accuracy can sometimes keep up with the fine-tuning approach. There are strong fluctuations in the *support* class, where the accuracy ranges between 43.9% and 82.8%. The confusion matrices of the best runs per approach are shown in Table 7.

Overall, fine-tuning achieves significantly better results with each approach than all zero-shot approaches. It can be observed that the zero-shot models have the most problems with predicting the *hide* class, while the fine-tuned model has there the highest accuracy. The zero-shot models are very dependent on the choice of the appropriate labels. The model with fine-tuning is independent of the choice of label, but dependent on the selection and distribution of the training data.

Table 6 – Confusion matrices with average accuracy of the zero-shot approaches.

		Prediction (% accuracy)					
		Employee labels			**ChatGPT labels**		
Model	**Label**	*no action*	*support*	*hide*	*no action*	*support*	*hide*
roberta	*no action*	70.7	29.3	0.1	78.9	17.7	3.4
	support	14.0	82.8	3.2	41.3	43.9	14.8
	hide	18.9	75.9	5.2	44.2	40.2	15.6
mDeBERTa	*no action*	80.7	16.7	2.5	76.6	20.2	3.2
	support	16.6	73.4	10.0	19.5	57.7	22.8
	hide	38.3	52.4	9.2	35.3	39.8	24.8

Table 7 – Confusion matrices with the highest accuracy per approach in the zero-shot scenario.

		Prediction (% accuracy)					
		Employee labels			**ChatGPT labels**		
Model	**Label**	*no action*	*support*	*hide*	*no action*	*support*	*hide*
roberta	*no action*	81.1	18.9	0.0	97.6	2.4	0.0
	support	12.7	82.5	4.8	54.0	24.6	21.4
	hide	22.2	70.6	7.1	59.5	6.3	34.1
mDeBERTa	*no action*	85.8	12.6	1.6	82.7	7.9	9.4
	support	14.3	84.1	1.6	25.4	27.8	46.8
	hide	50.8	34.9	14.3	19.0	12.7	68.3

4 Discussion

The poorer performance of zero-shot language models can be attributed, in part, to the nature of their application. The leading model of this experiment, mDeBERTa-v3-base-xnli-multilingual-nli-2mil7, was trained on datasets such as MultiNLI, Fever-NLI, and Adversarial-NLI. In Natural Language Inference (NLI) tasks, texts are evaluated for semantic coherence, aiding a language model in understanding language context. However, in this scenario, comments were classified not based on content but on a meta-level. Unlike content-related queries, which establish a contextual relationship, comments were categorized based on their inherent nature.

Furthermore, the diverse range of scenarios covered by a single class is reflected in the created labels. For the *hide* class, labels such as "offensive," "meaningless," and "spam" were assigned, while "meaningless" and "offensive" are not synonymous. This lack of semantic closeness between some comments of a class and labels might contribute to the poorer performance of zero-shot classification. Summarizing diverse comment content into a single class represents a weakness compared to the fine-tuning approach. Human-generated labels led to better results for both models, suggesting a possible advantage in human categorization. However, the Sofatutor employee had a considerable knowledge advantage over ChatGPT due to the previously executed manual classification.

Additionally, the level of complexity increased due to the fact that the comments are authored by children, resulting in incoherent, contradictory, or unconventional expressions with frequent spelling and grammar errors. The elongation of individual letters or the abundance of emojis often led to a *hide* classification in the manual procedure. These types of comments posed a significant challenge for a zero-shot classification.

Moreover, the brevity of comments presented a challenge for language models. While both fine-tuning and zero-shot approaches faced difficulties, the zero-shot model struggled more due to the lack of additional context for short texts.

However, challenges persisted, and the study highlighted the need for more practical research on the deployment of zero-shot models in real-world scenarios. Understanding the tasks and data that can be processed with high accuracy, as well as those requiring fine-tuning, is essential for effective utilization in various industries. Consideration should be given to the imperfect and unconventional nature of texts in real-world applications. Current research datasets typically originate from high-quality and/or synthetic sources, and the robustness of zero-shot models under special data conditions need further exploration.

5 Conclusion

This research aimed to test the accuracy of text classification using two different methods through a practical experiment. The classification was conducted using a language model with fine-tuning and two different zero-shot models. The study utilized comments from the E-Learning platform Sofatutor, categorized into three classes: *no action* (mostly praise or neutral statements), *support* (questions, errors in tasks, or constructive feedback), and *hide* (meaningless comments or insults). The language model xlmroberta-base was trained with four different training sizes, achieving classification accuracies between 82.32% and 86.54%.

For zero-shot classification, two different language models (roberta and mDeBERTa) and two label generation approaches were chosen. A Sofatutor employee and ChatGPT were consulted for suitable additions to the hypothesis statement, resulting in an average accuracy (averaged per label approach and model) between 46.2% and 54.5%, with the best result reaching 61.48%. Human-generated labels consistently led to better results in both models.

Overall predictions in the three different classes for the fine-tuning approach were balanced, with slight weaknesses in the *no action* class. In the zero-shot approach, weaknesses were observed in the *hide* class due to the diverse content falling into the class. The inherent difficulty of classifying non-homogeneous classes with a single label in zero-shot learning likely contributed to this issue. In contrast, fine-tuning does not encounter this challenge. Furthermore, the lower performance when using zero-shot learning could be attributed to the special linguistic structure of comments written by pupils (1st to 12th grade). The data quality differs from often used high-quality datasets in research and needs further investigation to strengthen this assumption.

While zero-shot models offer advantages for SMEs, such as ease of use and minimal implementation requirements, they are currently less accurate than fine-tuned models. Few-shot classification, requiring minimal training data, presents a promising approach that warrants further exploration. This research aims to contribute to understanding if zero-shot learning is applicable in business contexts. With a maximum accuracy of 61.48%, zero-shot classification is deemed too imprecise for practical use in Sofatutor. However, pre-sorting the *no action* class through a zero-shot model could be considered, as it does not lead to irreversible actions. Understanding the strengths and weaknesses of zero-shot learning is crucial for companies, and further practical research is needed.

References

[1] Chui, M., E. Hazan, R. Roberts, A. Singla, K. Smaje, A. Sukharevsky, L. Yee, and R. Zemmel: *The economic potential of generative ai: The next productivity frontier.* 2023. URL https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier.

[2] Demary, V. and H. Goecke: *Wie KMU Künstliche Intelligenz nutzen.* 2021. URL https://www.iwkoeln.de/studien/vera-demary-henry-goecke-wie-kmu-kuenstliche-intelligenz-nutzen.html.

[3] Meyer, L. and M. Seiz: *Künstliche Intelligenz im Mittelstand.* 2021. URL https://www2.deloitte.com/content/dam/Deloitte/de/Documents/Mittelstand/Erfolgsfaktorenstudie_K%C3%BCnstliche%20Intelligenz%20im%20Mittelstand.pdf.

[4] Zhang, D., S. Mishra, E. Brynjolfsson, J. Etchemendy, D. Ganguli, B. Grosz, T. Lyons, J. Manyika, J. C. Niebles, M. Sellitto, Y. Shoham, J. Clark, and R. Perrault: *The ai index 2021 annual report. Human-Centered AI Institute - Stanford University*, 2021.

[5] Vaswani, A., N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin: *Attention is all you need.* In *Neural Information Processing Systems*. 2017.

[6] Bommasani, R., D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. S. Chatterji, A. S. Chen, K. A. Creel, J. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. E. Gillespie, K. Goel, N. D. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. F. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. S. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. P. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. F. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. H. Roohani, C. Ruiz, J. Ryan, C. R'e, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. P. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. A. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, and P. Liang: *On the opportunities and risks of foundation models. ArXiv*, 2021. URL https://crfm.stanford.edu/assets/report.pdf.

[7] Brown, T. B., B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei: *Language models are few-shot learners. Advances in Neural Information Processing Systems*, 33, pp. 1877–1901, 2020.

[8] European Commission: *Sme definition.* 2003. URL https://single-market-economy.ec.europa.eu/smes/sme-definition_en.

[9] Blumenstein, T.: *Sofatutor pressebereich.* 2023. URL https://www.sofatutor.com/about/press.