Towards Interpreting Topic Models with ChatGPT

Emil Rijcken*§, Floortje Scheepers‡, Kalliopi Zervanou†,Marco Spruit †, Pablo Mosteiro§ Uzay Kaymak*
*Jheronimus Academy of Data Science, Eindhoven University of Technology, Eindhoven, The Netherlands
Email: e.f.g.rijcken@tue.nl, u.kaymak@ieee.org
‡Psychiatry, University Medical Centre Utrecht, Utrecht, The Netherlands
Email: f.e.scheepers-2@umcutrecht.nl
†Leiden Institute of Advanced Computer Science, Leiden University, Leiden, The Netherlands
Email: k.zervanou@liacs.leidenuniv.nl, m.r.spruit@liacs.leidenuniv.nl
§Department of Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands
Email: p.mosteiro@uu.nl

***Abstract*—Topic modeling has become a popular approach to identify semantic structures in text corpora. Despite its wide applications, interpreting the outputs of topic models remains challenging. This paper presents an initial study regarding a new approach to better understand this output, leveraging the large language model ChatGPT. Our approach is built on a three-stage process where we first use topic modeling to identify the main topics in the corpus. Then, we ask a domain expert to assign themes to these topics and prompt ChatGPT to generate human-readable summaries of the topics. Lastly, we compare the human- and machine-produced interpretations. The domain expert found half of ChatGPT's descriptions useful. This explorative work demonstrates ChatGPT's capability to describe topics accurately and provide useful insights if prompted accurately.**

***Index Terms*—Topic Modeling, LLM, ChatGPT, Electronic Health Records, Fuzzy Topic Models, Prompt Engineering**

I. Introduction

As the volume of textual data continues to grow, so does the significance of extracting valuable insights from large text corpora. Topic modeling is a technique for identifying latent semantic structures within vast volumes of text. Topic models are a group of unsupervised natural language processing algorithms returning a collection of topics from a corpus. Each topic is a set of words, aimed to be semantically related. Yet, interpreting a word set is not always straightforward. Often not all words in the word set seem semantically related, so the semantic theme of the word set is unclear. There is vast work on topic model evaluation [1]–[4], focusing on the quality of the produced topics, but little on topic model interpretation [5], [6], aimed at making sense of the produced topics.

The Natural Language Processing (NLP) field has recently seen numerous breakthroughs from Large Language Models (LLMs) such as BERT [7] and ChatGPT, which is based on the Generative Pre-training Transformer (GPT) model [8]. These models are based on deep learning [9] and Transformer models [10] and are trained on vast amounts of data. ChatGPT is the latest in a series of such LLMs and has caused excitement and controversy because it is one of the first models that can convincingly converse with its users in English and other languages on a wide range of topics. It learns autonomously from data, can write seemingly intelligent, human-like responses, and provide meaningful summaries of texts [11]. The rise of these models opens the doors to many new research applications, one of which is enhancing topic modeling.

However, to our knowledge, LLMs have not been used for interpreting topic models so far. In this work, we explore the potential of using ChatGPT, the state-of-the-art in LLMs, to interpret the output generated by topic models. We work with topics trained on Electronic Health Records (EHRs) and compare the interpretations from ChatGPT and a domain expert.

We follow a three-step approach. In the first step, we create topics generated by the popular probabilistic topic modeling algorithm Latent Dirichlet Allocation (LDA) [12]. Because of the nature of topic modeling, where words belong to different topics to some degree, it is intuitive to use a fuzzy topic model instead of a probabilistic model. For this reason, we compare LDA with a fuzzy topic modeling algorithm, Fuzzy Latent Semantic Analysis - Words (FLSA-W) [13]. In the second step, we prompt these topics to ChatGPT, to distil the essence of these topics into concise, easily digestible summaries. Simultaneously, we ask a similar question to a domain expert for the same topics. Then, in step three, we compare the generated summaries with those produced by a domain expert. In doing so, we aim to highlight the potential of LLMs in streamlining the interpretation process. By making topic modeling interpretation more intuitive, we aim to make topic modeling more accessible to a broader audience.

The outline of this paper is as follows. In Section II, we discuss topic modeling, focusing on various algorithms and the challenges of interpreting topics. In Section III, we discuss LLMs, focussing on their main components and capabilities in more detail. Section IV discusses the three-step approach and the data used for our experiments and evaluation. Then, we present the results and discuss the implications and limitations in Section V-A. Finally, Section VI concludes the paper.

II. Topic Modeling

Topic modeling is an unsupervised NLP task aiming to discover the underlying topics in a corpus of text and represent them as propensity distributions over a vocabulary. Most topic models return two matrices $\mathbf{P}(\mathbf{W}|\mathbf{T})$ & $\mathbf{P}(\mathbf{T}|\mathbf{D})$, the propensity of a word given a topic and the propensity of a topic

given a document, respectively. Then, the top-N words with the highest propensity per topic are typically taken to represent a topic. The most popular topic modeling algorithm is LDA [12]. It assumes that each document in a corpus is a mixture of topics and is a probability distribution over words. LDA iteratively assigns words to topics and topics to documents, optimizing the likelihood of the observed data. FLSA-W [13] is a fuzzy-clustering-based topic modeling algorithm. It starts by creating local- and global-term weights and then uses C-means clustering [14] on the V-matrix created by singular value decomposition to get the partition matrix. Lastly, the output matrices are obtained by matrix multiplications based on Bayes' Theorem. FLSA-W has been shown to outperform other algorithms in various evaluation metrics [15].

For evaluation, a coherence score is calculated that indicates how well the within-topic words semantically support each other. Various metrics exist to measure the coherence score [16], which can be based on counting co-occurrences [3] or by measuring the cosine similarity between words represented in a high-dimensional continuous space [17]. Alternatively, the diversity score reflects the uniqueness of the top-N words between different topics [4]. It is calculated by dividing the number of distinct words in the set of topics by the total number of topic words.

The product of the coherence- and the diversity score has been proposed as the interpretability score [4]. However, it is questionable whether this score actually reflects a topic's interpretability [18]. Topic interpretation can be challenging due to the lack of coherence amongst topic words [5]. Also, experiments have shown that the representation of topics as a bag of words might cause users to overlook important words, read too much into words, assume adjacent words go together, or incorrectly resolve polysemy [5]. For this reason, there is a need for adequate interpretation techniques.

III. Large Language Models

LLMs have recently emerged as a powerful tool in natural language processing. These models are based on deep learning techniques and Transformer models, such as BERT [7] GPT-3 [8], and ChatGPT. They have been used in various applications, including language translation, text summarization, and question-answering systems [7], [8], [19] and have shown remarkable performance in generating human-like responses.

Most LLMs use the Transformer framework for their architectures, which are deep-learning neural networks with Transformer blocks. These blocks commonly incorporate a multi-head self-attention module complemented by a fully connected feed-forward layer [10]. The distinctive feature of the multi-head self-attention mechanism is its ability to concentrate simultaneously on various segments of the input sequence, thereby discerning relationships among words. Furthermore, the fully connected feed-forward layer operates on each element of the input sequence in a parallel manner, enhancing the representations produced by the multi-head self-attention module [20]. The Transformer's success is partly due to its ability to capture long-range dependencies in the input sequence and to its ability to focus on the most relevant parts of the sequence when generating its output [10].

The Generative Pre-trained Transformers (GPT) models are a sequence of LMs that have each produced state-of-the-art results on various tasks. The first version, GPT, is a *left-to-right* model, meaning it is trained unidirectional, based on a deep Transformer decoder [21]. This model follows a *pre-train and fine-tune* paradigm [22]. During pre-training, an extensive collection of textual data is used to learn general language patterns and representations. Then, the LLMs are fine-tuned on various discriminative language understanding tasks.

However, there are several limitations to this fine-tuning approach. Evidence suggests that models overfit to training distributions and do not generalize well outside of it [23]–[25] and models are good on datasets, not so good at underlying tasks [26], [27]. Moreover, each task requires a large specific dataset, and one might end up with many versions of a similar model. To overcome this limitation, the GPT-2 model is pre-trained using a language modeling objective, but it performs no fine-tuning [28]. Instead, it uses textual prompts to perform zero-shot inference on various tasks. Ever since the switch from the *pre-train and fine-tune* paradigm to the *pre-train, prompt, and predict* paradigm [22], there is a growing literature on prompt engineering [29], [30]. After the GPT-2 proposal, evidence suggested that LM performance strongly depends on the scale and only weakly on the model shape and that LMs are more sample efficient [31]. For this reason, GPT-3 was trained at a much larger size than GPT-2, going from approximately 1.5 billion to 175 billion parameters [8]. Whereas GPT-2 was a zero-shot learner, GPT-3 performs best as a few-shot learner [8].

ChatGPT is GPT-3's successor and was specifically designed for conversation modeling. It uses 100 times fewer parameters than GPT-3, but its output is preferred over GPT-3's prompts.

This effectiveness seems to come from using Reinforcement Learning from Human Feedback (RLHF) [32]. The training procedure with RLHF consists of three steps.

1) Collect demonstration data and train a supervised policy.
2) Collect comparison data and train a reward model.
3) Optimize a policy against the reward model using reinforcement learning.

In the first step, a prompt is sampled from a prompt dataset. Then, a human labeller demonstrates the desired output behaviour, which is used to fine-tune the policy with supervised learning. In the second step, the prompt and several model outputs are sampled. Then, a labeller ranks the outputs from best to worst, and this data is used to train a reward model. In step three, a new prompt is sampled from the dataset. The policy from step one generates an output, and the reward model from step two calculates a reward. The reward is used to update the policy using proximal policy optimization [33]. This extra element changes the language model's objective from *predicting the next token on a webpage from the internet*, to *following the user's instructions helpfully and safely* [8], [28], [32], [34], [35].

IV. Data and Methodology

A. Data

The dataset used for training the topic models consists of Dutch clinical notes from the Psychiatry department of the UMCU. It comprises 4280 clinical notes, anonymized with DEDUCE [36] with an average length of 1481 words per document as discussed in [37], [38]. The original dataset contains labels for each document, as it was used for a classification task. However, because topic modeling is unsupervised, we only use the written notes and not the labels.

B. Methodology

This section describes our methodology to explore the potential of using ChatGPT to interpret topics generated by LDA and FLSA-W. We follow a three-step approach, as discussed below.

1) Step 1 - Generating Topics: We generate topics using the topic modeling algorithms, LDA [12] and FLSA-W [13] on a corpus of clinical notes from EHRs (Section IV-A). We train both algorithms so that they produce ten topics with twenty words per topic and use the Gensim [39] and FuzzyTM package [15] packages to train LDA and FLSA-W, respectively. FLSA-W is a fuzzy topic modeling algorithm that considers the degree of membership of a word to a topic. LDA is a widely used probabilistic topic modeling algorithm that infers the underlying topic structure of a corpus based on the distribution of words.

2) Step 2 - Interpreting Topics with a Domain Expert and ChatGPT: In this step, we ask a human domain expert and ChatGPT to interpret the topics generated in the first step. The domain expert is a psychiatrist working at a University Medical Center in the Netherlands. We asked the domain expert to assign a label to a topic in case she recognized one. To get useful outputs from ChatGPT, we had to be specific in our prompts. Initially, the prompt *"Which common denominator does the following set of words have <TOPICS>"*[1] led to meaningless topics, as it described the words to come from a medical setting in many cases. Although true, our goal was to seek more specific descriptions. Hence, we refined our prompts as follows: *"Given the fact that the following words all originate from the electronic patient record of a psychiatric department, what common denominator do the following words have? Please be as specific as possible. <TOPIC>"*. The first part of the prompt aims to prevent the output from providing known information. The last part was based on an iterative round in which we prompted ChatGPT to be more specific after providing the output. In addition to a single word, ChatGPT also provided an elaborate motivation for its answer.

[1]We used Dutch equivalents to the examples given.

3) Step 3 - Comparing Summaries Generated by ChatGPT and the Domain Expert: We compare the summaries generated by ChatGPT and the domain expert qualitatively. For each topic, we show the domain expert her own answer and the output provided by ChatGPT supplemented with the motivation. Then, we asked the domain expert the following questions.

1) Do you agree with ChatGPT's output? Answers:
   a) yes,
   b) no.
2) Which description is better? Answers:
   a) domain expert,
   b) ChatGPT,
   c) approximately equally good,
   d) not applicable.
3) How useful is ChatGPT's answer? Answer:
   a) not useful,
   b) not really useful,
   c) moderately useful,
   d) useful.

V. Results and Discussion

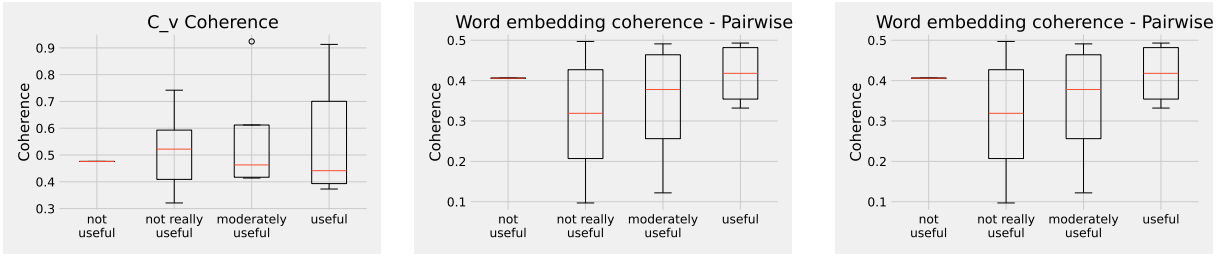
A. Usefulness of ChatGPT topics

Figure 2 shows the distribution of ratings assigned to ChatGPT's topic description, by the domain expert. 50 percent of the ratings are useful, and 50 percent are unuseful. Moreover, we find that:

1) The domain expert assigned a label to 75% of all topics.
2) In 20% of the unassigned topics, the domain expert found ChatGPT's description to be useful.
3) In 55% of all the topics, the domain expert did not agree with ChatGPT's topic description.
4) For 35%, of the topics, the domain expert considered her own description to be better than ChatGPT's description, while she found 30% of the topics to be equally good, and in 5% of the cases, ChatGPT's description was considered better.

B. Association with coherence

The ratings indicate that 50% of the descriptions provided by ChatGPT are perceived as useful by the domain expert. Aiming to understand the variability in the descriptions' usefulness, we assess whether there is a correlation between the usefulness and the coherence score. Figure 1 shows the distribution of coherence scores per topic over the ratings. The three coherence scores are c_v-coherence [3], and two word embedding-based approaches [17]. The c_v-coherence does not seem to correlate with the given ratings. However, the two word-embedding-based scores indicate a positive correlation between the coherence and rating scores. The average coherence score is higher for higher ratings, and the coherence variation is lower for useful topics. A contradiction between the c_v-coherence and the word embedding coherence was also found in [16]. This analysis indicates that topics with higher word-embedding-based coherence scores are more likely to be considered useful by the domain expert.



(a) (b) (c)

Fig. 1: Three boxplots showing the distributions for different rating categories and coherence metrics: (a) c_v-coherence (b) pairwise word embedding topic coherence, (c) centroid-based word embedding topic coherence.

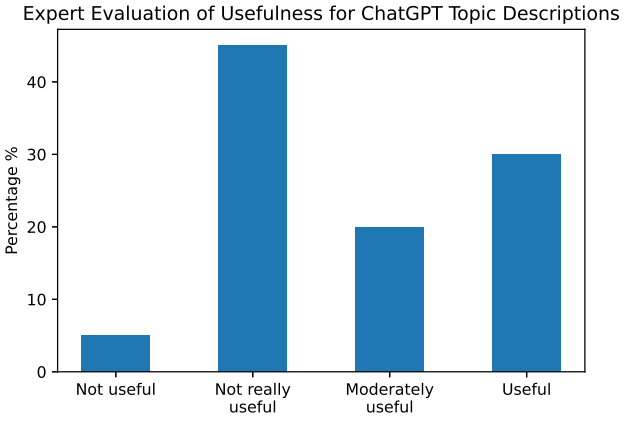


Fig. 2: A bar chart showing the distribution of domain expert ratings about the ChatGPT generated topic evaluation.

.

C. Topic analysis

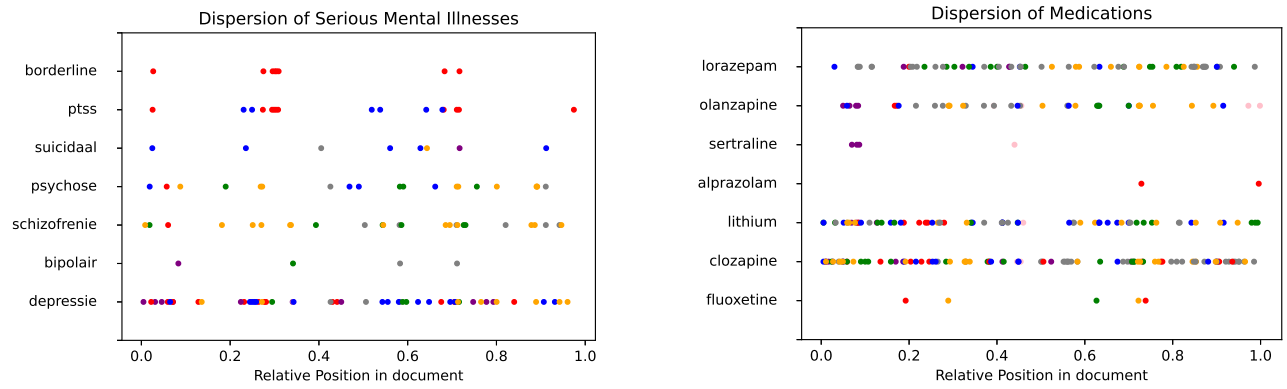
To better understand the variability in the descriptions' usefulness, we analyze a few topics in more depth in this section, where we consider one topic per description. Table I shows a topic per rating and the descriptions given by the domain expert and ChatGPT. We will use these descriptions and topics for our analysis in this section.

1) Useful: Here we discuss the topic from Table I about *Psychoses*. ChatGPT's motivation for this description is: *Words such as "person", "mother", "parents", "patient", "good", "father", "year", "admission", "issue" suggest that the words are related to a situation of treatment of psychological problems in a young person in the context of their family. Words such as "voices", "complaints", "sometimes", "hallucinations", "psychosis", "diagnostics", "examination", "anamnesis", "thoughts", "tells", "where" suggest that the words in related to an analysis of the symptoms and causes of psychosis, taking into account the role of parents and family in the history and diagnosis*. In this example, the domain expert and ChatGPT agree that the topic is about psychoses. However, ChatGPT refines the domain expert's description by specifying this topic is about psychoses with young people, and it explains why this seems to be the case.

2) Moderately useful: Amongst the topics considered to be moderately useful, the domain expert agreed with ChatGPT's description in 75% of the cases. We analyze the topic about *Anamnesis* from Table I here. ChatGPT's motivation for giving this description is: *This may include conducting research, collecting history, obtaining information from referrers and obtaining additional information about somatic problems. It is also stated whether the patient was previously known to the institution and whether there are any affective traits and intoxications. The patient's psychomotor skills, intelligence and legal orientation are also examined*. In this topic, the domain expert and ChatGPT seem to agree with each other. However, ChatGPT does not seem to add extra relevant information on top of the information provided by the domain expert.

3) Not really useful: The domain expert disagreed with all of the topics considered as *Not really useful*. We analyze the topic from Table I about *social skills* here. ChatGPT's motivation for its description is: *Words such as "person", "good", "problems", "risk", "contact", "adl", "function", "goes", "aggression", "psychic", "functions", "social" and "skills" suggest that the words are related to a situation of treating a person's psychological problems and/or aggression. Words such as "occupational", "conversation", "gives", "vpk", "ward", "going" and "attending" suggest that the treatment takes place in a professional setting, such as a ward of a psychiatric institution*. Here, the focus as assigned by ChatGPT and the domain expert varies greatly, and the example given by the domain expert is more specific than ChatGPT's answer. Because ChatGPT has been prompted with the information that the words originate from the electronic health records of a psychiatric department, it seems unable to interpret the topics beyond this information.

4) Not useful: In one topic, the domain expert disagrees with ChatGPT and finds the output not useful (the topic about *serios mental illnesses in Table I*). The motivation



(a) (b)

Fig. 3: Dispersion plots of serious mental illnesses (a) and medications (b) across 11 documents, each represented by a different color. The plots show the frequency of mentions for each topic at their relative position in the document.

provided by ChatGPT for this topic was: *Words such as "mg", "medication", "symptoms", "effect", "clozapine", and "lithium" suggest that the words are related to the medication treatment of mental illness. This indicates that there is psychopharmacology.* To better understand this topic, we analyze the associated documents. We select all 11 documents from the corpus containing all 20 topic words. Then, we prompt ChatGPT to provide typical serious mental illnesses or psychiatric medications. We look at how often both sets of words appear in these 11 documents. Figure 3 shows the relative positions of words in the documents, where each colour represents a document. The high prevalence of words from both groups (*Serious Mental Illnesses* and *psychopharmacology*) indicates both descriptions could be considered relevant descriptions. Yet, ChatGPT focuses on different aspects of the descriptions. The medications are often prescribed to patients with serious mental illnesses. Hence, the difference in descriptions makes sense, and both descriptions seem accurate, although one is considered less useful by the domain expert.

D. Comparison of LDA and FLSA-W

The domain expert assigned a label to all LDA topics, but could not assign a label to three topics produced by FLSA-W. This might indicate that the topics produced by LDA are better interpretable for the domain expert. However, the different coherence scores indicate the topics produced by both models are approximately equally coherent (Table II).

E. Limitations

The results in terms of the percentage of agreement between the domain expert and ChatGPT do not portray the machine in a positive light. We deem this agreement due to several factors.

- We rely on the evaluation of one domain expert only. This precludes the assessment of potential sources of error, such as biased judgments, inaccuracies in the data, inter-annotator agreement, and other significant metrics that are instrumental in gauging the quality of annotation.
- The questions asked in our work might introduce an extra bias knowing that the texts came from ChatGPT.

Moreover, our experiments are conducted on a small scale, with a small number of topics, and we have only evaluated a few topics. For this reason, these results are preliminary. Lastly, ChatGPT and other LLMs are large and non-deterministic algorithms. Although there is growing literature about LLM explainbility [40]–[42], there is no guarantee that the produced output is truthful. For this reason, the topic explanations provided by ChatGPT should be used to support topic modeling interpretation, but should not be seen as an objectively truthful answer.

VI. Conclusion

Summarizing, this paper explores a novel approach for interpreting topic modeling output by leveraging the large language model ChatGPT. The approach involves a three-stage process where topics are identified using topic modeling, themes are assigned by a domain expert and ChatGPT, and the domain expert compares the descriptions. Although we find discrepancies between human- and machine descriptions, the results indicate that ChatGPT might be used to describe topics and provide useful insights. Yet, the experiments are conducted on a small scale with a small number of topics and one domain expert. Furthermore, we have only analyzed a small number of topics in more depth. For this reason, the results are preliminary. Also, ChatGPT is not deterministic, so the results' reproducibility could be poor. Yet, the descriptions provided by ChatGPT are useful. For this reason, we believe ChatGPT is a promising model to support topic modeling interpretation. This work aims to inspire peers to study ChatGPT further for topic interpretations. In future work, we intend to study an iterative customization process per prompt to obtain more specific topic descriptions. Moreover, we plan to include more domain experts, change the questions to reduce bias and conduct our experiments on a larger scale.

Rating	Domain Expert	ChatGPT	Topic
Useful	Psychoses	Treatment of psychosis in young people, including the role of parents and family in the history and diagnosis.	*'person', 'mother', 'parents', 'patient', 'good', 'father', 'year', 'admission', 'issue', 'voices', 'complaints', 'sometimes', 'hallucinations', 'psychosis', 'diagnostics', 'examination', 'anamnesis', 'thoughts', 'tells', 'whereby'*
Moderately useful	Anamnesis	Examination and admission of a patient, taking into account anamnesis, referrer, additional somatic and psychological factors, and legal orientation.	*'examination', 'recording', 'anamnesis', 'referrer', 'additional', 'somatic', 'known', 'affect', 'year-old', 'preserved', 'means', 'recorded', 'intact ', 'attract', 'intoxications', 'consciousness', 'psychomotor', 'intelligence', 'legal', 'orientation'*
Not really useful	Social skills	Treatment of psychological problems and aggression in a person and their functioning in daily activities.	*'person', 'good', 'problems', 'risk', 'contact', 'general activities of daily living', 'functioning', 'goes', 'aggression', 'psychic', 'functions', 'social', 'skills ', 'occupational', 'conversation', 'gives', 'nurse', 'department', 'going', 'present'*
Not useful	Serious mental illnesses	Psychopharmacology	*'mg', 'admission', 'dp', 'organization', 'medication', 'person', 'treatment', 'complaints', 'ect', 'patient', 'day', 'clozapine', 'effect', 'lithium', 'good', 'disorder', 'psychotic', 'depression', 'start', 'burden'*

The original word was a Dutch abbreviation. We use multiple words here to maintain the original meaning.

TABLE I: Topic information per rating. Each row shows an example of a topic and its descriptions for each rating. Note that ChatGPT's description and the topics are both translated from Dutch.

	$\mathbf{C_v}$	$\mathbf{WEC_{PW}}$	$\mathbf{WEC_{CENT}}$
FLSA-W	0.517	0.349	0.745
LDA	0.553	0.349	0.655

TABLE II: The average coherence scores per model coherence metric.

ACKNOWLEDGMENT

The authors acknowledge the COVIDA funding provided by the strategic alliance of TU/e, WUR, UU, and UMC Utrecht. They have no competing interests to declare relevant to this article's content.

REFERENCES

[1] J. Chang, S. Gerrish, C. Wang, J. Boyd-Graber, and D. Blei, "Reading tea leaves: How humans interpret topic models," *Advances in neural information processing systems*, vol. 22, 2009.

[2] J. H. Lau, D. Newman, and T. Baldwin, "Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Gothenburg, Sweden: Association for Computational Linguistics, Apr. 2014, pp. 530–539. [Online]. Available: https://aclanthology.org/E14-1056

[3] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proceedings of the eighth ACM International Conference on Web Search and Data Mining*, 2015, pp. 399–408.

[4] A. B. Dieng, F. J. R. Ruiz, and D. M. Blei, "Topic modeling in embedding spaces," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 439–453, 2020. [Online]. Available: https://aclanthology.org/2020.tacl-1.29

[5] T. Y. Lee, A. Smith, K. Seppi, N. Elmqvist, J. Boyd-Graber, and L. Findlater, "The human touch: How non-expert users perceive, interpret, and fix topic models," *International Journal of Human-Computer Studies*, vol. 105, pp. 28–42, 2017.

[6] A. Marchetti and P. Puranam, "Interpreting topic models using prototypical text: From 'telling'to'showing'," *Social Science Reseach Network*, 2020.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[8] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[9] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[11] E. A. van Dis, J. Bollen, W. Zuidema, R. van Rooij, and C. L. Bockting, "Chatgpt: five priorities for research," *Nature*, vol. 614, no. 7947, pp. 224–226, 2023.

[12] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *The Journal of Machine Learning research*, vol. 3, pp. 993–1022, 2003.

[13] E. Rijcken, F. Scheepers, P. Mosteiro, K. Zervanou, M. Spruit, and U. Kaymak, "A comparative study of fuzzy topic models and LDA in terms of interpretability," in *Proceedings of the 2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2021.

[14] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. Springer Science & Business Media, 2013.

[15] E. Rijcken, P. Mosteiro, K. Zervanou, M. Spruit, F. Scheepers, and U. Kaymak, "FuzzyTM: a software package for fuzzy topic modeling," in *2022 IEEE international conference on fuzzy systems (FUZZ-IEEE)*, 2022.

[16] E. Rijcken, K. Zervanou, M. Spruit, P. Mosteiro, F. Scheepers, and U. Kaymak, "Exploring embedding spaces for more coherent topic modeling in electronic health records," in *IEEE International Conference on Systems, Man, and Cybernetics*, 2022.

[17] R. Ding, R. Nallapati, and B. Xiang, "Coherence-aware neural topic modeling," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 830–836. [Online]. Available: https://aclanthology.org/D18-1096

[18] E. Rijcken, U. Kaymak, F. Scheepers, P. Mosteiro, K. Zervanou, and M. Spruit, "Topic modeling for interpretable text classification from EHRs," *Frontiers in Big Data*, vol. 5, 2022. [Online]. Available: https://www.frontiersin.org/article/10.3389/fdata.2022.846930

[19] S. Liu, X. Zhang, S. Zhang, H. Wang, and W. Zhang, "Neural machine reading comprehension: Methods and trends," *Applied Sciences*, vol. 9, no. 18, p. 3698, 2019.

[20] T. Lin, Y. Wang, X. Liu, and X. Qiu, "A survey of transformers," *AI Open*, 2022.

[21] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training," *OpenAI blog*, 2018.

[22] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–35, 2023.

[23] D. Yogatama, C. d. M. d'Autume, J. Connor, T. Kocisky, M. Chrzanowski, L. Kong, A. Lazaridou, W. Ling, L. Yu, C. Dyer *et al.*,

"Learning and evaluating general linguistic intelligence," *arXiv preprint arXiv:1901.11373*, 2019.
[24] R. T. McCoy, E. Pavlick, and T. Linzen, "Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference," *arXiv preprint arXiv:1902.01007*, 2019.
[25] D. Hendrycks, X. Liu, E. Wallace, A. Dziedzic, R. Krishnan, and D. Song, "Pretrained transformers improve out-of-distribution robustness," *arXiv preprint arXiv:2004.06100*, 2020.
[26] S. Gururangan, S. Swayamdipta, O. Levy, R. Schwartz, S. R. Bowman, and N. A. Smith, "Annotation artifacts in natural language inference data," *arXiv preprint arXiv:1803.02324*, 2018.
[27] T. Niven and H.-Y. Kao, "Probing neural network comprehension of natural language arguments," *arXiv preprint arXiv:1907.07355*, 2019.
[28] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.
[29] S. Arora, A. Narayan, M. F. Chen, L. J. Orr, N. Guha, K. Bhatia, I. Chami, F. Sala, and C. Ré, "Ask me anything: A simple strategy for prompting language models," *arXiv preprint arXiv:2210.02441*, 2022.
[30] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, and D. C. Schmidt, "A prompt pattern catalog to enhance prompt engineering with chatgpt," *arXiv preprint arXiv:2302.11382*, 2023.
[31] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*, 2020.
[32] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *arXiv preprint arXiv:2203.02155*, 2022.
[33] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.
[34] R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du *et al.*, "Lamda: Language models for dialog applications," *arXiv preprint arXiv:2201.08239*, 2022.
[35] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," *J. Mach. Learn. Res*, vol. 23, pp. 1–40, 2021.
[36] V. Menger, F. Scheepers, L. van Wijk, and M. Spruit, "DEDUCE: A pattern matching method for automatic de-identification of dutch medical text," *Telematics and Informatics*, vol. 35, no. 4, pp. 727–736, 2018.
[37] P. Mosteiro, E. Rijcken, K. Zervanou, U. Kaymak, F. Scheepers, and M. Spruit, "Making sense of violence risk predictions using clinical notes," in *Proceedings of the International Conference on Health Information Science*. Springer, 2020, pp. 3–14.
[38] ——, "Machine learning for violence risk assessment using Dutch clinical notes," *Journal of Artificial Intelligence for Medical Sciences*, vol. 2, no. 1-2, pp. 44–54, 2021.
[39] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, Valletta, Malta, May 2010, software available from http://radimrehurek.com/gensim. [Online]. Available: http://is.muni.cz/publication/884893/en
[40] S. Stevens and Y. Su, "An investigation of language model interpretability via sentence editing," in *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 435–446. [Online]. Available: https://aclanthology.org/2021.blackboxnlp-1.34
[41] M. Szczepański, M. Pawlicki, R. Kozik, and M. Choraś, "New explainability method for bert-based model in fake news detection," *Scientific reports*, vol. 11, no. 1, p. 23705, 2021.
[42] V. W. Anelli, G. M. Biancofiore, A. De Bellis, T. Di Noia, and E. Di Sciascio, "Interpretability of bert latent space through knowledge graphs," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022, pp. 3806–3810.