Lexicon Annotation with LLM: a proof of concept with ChatGPT

Francisco S. Marcondes[1], Adelino de C. O. S. Gala[1], Manuel Rodrigues[1], José João Almeida[1] and Paulo Novais[1]

[1]ALGORITMI Research Centre/LASI, University of Minho, Braga, Portugal
francisco.marcondes@algoritmi.uminho.pt, adelinogala@gmail.com,
manuel.rodrigues@algoritmi.uminho.pt, jj@di.uminho.pt, pjon@di.uminho.pt

Abstract. Lexicon annotation is a critical yet time-consuming task that can hold back the progress of language-intensive projects. This paper explores the potential of Large Language Models (LLMs) to automate lexicon annotation, traditionally performed by humans. We present a proof of concept by evaluating ChatGPT's performance on annotating VADER's sentiment lexicon. Our findings demonstrate that ChatGPT achieves fair performance in this task, suggesting that LLMs can operate as a valuable tool for initial annotations, with subsequent refinements by domain specialists. This approach could significantly accelerate lexicon development and maintenance while balancing efficiency and accuracy. Our study provides insights into the capabilities and limitations of LLMs in lexicon annotation, leading the way for further research in automating linguistic resources development.

Keywords: LLMs, Lexicon annotation, ChatGPT, NLP.

1 Introduction

Lexicon annotation is a fundamental task in Natural Language Processing (NLP) that involves assigning sentiment scores or polarity labels (e.g., positive, negative, neutral) to words or phrases. These annotated lexicons serve as valuable resources for various NLP applications, including sentiment analysis, opinion mining, and text classification. The importance of lexicon annotation is underscored by its widespread use in sentiment analysis tools like VADER (Valence Aware Dictionary and sEntiment Reasoner). VADER, a lexicon-based model, relies on a pre-defined lexicon of words with associated sentiment scores to determine the overall sentiment of a text. However, manually creating and maintaining such lexicons consumes many resources and can lead to inconsistencies. This time-consuming and resource-intensive process often hinders the progress of language-intensive projects. Finding efficient alternatives is crucial for advancing NLP research and applications.

The concept of lexicon annotation is not new and has been explored in various contexts. In (Krista et al., 2013), the use of crowd-sourcing platforms like Amazon Mechanical Turk is examined to gather sentiment annotations from multiple human annotators. However, this approach can be costly and may still suffer from subjectivity issues.

The use of LLMs like ChatGPT represents a novel and promising direction in lexicon annotation, as it leverages the vast knowledge and language understanding capabilities of these models to automate a task that has traditionally relied on human judgment. Using LLMs for lexicon annotation is of paramount importance in advancing this field of research. Lexicon annotation is a fundamental task in NLP. However, as mentioned, it is a labor-intensive and time-consuming process when done manually. LLMs, with their ability to process and understand vast amounts of textual data, offer a promising solution to automate and streamline this task. By leveraging LLMs, researchers and practitioners can significantly reduce the time and effort required for lexicon annotation, enabling them to create more comprehensive and up-to-date lexicons. This, in turn, can lead to improved performance in various NLP applications, such as sentiment analysis, opinion mining, and text classification. Furthermore, the use of LLMs can help to mitigate the potential for human bias in annotation, ensuring greater consistency and objectivity in the resulting lexicons.

Considering modern sentiment analysis, it is possible to suggest that there are currently two widely used implementations with similar performance, as shown in the table 1: VADER (Hutto and Gilbert,2014) and BERT (Devlin et al., 2018); the former being lexicon-based and the latter a transformer model. Currently, ChatGPT is being considered for annotation (Gilardi et al., 2023; Huang et al.,2023) and other NLP tasks (Qin et al., 2023), which would, for example, allow VADER to improve its performance by expanding and updating its lexicon.

As a point of reference, in 2014 COVID was not yet a negative word for the public, and there was no King Charles III. However, there is no economically efficient way of both checking the sentiment of words already in the lexicon and expanding the lexicon with new words. In this sense, to the extent that ChatGPT (or LLMs in general) provides reasonable performance, annotation could be done on a large scale, possibly involving a few specialists for performing adjustments.

The objective of this paper is to evaluate the use of ChatGPT for annotating sentiment lexicons. The VADER's lexicon is considered for case study. The aim of this paper is not to settle the question, but to provide a proof of concept that enable the discussion of the potential of such an approach to address the issues presented. In this sense, where they provide useful insights, qualitative evaluations are included, even if based on small data sets. Python modules used are: pandas, numpy, scipy, sklearn, and matplotlib.

Table 1 - VADER (V) and BERT (B) comparison for sentiment analysis. Ref.: (1) (Majidpour and Al-Barznji, 2022); (2) (Nair et al., 2021); (3) (Xie et al., 2023); (4) (Waterman, 2022); and (5) (Fontes et al., 2023).

Ref.	Accuracy		Recall		Precision		F1	
	V	B	V	B	V	B	V	B
(1)	0.75	0.72	0.74	0.71	0.75	0.72	0.75	0.71
(2)	0.88	‘.92	0.94	0.93	0.90	0.94	0.92	0.88

(3)	0.55	0.70	0.49	0.65	-	-	0.51	0.67
(4)	0.39	0.39	0.39	0.39	0.41	0.42	0.51	0.53
(5)	0.78	0.90	0.78	0.99	0.77	0.81	0.78	0.89
μ	0.67	0.73	0.67	0.73	0.71	0.72	0.69	0.74
Δ	0.06		0.06		0.01		0.05	

2 Annotating VADER's Lexicon

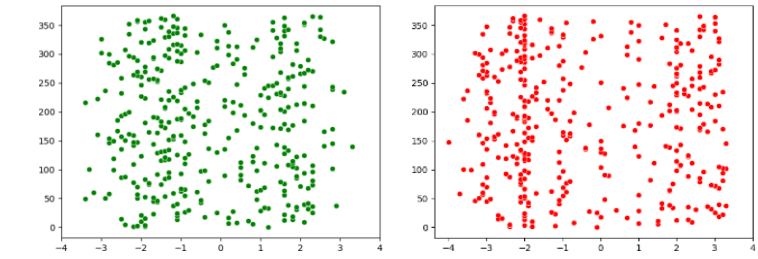
The VADER lexicon is a set of words collected from Twitter and manually annotated by ten raters using a Likert scale ranging from extremely negative −4 to extremely positive +4 (Hutto and Gilbert, 2014). Each word in the lexicon is composed by the ten ratings, together with these ratings mean and standard deviation. Only the mean value is actually used to calculate the sentiment. VADER's lexicon has 14 repeated words which are removed, leaving only the first value; misspelled words, emoticons, etc. which are left as they are. To assess ChatGPT's ability to annotate sentiments, a sample of 367 words from VADER's lexicon is queried (the lexicon's "population" is 7503, so confidence is 0.95 and error is 0.05). Note that ChatGPT allows variation in its output (the heat of the model cannot be changed). Then, to reproduce the methodology used on VADER, ten generation rounds are run (see prompt 1), each on a single, new chat session to avoid the context window; then their mean and standard deviation are calculated. For reference, a PCA analysis of the generations against the mean shows that the value becomes negligible after the ninth generation (number of adopters for the next experiments). Also, the comparison with a random number gives a low p-value for both the VADER and ChatGPT generations, which shows that they are not random sequences.

Prompt 1 *Classify the words sentiments score using the following model: [extremely negative = -4, very negative = -3, moderately negative = -2, slightly negative = -1, neutral = 0,076 slightly positive = 1, moderately positive = 2, very positive = 3, extremely positive = 4] Provide the results in a word comma format. Provide the resulting vertical list only with the words and scores. No descriptions.*

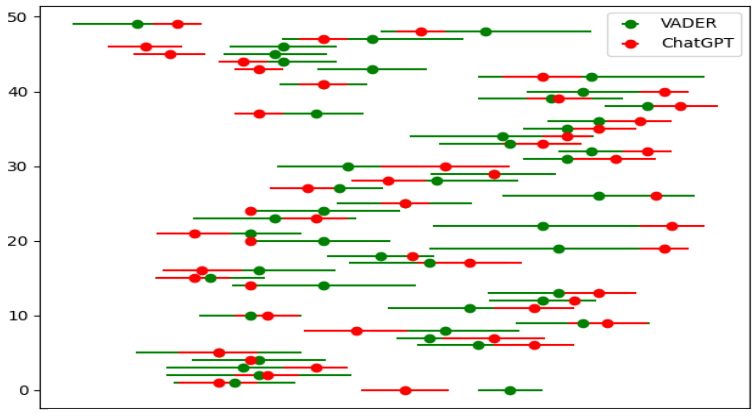
Table 2 - Data Excerpt (ChatGPT)

word	ChatGPT	VADER	Random
pityingly	-1.0	-1.0	-0.6
poisoned	-3.2	-2.2	0.3
poisoners	-3.2	-3.1	0.9
poverty	-2.7	-2.3	0.6

For an illustration, table 2 presents an excerpt of ChatGPT and random results compared to VADER. Refer to figures 1a and 1b for a dot-plot visualization of data; also, figure 1c shows a comparison including the standard deviation.



(a) VADER, baseline. (b) Ordered dataset, prompt 1.



(c) Excerpt of Ordered dataset, prompt 1 (n=50).

Figure 1 - Data visualization.

The mean distance of ChatGPT's results from VADER's is of −0.078 ± 0.874, the of metrics can be found in table 3 (this table summarizes the results of the efforts still to be presented in the paper). Since the error is of about 1, it is possible to say that ChatGPT may not being able of differentiating fine-grained sentiment scale such as moderately and slightly but, also that error margin that could also be found on a different annotation of VADER's lexicon (the average standard deviation in VADER is 0.921), thus, ChatGPT shown a possible result.

Table 3- Summary Table. Legend, see (Kunapuli, 2023): P. means *prompt*; Difference is the mean difference of target and generation; MAE means *mean absolute error*; and MSE means *mean squared error*.

Data	P.	Difference	MAE	MSE	R2
random		−0.277 ± 1.847	1.634	3.684	−0.281
Ordered	1	−0.078 ± 0.874	0.715	0.768	0.733
Shuffled	1	−0.039 ± 0.812	0.660	0.659	0.771
Shuffled	2	−0.019 ± 0.914	0.684	0.834	0.710
Shuffled	3	0.002 ± 0.781	0.616	0.608	0.789

In addition, ChatGPT accurately identify the polarities. There were 22 out of 367 words

(≈ 6%) situations where ChatGPT output inverted the polarity presented in VADER, see table 4a for an excerpt. Consider the word *tricking*, asking ChatGPT to produce a sentence with it, outputs:

> *She was wary of his smooth talk, fearing he might be tricking her into believing his false promises.* (i)

Note that when the word is scored out of context, as it was, humans may refer to the lemma of the word, where *trick* is certainly slightly positive, but its inflection such as tricking is often used in a deceptive context, resulting in a negative semantics. Then it is possible to suggest that ChatGPT scored this word better than the human annotators.

There are also issues related to the pragmatics (Qiu et al., 2023). As can be seen in table 2, the words *poisoned* and *poisoners* received the same score from ChatGPT, while the second is rated worse by humans. This is due to this word being close in the embedding space, whose pragmatics scapes ChatGPT. Consider then table 4b, since all these words are close to each other in the embedding space, ChatGPT scores them with the same value (a slight difference is seen by requesting ChatGPT to use a continuous scale). This is suggesting that the sequence of words might be defining the context used by the ChatGPT to scoring the words.

A subtlety to be noticed is the "straight line" on the −2 value in figure 1b, for an excerpt see table 5a. VADER's lexicon is lexicographically ordered, given the attention window, this reinforces the claim that the order of the words are defining the context, then driving the result.

By shuffling the data set (Lew, 2023) and following the same procedure as before, it is obtained a slight variation, see table 5a yet the "straight lines" increased, see figure 5b, and the average error de- creased, see table 3. This drives the discussion to the prompt.

Table 4 – Data Excerpts

word	ChatGPT	Vader
robing	0.6	-1.5
Startling	-0.7	0.3
toughies	0.2	-0.6
tricking	-2.0	0.1
whitewash	-1.4	0.1
grossular	0.0	-0.3
supremacy	-1.1	0.2
sillibub	1.0	-0.1
<:	-0.9	2.1

(a) Data Excerpt of words with Inverted Polarities.

word	Discrete	ChatGPT	Vader
poison	-3	-3.2	-2.5
poisoned	-3	-3.4	-2.2
poisoner	-3	-3.1	-2.7
poisoners	3	-3.2	-3.1
poisoning	3	-3.3	-2.8
poisonings	3	-3.4	-2.4
poisonous	3	-3.0	-2.7
poisonously	3	-3.2	-2.9
poisons	3	-3.1	-2.7

(b) ChatGPT´s output of similar words (single generation).

Roughly, the transformer is trained to identify embeddings that usually follow a sequence of embeddings. Then, given a sequence of embeddings as input, with each token serving as a context for the others, it reweights the embeddings of these tokens

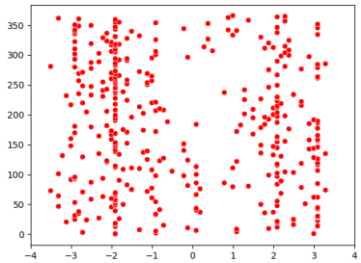
(Jurafsky and Martin, 2023). This in turn helps to determine the semantics of these tokens and to calibrate the probability of the next one. The idea is then to ask ChatGPT to make a sentence using the word to guide the scoring context, resulting in prompt 2 (the shuffled dataset was used).

Prompt 2 *Create a different sentence for each word. Then classify the word sentiments score using the following model: [extremely negative=-4, very negative=-3, moderately negative=-2, slightly negative=-1, neutral=0, slightly positive=1, moderately positive=2, very positive=3, extremely positive=4]. Provide the results in a word comma format. Pretty print the result in the format [sentence, word, scores]. No descriptions, no brackets.*

Table 5 - The "straight line" issue

word	O.	V.	S.
smuggles	-2.0	-1.7	-2.0
sobs	-2.0	-2.5	-2.1
trickers	-2.0	-1.4	-1.8
tricking	-2.0	0.1	-1.7
weary	-2.0	-1.1	-2.0
insipid	-2.0	-2.0	-2.0
obsessively	-2.0	-0.4	-2.0
paranoids	-2.0	-1.6	-2.0

(a) Data on the "straight line" (at −2) from the figure 1c. Legend: O. means ordered; V. VADER and S. Shuffled.



(b) ChatGPT's generation of shuffled data using prompt 1. The "straight lines" can be seen on −3, −2, 2 and 3.

As a result, the visualization improved, see figure 2a, but the error increased, see table 3. This indicates that these "straight lines" are not necessarily a problem. ChatGPT produced repeated sentences, but in not a significative amount. Also, it refused to generated sentences to blacklisted words, such as *molesters*, in the first run.

2.1 Experiments

Finally, another explainability effort (Deldjoo, 2023) has to be made in order to understand the output. In this sense, a series of experiments are carried out, but on a very small sample (n = 6) and with a single generation. The intention is to build an intuition about the behaviour of ChatGPT. The experiments are divided into two parts. The first aims to understand how the Likert scale actually participates in the prompt 1 and the second to understand the meaning of the numbers (the experiments are all variations upon prompt 1). The first set of experiments are:

1. replace the scale by the interval (−4.0..4.0)
2. replace the previous interval by (1.0..9.0)
3. replace the previous interval by (9.0..1.0)
4. replace the previous interval by (A..I)

The result of the first set of experiments is shown in the table 6. As can be seen, the model behaves correctly in experiments #1 and #2, but loses reference in experiments #3 and #4. This suggests that the scoring is based on associating negative scores with lower numbers. If the order is reversed or replaced by letters, the coherence of the model cannot be observed.

Table 6 - Prompt Scrutiny I (expected variation, ±1)

word	Vader	ChatGPT	Experiments			
			#1	#2	#3	#4
pityingly	-1.0	-1.0	-2.5	-2.5 (3.5)	-2.5 (7.5)	3.0 (H)
thrilling	2.1	3.0	3.5	3.2 (8.2)	-4.0 (9.0)	0.0 (E)
poisoned	-2.2	-3.2	-3.5	-3.8 (2.8)	2.0 (3.0)	2.0 (G)
assure	1.4	1.8	2.0	1.7 (6.7)	-3.5 (8.5)	-3.0 (B)
discards	-1.0	-2.0	-1.5	-1.1 (4.1)	3.0 (2.0)	-1.0 (D)
challenge	0.3	1.0	1.5	2.3 (7.3)	-1.5 (6.5)	-2.0 (C)

The second part of the experiments are:

5. remove the word “sentiment”
6. replace the word “classify” by “score”
7. replace “score” related words by “classify”
8. include the word “sentiment” by “classify”
9. replace the word “sentiment” by the scale

The result of the second set of experiments is shown in the table 7. As can be seen, removing the word “sentiment” and using the word “score”, the numbers are coherent, but replacing “score” with “classify”, the model tags each word with a number. A *tentative* explanation is therefore that the scoring word, whose embedding space is related to numbers, associates the words with the numbers, with words of negative valence being associated with lower numbers. Asking to classify “sentiment” had no apparent effect. Asking for classification based on a Likert scale made the model work properly.

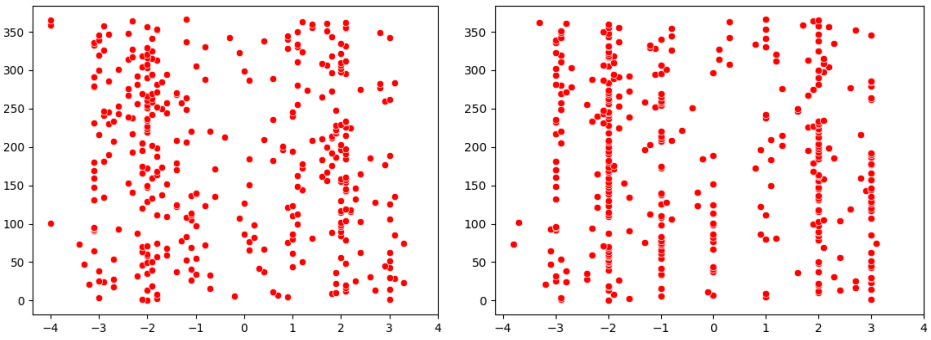
The output of experiment #9 seems to be the most controlled output. An additional evaluation, following the same procedure as the previous one, using prompt 3 with shuffled data is then worth- while. Again, the number of “straight lines” in- creased, see figure 2b, while the average error de- creased, see table 3. It is not clear why the model produces these lines, perhaps it is a regressive nature of the classification. Nevertheless, it shows that the prompt 3 produces more stable results than the previous one. As it also produced the lower average error, it can be considered as a baseline prompt to compare with in future work. For com- parison, a data extract from VADER, prompt 2 and prompt 3 is shown in figure 2c.

Prompt 3 *Classify the words using the following model [extremely negative=-4, very negative=-3, moderately negative=-2, slightly negative=-1, neutral=0, slightly positive=1, moderately*

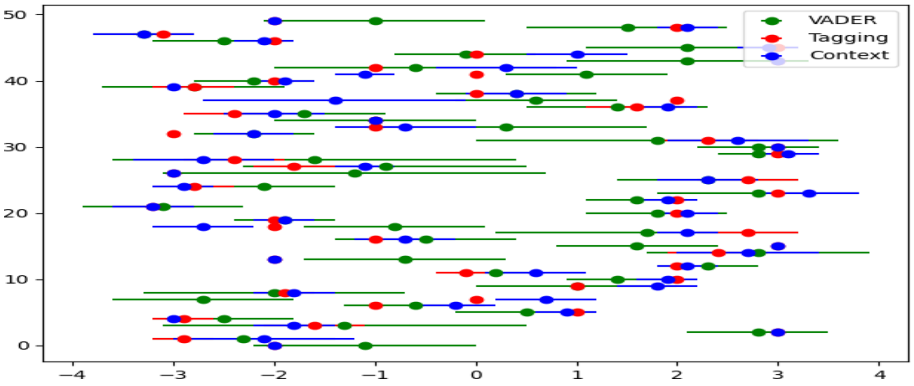
positive=2, very positive=3, extremely positive=4]. Provide the results in a word comma format. Provide the resulting vertical list only with the words and classification numbers. No descriptions.

Table 7 - Prompt Scrutiny II (expected variation, ±1)

word	VADER	ChatGPT	Experiments				
			#5	#6	#7	#8	#9
pityingly	-1.0	-1.0	2.0	-3.0	-2.0	2.5	-1.0
thrilling	2.1	3.0	3.5	2.5	4.0	3.8	3.0
poisoned	-2.2	-3.2	-3.0	-2.7	-4.0	-3.2	-3.0
assure	1.4	1.8	2.5	1.8	3.0	2.0	2.0
discards	-1.0	-2.0	-1.5	-1.5	2.0	-1.5	-2.0
challenge	0.3	1.0	2.0	3.2	1.0	2.7	2.0



(a) Shuffled dataset, prompt 2 (b) Shuffled dataset, prompt 3



(c) Excerpt of Shuffled dataset, prompt 2 and 3 (n=50.

Figure 2 - Data visualization.

3 Results and final remarks

The developed work shows that ChatGPT can be used as a productivity tool for annotating sentiment lexicons, as far as it receives a final adjustment from a specialist. The justification for this approach is that, given the results presented in the table 1, it may not be necessary to use an LLM for most tasks. However, training, running and storing these models has a high environmental footprint (Mok, 2023; Li et al., 2023; Maslej et al., 2023), VADER's footprint is comparatively negligible (Lannelongue et al., 2021). However, keeping VADER's lexicon up to date has not only economic costs, but also the social costs inherent in crowd working (Casleret al., 2013; Moreschi et al., 2020). This can then be seen as a compromise solution.

On starting this paper, a major concern was whether the numbers delivered by ChatGPT are random. The several endeavours performed in this paper shown that they are coherent. Another one is the impact of the prompt. Data shown that the prompt pushes the model to suitable or unsuitable contexts, then prompt design is an art that needs to be mastered. Note also that this paper performed a daring endeavour either by not seeking for broad grained classification and all prompts used are zero-shot learning (Zhang et al., 2023).

Note that response bias in Likert scales is a known fact (Van Vaerenbergh and Thomas, 2013). One of these phenomena is "straightlining" (Schonlau and Toepoel, 2015) as described in the paper.

As a limitation, this paper evaluates a single lexicon (VADER) against a single LLM (ChatGPT). The evaluation sample used in VADER is not significant, but also subjective. So not a gold standard. ChatGPT is a proprietary tool with unknown inner workings in a constantly changing setup. Reproducibility in chatGPT is constrained by these issues. No particular risks, other than those associated with VADER and ChatGPT, see (Marcondeset al., 2024), could be identified for this work. It took about 120 runs in ChatGPT.

4 Conclusions

This work bridges the gap between traditional lexicon-based approaches and cutting-edge AI technologies, potentially offering a new paradigm for lexicon development and maintenance. It provides a rigorous, quantitative assessment of ChatGPT's annotation capabilities, comparing it directly with the well-established VADER lexicon. This empirical approach offers valuable insights into the strengths and limitations of LLMs in fine-grained sentiment scoring tasks.

Through detailed experimentation with various prompts and scoring scales, this work contributes significant methodological insights for effectively leveraging LLMs in NLP tasks. This exploration of prompt engineering is particularly valuable given the growing importance of LLMs in AI research. A critical challenge in sentiment analysis is also adressed: maintaining the computational efficiency of lexicon-based methods while

improving accuracy and adaptability. It presents a potential middle ground between resource-intensive deep learning models and static lexicons.
As a conclusion, it might be said that ChatGPT competently annotated the sample of VADER's lexicon. As presented in the paper, there are situations where ChatGPT provided a better annotation and others when humans provided a better annotation. Therefore, ChatGPT can be used as a productivity tool for annotating sentiment lexicons, as far as it receives a final adjustment from a specialist.

This work has been supported by FCT - Fundação para a Ciência e Tecnologia within the R&D Units Project Scope: UIDB/00319/2020.

References

1. Krista Casler, Lydia Bickel, and Elizabeth Hackett. 2013. Separate but equal? a comparison of participants and data gathered via amazon's mturk, social media, and face-to-face behavioral testing. Comput- ers in human behavior, 29(6):2156–2160.
2. Yashar Deldjoo. 2023. Fairness of chatgpt and the role of explainable-guided prompts. arXiv preprint arXiv:2307.11761.
3. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understand- ing. arXiv preprint arXiv:1810.04805.
4. Tânia Fontes, Francisco Murçós, Eduardo Carneiro, Joel Ribeiro, and Rosaldo JF Rossetti. 2023. Leveraging social media as a source of mobility intelligence: An nlp-based approach. IEEE Open Journal of Intelli- gent Transportation Systems.
5. Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. Proceedings of the National Academy of Sciences, 120(30):e2305016120.
6. Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. arXiv preprint arXiv:2302.07736.
7. Clayton Hutto and Eric Gilbert. 2014. Vader: A parsi- monious rule-based model for sentiment analysis of social media text. In Proceedings of the international AAAI conference on web and social media, volume 8, pages 216–225.
8. Dan Jurafsky and James H. Martin. 2023. Speech and Language Processing, third edition. draft (https://web.stanford.edu/˜jurafsky/slp3/).
9. G. Kunapuli. 2023. Ensemble Methods for Machine Learning. Manning.
10. Loïc Lannelongue, Jason Grealey, and Michael In- ouye. 2021. Green algorithms: quantifying the carbon footprint of computation. Advanced science, 8(12):2100707.
11. Robert Lew. 2023. Chatgpt as a cobuild lexicographer. Humanities and Social Sciences Communications, 10(1).
12. Pengfei Li, Jianyi Yang, Mohammad A Islam, and Shaolei Ren. 2023. Making ai less" thirsty": Uncovering and addressing the secret water footprint of ai models. arXiv preprint arXiv:2304.03271.
13. Jafar Majidpour and Kamal Al-Barznji. 2022. Opinions for receiving covid-19 vaccines based on sentiment analysis. Journal of Pharmaceutical Negative Re- sults, pages 1648–1659.
14. Francisco S. Marcondes, José João Almeida, and Paulo Novais. 2024. An exploratory design science research on troll factories. Integr. Comput.-Aided Eng.,31(1):95–115.
15. Nestor Maslej, Loredana Fattorini, Erik Brynjolfs son, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Helen Ngo, Juan Carlos Niebles,Vanessa Parli, et al. 2023. Artificial intelligence index report 2023. arXiv preprint arXiv:2310.03715.

16. Aaron Mok. 2023. Chatgpt could cost over $700,000 per day to operate. microsoft is reportedly trying to make it cheaper.https://www.businessinsider.com/how-much-chatgpt-costs-openai-to-run-estimate-report-2023- 4.
17. Bruno Moreschi, Gabriel Pereira, and Fabio G Cozman. 2020. The brazilian workers in amazon mechanical turk: dreams and realities of ghost workers. Contracampo, 39(1).
18. Anu J Nair, G Veena, and Aadithya Vinayak. 2021. Comparative study of twitter sentiment on covid-19 tweets. In 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), pages 1773–1778. IEEE.
19. Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? arXiv preprint arXiv:2302.06476
20. Zhuang Qiu, Xufeng Duan, and Zhenguang Cai. 2023. Does chatgpt resemble humans in processing implicatures? In Proceedings of the 4th Natural LogicMeets Machine Learning Workshop, pages 25–34.
21. Matthias Schonlau and Vera Toepoel. 2015. Straightlining in web survey panels over time. In Survey Research Methods, volume 9, pages 125–137.
22. Yves Van Vaerenbergh and Troy D Thomas. 2013. Response styles in survey research: A literature review of antecedents, consequences, and remedies. International journal of public opinion research, 25(2):195 -217.
23. Daniël Waterman. 2022. Predicting Twitter Sentimenton the Russo-Ukrainian was using Lexicon-based and Transformer Models. Ph.D. thesis, Tilburg U.
24. Tianyi Xie, Yaorong Ge, Qian Xu, and Shi Chen. 2023. Public awareness and sentiment analysis of covid-related discussions using bert-based infoveillance. AI, 4(1):333–347.
25. Lingyu Zhang, Ting Hua, Yilin Shen, and Hongxia Jin. 2023. Prompt tuning for zero-shot compositional learning. arXiv preprint arXiv:2312.02191.