

# Leveraging Large Language Models and Weak Supervision for Social Media data annotation: an evaluation using COVID-19 self-reported vaccination tweets

Ramya Tekumalla and Juan M. Banda

Georgia State University, Atlanta, GA, 30328, USA  
{rtekumalla1, jbanda}@gsu.edu

**Abstract.** The COVID-19 pandemic has presented significant challenges to the healthcare industry and society as a whole. With the rapid development of COVID-19 vaccines, social media platforms have become a popular medium for discussions on vaccine-related topics. Identifying vaccine-related tweets and analyzing them can provide valuable insights for public health researchers and policymakers. However, manual annotation of a large number of tweets is time-consuming and expensive. In this study, we evaluate the usage of Large Language Models, in this case GPT-4 (March 23 version), and weak supervision, to identify COVID-19 vaccine-related tweets, with the purpose of comparing performance against human annotators. We leveraged a manually curated gold-standard dataset and used GPT-4 to provide labels without any additional fine-tuning or instructing, in a single-shot mode (no additional prompting).

**Keywords:** Large language models, GPT, weak supervision, social media data, Twitter.

## 1 Introduction

### 1.1 A Subsection Sample

The widespread adoption of social media platforms has led to an explosion of user-generated content, making them valuable sources of real-time information [1]. Social media platforms have become a valuable resource for studying public health issues [2], including the COVID-19 pandemic. Social media platforms like Twitter have a vast user base, representing diverse demographics and geographic locations. Analyzing vaccination sentiment data from such platforms allows for a more comprehensive understanding of public opinion, as it encompasses a wide range of perspectives. Twitter, in particular, has emerged as a platform where individuals share their personal experiences, including vaccination updates [3]. By analyzing the data, public health officials, policymakers, and researchers can gauge the overall sentiment towards vaccines, identify trends, and make informed decisions to address concerns or misconceptions. Analyzing self-reported vaccination tweets can provide valuable insights

into vaccine sentiment, vaccine uptake, and vaccine-related concerns among the general population. However, manually annotating large volumes of social media data is labor-intensive and time-consuming, requiring domain experts to label the data accurately.

Weak supervision [4] techniques have emerged as a powerful approach for data annotation, offering significant advantages in terms of scalability [5], cost-effectiveness [6], and flexibility [7]. Traditional methods of data annotation often rely on manual labeling, which can be time-consuming, expensive, and limited in terms of the volume of labeled data that can be produced. In contrast, weak supervision techniques leverage various sources of supervision to automatically generate labeled data, reducing the manual effort required while maintaining reasonable accuracy. Scalability is one of the primary advantages of weak supervision. With the exponential growth of data, manually labeling vast amounts of data becomes impractical and expensive. Weak supervision allows for the rapid annotation of large datasets by leveraging existing resources such as heuristics, rules, or readily available weak labels [8]. These weak signals can be automatically applied to unlabeled data, effectively increasing the amount of labeled data available for training and development of robust machine learning models. Cost-effectiveness is another key benefit of weak supervision techniques. Manual data annotation often requires skilled human annotators, which can be costly and time-consuming. In contrast, weak supervision reduces the reliance on manual annotation efforts, thus reducing costs. Although weakly supervised labels may not be as accurate as manually annotated labels, they can still provide valuable insights and improve the performance of machine learning models. By combining weakly supervised labels with a smaller amount of manually labeled data, comparable results can be achieved at a fraction of the cost [9]. Additionally, traditional data annotation methods often require significant upfront effort to design annotation schemas, guidelines, and quality control processes. These rigid procedures can be challenging to adapt as new data sources or requirements emerge [10]. In contrast, weak supervision provides a more agile and adaptable approach to data annotation. Weakly supervised labels can be easily generated or modified based on changing needs, enabling rapid iteration and refinement of models in response to evolving data or domain-specific requirements.

Large language models (LLMs), such as GPT-3 [11], have revolutionized natural language processing and transformed various applications across multiple domains. These models employ deep learning techniques to generate coherent and contextually relevant text, making them invaluable for tasks like language translation, text summarization, and conversational agents. Their effectiveness is attributed to the vast amount of pre-training data and the ability to capture complex linguistic patterns. This work assesses the effectiveness of Language Models (LLMs) (GPT-3.5 and GPT-4 (March 23 version)), in conjunction with weak supervision, for the identification of COVID-19 vaccine-related tweets. The primary objective is to compare the performance of LLMs against human annotators. To achieve this, we utilized an expertly curated gold-standard dataset and employed GPT3.5 and GPT-4 to generate labels in a single-shot mode, without resorting to additional fine-tuning or explicit instructions.

## 2 Related Works

In the past weak supervision has demonstrated successful results in clinical text classification [12], multi-language sentiment classification [13], generating training sets for phenotype models [14], information retrieval [15], identifying drugs from Twitter [16–18], classifying different kinds of epidemics [19], natural disasters [20, 21] and several health applications [22–24]. In this aspect, LLMs have been effectively utilized to leverage weak supervision techniques, automating data annotation processes by generating or modifying labels based on the model's pre-trained knowledge and heuristics. LLMs, such as BERT [25] and GPT [26], have shown impressive performance in various natural language processing tasks, including sentiment analysis, named entity recognition, and text classification. These models can be fine-tuned on domain-specific datasets, enabling them to learn specific patterns and characteristics of the data. By leveraging pre-trained LLMs, researchers can automate or assist in the annotation process, significantly reducing the human effort required for data labeling. The evolution of LLMs has been marked by significant milestones, with BERT acting as a groundbreaking advancement. BERT introduced the concept of pretraining and fine-tuning, revolutionizing the field of NLP. By pretraining models on large corpora of text data and fine-tuning them on specific downstream tasks, BERT achieved state-of-the-art performance on a wide range of NLP benchmarks. BERT served as a foundation and inspiration for the development of numerous pre-trained models like GPT [26], AIBERT [27], RoBERTa [28], DistilBERT [29], ELECTRA [30], XLNet [31], T5 [32], MegatronLM [33], BART [34], CamemBERT [35]. These models have leveraged the success of BERT's architecture and training techniques to and improved the model by tackling various limitations like performance, optimization, reduction in training size. As a result, several other domain specific pre-trained models like Covid-Twitter-BERT [36], BioBERT [37], SciBERT [38], ClinicalBERT [39], LegalBERT [40], FinBERT [41–43] emerged. Building upon the success of BERT, subsequent models such as GPT-2 [44] and GPT-3 [11] further pushed the boundaries of LLM capabilities. GPT-2 demonstrated impressive language generation abilities, while GPT-3 introduced even larger model sizes and showcased the potential for diverse applications. GPT-3.5 is a transitional model, which further refines the AI's capabilities of GPT-3, and is known for nuanced understanding and contextual response generation. GPT-4, introduces a major leap with significant improvements in model size, training data, and comprehension abilities. GPT-4 is designed to better handle ambiguities and complexities in natural language, generating more coherent, relevant, and detailed responses.

In the aspect of data labeling, LLMs have emerged as a promising solution to address these challenges by automating or assisting in the data annotation process. With their language understanding capabilities, LLMs can be employed to generate annotations or suggest labels for a given input, a technique known as active learning [45]. This approach allows human annotators to focus their efforts on more challenging or uncertain instances, thereby improving the efficiency and quality of the annotation process. Previous research has demonstrated that 35-40% of the crowd workers widely use LLMs for text related data annotation tasks [46]. In a study conducted by Gi-

lardi et.al, Chat-GPT outperformed Crowd-Workers for text annotation tasks [47]. To improve the precision of ChatGPT as the hallucination is one of the limitations of LLMs, He et.al. designed a two step approach to explain why the text was labeled [48]. LLMs have demonstrated success in various data annotation tasks [49], sentiment analysis [50], text categorization, linguistic annotations [51], multi-linguistic data annotation [52] and social computing [53].

This work examines the role of LLMs in data annotation, discussing the benefits, limitations, and potential future directions. The advancements in LLMs have not only transformed NLP tasks but have also had a profound impact on human tasks that involve language understanding and generation. LLMs have been integrated into various applications, ranging from chatbots and virtual assistants to language translation and content generation. In human-computer interaction scenarios, LLM-based systems have enabled more natural and effective communication, bridging the gap between machines and humans. However, the increasing reliance on LLMs also raises important ethical and societal considerations, such as potential biases and the responsible deployment of AI technologies [54]. LLMs exhibit non-deterministic behavior, similar to human coders, where identical input can produce varying outputs [55, 56]. Hence, it is crucial to exercise caution when utilizing LLMs to ensure consistent and reliable results.

### 3 Methods

#### 3.1 Datasets Used

##### 3.1.1 Gold standard dataset

We collected a dataset of tweets related to COVID-19 vaccines by filtering related keywords, from one of the largest COVID-19 Twitter datasets [57] available. After filtering, this dataset consists of 2,454 self reported vaccination confirmation tweets and 19,946 vaccine chatter tweets. The complete dataset was manually curated by two medical students, having a Cohen Kappa score inter annotator agreement of 0.82 with a third annotator resolving all conflicts. This dataset was used in the Social Media Mining for Health 2022 shared tasks [58]. With the annotation task consuming over 200 human hours, it is vital to try to identify additional techniques to attempt to streamline this process.

##### 3.1.2 Silver standard dataset

While weak supervision has shown promise in the area of social media mining [17, 59], we extracted an additional dataset, not manually curated, which consists of tweets selected by a weak labeling heuristic consisting of expressions like “vaccine causes”, “I was vaccinated”, “I got Moderna”, and similar. This weakly-supervised, or ‘silver standard’, consists of 750,000 randomly sampled (from a larger set of 12 million) tweets with an unidentifiable mixture of both classes. The rationale for doing so is

that researchers have shown that data augmentation using weak supervision leads to better and more generalizable models, than when only using gold standard data [60, 61]. Note that none of the 750,000 randomly sampled tweets used in this dataset do not have any overlap with the gold standard data.

### 3.2 Additional language models used

Besides the previously mentioned GPT-4 and GPT-3.5, we fine tuned COVID-Twitter-BERT [36] and BERTweet [62] with the GPT-labeled silver-standard data, for downstream tweet classification. Note that the class imbalance from the gold-standard dataset is roughly 1 to 8, between self reported vaccination tweets and vaccine chatter tweets. This was also found to be similarly the same in the GPT-labeled silver-standard data, making the fine-tuning and evaluation comparable.

### 3.3 Evaluation set-up

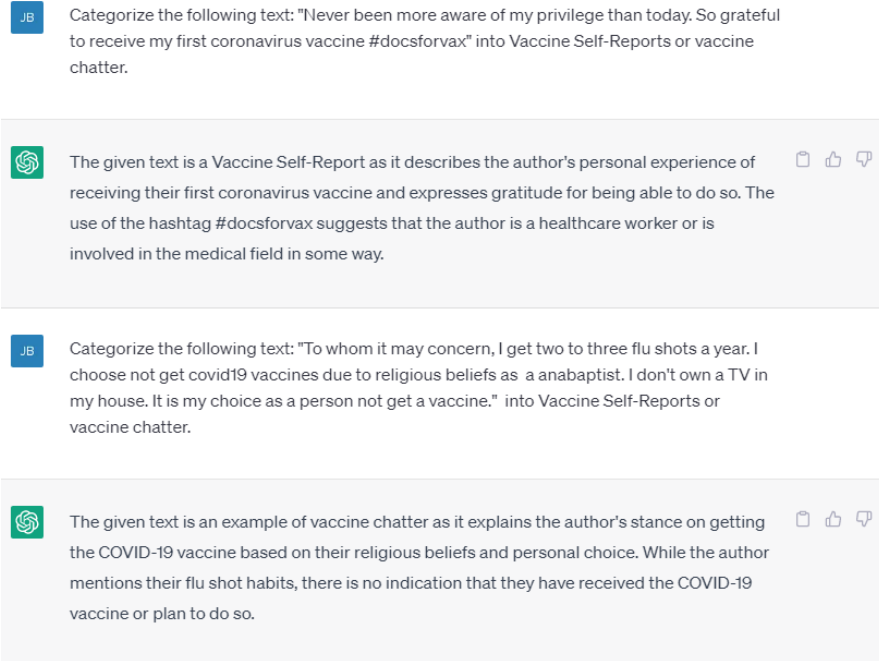
#### 3.3.1 LLM performance in annotating data

We evaluate the performance of GPT-4 and GPT-3.5 on the labeling of the gold-standard data. This evaluation will assess how good are LLMs in labeling data when compared to a set of medical professionals. As one of the most resource-expensive parts of generating datasets, if human annotation/labeling can be aided or streamlined, there is great value in leveraging LLMs in these types of tasks. Leveraging the Open AI API for both GPT-3.5 and GPT-4, we used the following prompt: “Categorize the following text: XXXXXXXXXXXX into vaccine self-reports or vaccine chatter. Figure 1 shows a sample output of the GPT-4 prompting on the chat.openai.com website. This evaluation made 22,400 API calls to each GPT-4 and GPT-3.5 models.

#### 3.3.2 LLM to improve weakly supervised dataset creation

In these evaluations, we leverage GPT-4/GPT-3.5 to attempt to ‘properly’ label the silver-standard data and then fine-tune BERT-like models to classify the gold-standard data. The creation of silver-standard datasets has gained popularity in the field of NLP with many groups building systems that leverage silver-standard data to enhance their training sets and achieve state-of-the-art results in a variety of NLP shared tasks [58, 63]. Using the same prompt for the first evaluation, we made a total of 750,000 API calls to each GPT model to label this silver-standard dataset.

With these evaluations we aim to answer two questions: a) is GPT-4/GPT-3.5 enough to annotate data with similar quality than a human expert, b) could we leverage both weak supervision and GPT-4/GPT-3.5 to quickly and scalably annotate large amounts of data with near-expert level performance. We would call these datasets: electrum datasets, which are a mixture of gold and silver standard-data.



**Fig 1:** Sample GPT-4 prompts to evaluate the created datasets.

## 4 Results

Before we introduce the actual results of our analysis, we would like to present a cost analysis of how much it would cost to run the data annotation tasks leveraging the GPT models and other traditional sources. We sent a total of 1,544,800 API calls, with a total cost of \$2,743.40 USD. While this price might seem high, note that we annotated a total of 1,544,800 tweets, which would be time and cost prohibitive to do hiring humans and paying them a fair wage. Even using a service like Amazon Sage-Maker Ground Truth, would cost around \$52,896.00 USD for the same task. Leveraging Amazon Mechanical Turk would cost \$37,075.20 USD for the same number of text classification tasks [64]. There is clear value in evaluating if we can leverage such a resource for data annotation, this would particularly help resource constrained researchers that can not afford to pay expert annotators. The second aspect is scale, while ~1.5 Million API calls are done fairly quickly, nobody to our knowledge has manually annotated any dataset this large.

#### 4.1.1 Results for LLM performance in annotating data

In Table 1 we showcase the annotating performance of both GPT-4 and GPT-3.5. It is not surprising that GPT-4 outperformed GPT-3.5 by nearly 10% for the self reported vaccination tweets category, the more interesting one, and marginally for vaccine chatter. While vaccine chatter is more easily identified, nearly 90% for both models, GPT-3.5's 71.11% performance on the self reported class, and 80.81% for GPT-4 are promising numbers. However, once larger amounts of data are annotated this way, this would lead to a considerable amount of noise to be added. These results are still promising as there was no additional prompting or fine-tuning performed, so the zero-shot results are pretty solid.

**Table 1.** Correct tweet labeling results for GPT models.

| Label                     | GPT-4  | %      | GPT-3.5 | %      |
|---------------------------|--------|--------|---------|--------|
| Self reported vaccination | 1,983  | 80.81% | 1,745   | 71.11% |
| Vaccine chatter           | 18,541 | 92.96% | 17,842  | 89.45% |

We look at the inter-annotator agreement between both GPT models using Cohen Kappa coefficient [65] and the human annotators. We evaluate this to get insights into how much the correctly labeled tweets diverge between models. The inter annotator agreement between GPT models was 0.79 (p-value < 0.0001), which is considered substantial [66]. In comparison, the human Cohen Kappa score inter annotator agreement was of 0.82, with a p-value < 0.0001, which is considered near perfect agreement. Objectively, the difference is not much, 0.03, however it does show that humans agree slightly better than the GPT models. Note that our human annotators worked independently and did not know or communicated with each other.

#### 4.1.2 Results for LLM to improve weakly supervised dataset creation

In the second evaluation, GPT-4 labeled 68,561 tweets as vaccine self-reports and 681,439 tweets as vaccine chatter. GPT-3.5 labeled 66,288 tweets as vaccine self-reports and 683,712 as vaccine chatter. While it might seem that GPT-4 labels more tweets, we are not sure they are correctly labeled and they have not been annotated by a human. Due to this fact, there are no comments on accuracy, the idea behind this exercise is to then feed this data as part of the fine-tuning step for the previously identified BERT-like models.

After fine-tuning COVID-Twitter-BERT and BERTweet, Table 2 shows the correct tweet labeling results achieved. It is very interesting to see that a fine-tuned COVID-Twitter-BERT performs marginally better than GPT-4 (and GPT-3.5) at labeling both tweet classes. While the improvement is marginal, it goes to show that a properly fine-tuned model does outperform a more complex model, at least in this scenario. Another interesting finding is that BERTweet performs slightly worse than GPT-4,

but better than GPT-3.5. This is most likely due to the training data for BERTweet not being focused on COVID-related tweets.

**Table 2.** Correct tweet labeling results for BERT models.

| Label                     | COVID-Twitter-BERT | %      | BERTweet | %      |
|---------------------------|--------------------|--------|----------|--------|
| Self reported vaccination | 2,045              | 83.33% | 1,897    | 77.30% |
| Vaccine chatter           | 19,012             | 95.32% | 18,457   | 92.53% |

In order to assess the actual labeling agreement between our top two models (GPT-4 and COVID-Twitter-BERT) we measured the Cohen Kappa score, which was quite surprising to learn that it was 0.85 with a p-value  $< 0.0001$ . This means that both models have a high level of agreement in which tweets they labeled, even more so than humans. Additionally, we calculated the Fleiss Kappa statistic [67] between all annotators, showing that we have a score of 0.76 with a p-value  $< 0.0001$ . This showcases that both the models and the humans mostly agree on what class the tweets should be labeled.

## 5 Conclusion

In conclusion, our study has several important findings:

GPT models perform fairly well, in a zero-shot, task of properly labeling social media data, tweets in this case. However, at larger scales the number of incorrect classifications might start becoming problematic, particularly depending on the downstream task that said data will be used for.

- When leveraging GPT models alongside weak supervision techniques to identify ‘silver-standard’ data, we can use data augmentation with higher confidence. These resulting ‘electrum datasets’ could be leveraged for further fine-tuning with potentially a considerable amount of less noise than just using weak supervision alone.
- Fine-tuned BERT models are still not obsolete, as we showed them outperforming GPT-4 for labeling social media data, self reported vaccine tweets in this case. While this comparison might be unfair, the point we show is that combining approaches leads to better results.
- Lastly, we show with our cost analysis that it is very cost effective to label data using GPT models, and that the results data is usable for downstream tasks. While we would continue to use human annotators to label data for our NER tasks, we can consider labeling less data to have equally or better performing systems in downstream tasks.




While we show that GPT models perform well, this work does not advocate for the replacing of human labeled data with GPT-annotated data. Our argument is to show that leveraging multiple approaches together, and fine-tuning, leads to potentially better and more generalizable results. The limitations of our work are clear: we only used one task - self reported vaccine tweet labeling, we only fine-tuned two different BERT models, and we did not evaluate how large our ‘electrum dataset’ should be to fine-tune a model enough to achieve solid performance. All these are future research directions that would greatly inform the community.

## References

1. Agichtein, E., Castillo, C., Donato, D., Gionis, A., Mishne, G.: Finding high-quality content in social media. In: *Proceedings of the 2008 International Conference on Web Search and Data Mining*. pp. 183–194. Association for Computing Machinery, New York, NY, USA (2008). <https://doi.org/10.1145/1341531.1341557>.
2. Pershad, Y., Hangge, P.T., Albadawi, H., Oklu, R.: Social Medicine: Twitter in Healthcare. *J. Clin. Med. Res.* 7, (2018). <https://doi.org/10.3390/jcm7060121>.
3. Xue, J., Chen, J., Hu, R., Chen, C., Zheng, C., Su, Y., Zhu, T.: Twitter Discussions and Emotions About the COVID-19 Pandemic: Machine Learning Approach. *J. Med. Internet Res.* 22, e20550 (2020). <https://doi.org/10.2196/20550>.
4. Ratner, A., Bach, S., Varma, P., Ré, C.: Weak supervision: the new programming paradigm for machine learning. *Hazy Research*. Available via <https://dawn.cs.berkeley.edu/> (2019).
5. Cutler, J., Culotta, A.: Using weak supervision to scale the development of machine-learning models for social media-based marketing research. *Applied Marketing Analytics*. 5, 159–169 (2019).
6. Chandra, A.L., Desai, S.V., Balasubramanian, V.N., Ninomiya, S., Guo, W.: Active learning with point supervision for cost-effective panicle detection in cereal crops. *Plant Methods*. 16, 34 (2020). <https://doi.org/10.1186/s13007-020-00575-8>.
7. Shin, C., Li, W., Vishwakarma, H., Roberts, N., Sala, F.: Universalizing Weak Supervision, <http://arxiv.org/abs/2112.03865>, (2021).
8. Ratner, A., De Sa, C., Wu, S., Selsam, D., Ré, C.: Data Programming: Creating Large Training Sets, Quickly. *Adv. Neural Inf. Process. Syst.* 29, 3567–3575 (2016).
9. Zhang, J., Hsieh, C.-Y., Yu, Y., Zhang, C., Ratner, A.: A Survey on Programmatic Weak Supervision, <http://arxiv.org/abs/2202.05433>, (2022).
10. Munro, R., Monarch, R.: Human-in-the-Loop Machine Learning: Active Learning and Annotation for Human-centered AI. *Simon and Schuster* (2021).
11. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Others: Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* 33, 1877–1901 (2020).
12. Wang, Y., Sohn, S., Liu, S., Shen, F., Wang, L., Atkinson, E.J., Amin, S., Liu, H.: A clinical text classification paradigm using weak supervision and deep representation. *BMC Med. Inform. Decis. Mak.* 19, 1 (2019). <https://doi.org/10.1186/s12911-018-0723-6>.
13. Deriu, J., Lucchi, A., De Luca, V., Severyn, A., Müller, S., Cieliebak, M., Hofmann, T., Jaggi, M.: Leveraging Large Amounts of Weakly Supervised Data for Multi-Language Sentiment Classification. In: *Proceedings of the 26th International Conference on World*

- Wide Web. pp. 1045–1052. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE (2017). <https://doi.org/10.1145/3038912.3052611>.
14. Agarwal, V., Podchiyska, T., Banda, J.M., Goel, V., Leung, T.I., Minty, E.P., Sweeney, T.E., Gyang, E., Shah, N.H.: Learning statistical models of phenotypes using noisy labeled training data. *J. Am. Med. Inform. Assoc.* 23, 1166–1173 (2016). <https://doi.org/10.1093/jamia/ocw028>.
  15. Zamani, H., Bruce Croft, W.: On the Theory of Weak Supervision for Information Retrieval, <http://dx.doi.org/10.1145/3234944.3234968>, (2018). <https://doi.org/10.1145/3234944.3234968>.
  16. Tekumalla, R., Asl, J.R., Banda, J.M.: Mining Archive. org’s Twitter Stream Grab for Pharmacovigilance Research Gold. In: Proceedings of the International AAAI Conference on Web and Social Media. pp. 909–917 (2020).
  17. Tekumalla, R., Banda, J.M.: Using weak supervision to generate training datasets from social media data: a proof of concept to identify drug mentions. *Neural Comput. Appl.* (2021). <https://doi.org/10.1007/s00521-021-06614-2>.
  18. Tekumalla, R., Banda, J.M.: An Enhanced Approach to Identify and Extract Medication Mentions in Tweets via Weak Supervision. In: Proceedings of the BioCreative VII Challenge Evaluation Workshop (2021).
  19. Tekumalla, R., Banda, J.M.: Identifying epidemic related Tweets using noisy learning. In: Proceedings of LatinX in Natural Language Processing Research Workshop at NAACL 2022.
  20. Tekumalla, R., Banda, J.M.: TweetDIS: A Large Twitter Dataset for Natural Disasters Built using Weak Supervision. In: 2022 IEEE International Conference on Big Data (Big Data). pp. 4816–4823 (2022). <https://doi.org/10.1109/BigData55660.2022.10020214>.
  21. Tekumalla, R., Banda, J.M.: An Empirical Study on Characterizing Natural Disasters in Class Imbalanced Social Media Data using Weak Supervision. In: 2022 IEEE International Conference on Big Data (Big Data). pp. 4824–4832 (2022). <https://doi.org/10.1109/BigData55660.2022.10020594>.
  22. Saab, K., Dunnmon, J., Ré, C., Rubin, D., Lee-Messer, C.: Weak supervision as an efficient approach for automated seizure detection in electroencephalography. *NPJ Digit Med.* 3, 59 (2020). <https://doi.org/10.1038/s41746-020-0264-0>.
  23. Fries, J.A., Varma, P., Chen, V.S., Xiao, K., Tejeda, H., Saha, P., Dunnmon, J., Chubb, H., Maskatia, S., Fiterau, M., Delp, S., Ashley, E., Ré, C., Priest, J.R.: Weakly supervised classification of aortic valve malformations using unlabeled cardiac MRI sequences. *Nat. Commun.* (2019). <https://doi.org/10.1101/339630>.
  24. Saab, K., Dunnmon, J., Goldman, R., Ratner, A., Sagreiya, H., Ré, C., Rubin, D.: Doubly Weak Supervision of Deep Learning Models for Head CT. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. pp. 811–819. Springer International Publishing (2019). [https://doi.org/10.1007/978-3-030-32248-9\\_90](https://doi.org/10.1007/978-3-030-32248-9_90).
  25. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, <http://arxiv.org/abs/1810.04805>, (2018).
  26. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training, <https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf>, last accessed 2023/06/17.
  27. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: ALBERT: A Lite BERT for Self-supervised Learning of Language Representations, <http://arxiv.org/abs/1909.11942>, (2019).

28. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A Robustly Optimized BERT Pretraining Approach, <http://arxiv.org/abs/1907.11692>, (2019).
29. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, <http://arxiv.org/abs/1910.01108>, (2019).
30. Clark, K., Luong, M.-T., Le, Q.V., Manning, C.D.: ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators, <http://arxiv.org/abs/2003.10555>, (2020).
31. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V.: Xlnet: Generalized autoregressive pretraining for language understanding. *Adv. Neural Inf. Process. Syst.* 32, (2019).
32. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* 21, 5485–5551 (2020).
33. Shueybi, M., Patwary, M., Puri, R., LeGresley, P., Casper, J., Catanzaro, B.: Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism, <http://arxiv.org/abs/1909.08053>, (2019).
34. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, <http://arxiv.org/abs/1910.13461>, (2019).
35. Martin, L., Muller, B., Suárez, P.J.O., Dupont, Y., Romary, L., de la Clergerie, É.V., Seddah, D., Sagot, B.: CamemBERT: a Tasty French Language Model, <http://arxiv.org/abs/1911.03894>, (2019).
36. Müller, M., Salathé, M., Kummervold, P.E.: COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter, <http://arxiv.org/abs/2005.07503>, (2020).
37. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics.* 36, 1234–1240 (2020). <https://doi.org/10.1093/bioinformatics/btz682>.
38. Beltagy, I., Lo, K., Cohan, A.: SciBERT: A Pretrained Language Model for Scientific Text, <http://arxiv.org/abs/1903.10676>, (2019).
39. Huang, K., Altosaar, J., Ranganath, R.: ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission, <http://arxiv.org/abs/1904.05342>, (2019).
40. Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., Androutsopoulos, I.: LEGAL-BERT: The Muppets straight out of Law School, <http://arxiv.org/abs/2010.02559>, (2020).
41. Liu, Z., Huang, D., Huang, K., Li, Z., Zhao, J.: Finbert: A pre-trained financial language representation model for financial text mining. In: *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*. pp. 4513–4519 (2021).
42. Yang, Y., Uy, M.C.S., Huang, A.: FinBERT: A Pretrained Language Model for Financial Communications, <http://arxiv.org/abs/2006.08097>, (2020).
43. Araci, D.: FinBERT: Financial Sentiment Analysis with Pre-trained Language Models, <http://arxiv.org/abs/1908.10063>, (2019).
44. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language Models are Unsupervised Multitask Learners, <https://openai.github.io/2020/05/27/GPT%E6%8A%80%E6%9C%AF%E5%88%9D%E6%8E%A2/language-models.pdf>, last accessed 2023/06/17.

45. Settles, B.: Active learning literature survey. University of Wisconsin-Madison Department of Computer Sciences (2009).
46. Veselovsky, V., Ribeiro, M.H., West, R.: Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks, <http://arxiv.org/abs/2306.07899>, (2023).
47. Gilardi, F., Alizadeh, M., Kubli, M.: ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks, <http://arxiv.org/abs/2303.15056>, (2023).
48. He, X., Lin, Z., Gong, Y., Jin, A.-L., Zhang, H., Lin, C., Jiao, J., Yiu, S.M., Duan, N., Chen, W.: AnnoLLM: Making large language models to be better crowdsourced annotators, <http://arxiv.org/abs/2303.16854>, (2023).
49. Møller, A.G., Dalsgaard, J.A., Pera, A., Aiello, L.M.: Is a prompt and a few samples all you need? Using GPT-4 for data augmentation in low-resource classification tasks, <http://arxiv.org/abs/2304.13861>, (2023).
50. Huang, F., Kwak, H., An, J.: Is ChatGPT better than human annotators? Potential and limitations of ChatGPT in explaining implicit hate speech, <http://arxiv.org/abs/2302.07736>, (2023).
51. Yu, D., Li, L., Su, H., Fuoli, M.: Using LLM-assisted Annotation for Corpus Linguistics: A Case Study of Local Grammar Analysis, <http://arxiv.org/abs/2305.08339>, (2023).
52. Kuzman, T., Mozetic, I., Ljubešić, N.: Chatgpt: Beginning of an end of manual linguistic data annotation? use case of automatic genre identification. arXiv e-prints, pages arXiv--2303. (2023).
53. Zhu, Y., Zhang, P., Haq, E.-U., Hui, P., Tyson, G.: Can ChatGPT Reproduce Human-Generated Labels? A Study of Social Computing Tasks, <http://arxiv.org/abs/2304.10145>, (2023).
54. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? . In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. pp. 610–623. Association for Computing Machinery, New York, NY, USA (2021). <https://doi.org/10.1145/3442188.3445922>.
55. Reiss, M.V.: Testing the Reliability of ChatGPT for Text Annotation and Classification: A Cautionary Remark, <http://arxiv.org/abs/2304.11085>, (2023).
56. Beware the hype: ChatGPT didn't replace human data annotators, <https://news.techworkerscoalition.org/2023/04/04/issue-5/>, last accessed 2023/06/17.
57. Banda, J.M., Tekumalla, R., Wang, G., Yu, J., Liu, T., Ding, Y., Artemova, E., Tutubalina, E., Chowell, G.: A Large-Scale COVID-19 Twitter Chatter Dataset for Open Scientific Research—An International Collaboration. *Epidemiologia*. 2, 315–324 (2021). <https://doi.org/10.3390/epidemiologia2030024>.
58. Weissenbacher, D., Banda, J., Davydova, V., Estrada Zavala, D., Gasco Sánchez, L., Ge, Y., Guo, Y., Klein, A., Krallinger, M., Leddin, M., Magge, A., Rodriguez-Esteban, R., Sarker, A., Schmidt, L., Tutubalina, E., Gonzalez-Hernandez, G.: Overview of the Seventh Social Media Mining for Health Applications (#SMM4H) Shared Tasks at COLING 2022. In: Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task. pp. 221–241. Association for Computational Linguistics, Gyeongju, Republic of Korea (2022).
59. Tekumalla, R., Asl, J.R., Banda, J.M.: Mining Archive.org's Twitter Stream Grab for Pharmacovigilance Research Gold. *ICWSM*. 14, 909–917 (2020). <https://doi.org/10.1609/icwsml.v14i1.7357>.
60. Solmaz, G., Cirillo, F., Maresca, F., Kumar, A.G.A.: Label Augmentation with Reinforced Labeling for Weak Supervision, <http://arxiv.org/abs/2204.06436>, (2022).

61. Robinson, J., Jegelka, S., Sra, S.: Strength from Weakness: Fast Learning Using Weak Supervision. In: Iii, H.D. and Singh, A. (eds.) *Proceedings of the 37th International Conference on Machine Learning*. pp. 8127–8136. PMLR (13--18 Jul 2020).
62. Nguyen, D.Q., Vu, T., Tuan Nguyen, A.: BERTweet: A pre-trained language model for English Tweets. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. pp. 9–14. Association for Computational Linguistics, Online (2020). <https://doi.org/10.18653/v1/2020.emnlp-demos.2>.
63. Magge, A., Klein, A., Miranda-Escalada, A., Ali Al-Garadi, M., Alimova, I., Miftahutdinov, Z., Farre, E., Lima López, S., Flores, I., O’Connor, K., Weissenbacher, D., Tutubalina, E., Sarker, A., Banda, J., Krallinger, M., Gonzalez-Hernandez, G.: Overview of the Sixth Social Media Mining for Health Applications (#SMM4H) Shared Tasks at NAACL 2021. In: *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*. pp. 21–32. Association for Computational Linguistics, Mexico City, Mexico (2021). <https://doi.org/10.18653/v1/2021.smm4h-1.4>.
64. AWS Pricing Calculator, <https://calculator.aws/#/addService/SageMakerGroundTruth>, last accessed 2023/06/22.
65. Cohen, J.: A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.* 20, 37–46 (1960). <https://doi.org/10.1177/001316446002000104>.
66. McHugh, M.L.: Interrater reliability: the kappa statistic. *Biochem. Med.* . 22, 276–282 (2012). <https://doi.org/10.1016/j.jocd.2012.03.005>.
67. Fleiss, J.L.: Measuring nominal scale agreement among many raters. *Psychol. Bull.* 76, 378–382 (1971). <https://doi.org/10.1037/h0031619>.