ИНФОРМАТИКА И КОМПЮТЪРНИ НАУКИ
INFORMATICS AND COMPUTER SCIENCES

COMPARATIVE ANALYSIS OF TOPIC MODELING AND LARGE LANGUAGE MODELS IN EXTRACTING INSIGHTS FROM SOCIAL MEDIA CONTENT

Vitali Chaiko
University of Library Studies and Information Technologies

Abstract: *The exponential growth of textual data, driven by communication technologies, presents a challenge in extracting valuable insights. This study focuses on evaluating the effectiveness of topic modeling algorithms BERTopic and Top2Vec compared to Large Language Models (LLMs), especially ChatGPT and Google Bard in the context of social media analysis. Specifically, it investigates the capability of these models in extracting topics from a corpus of Twitter data regarding the company Amazon. The methodology involves preprocessing a subset of more than 720,000 tweets, followed by topic model training and prompt engineering for LLMs. The study develops quantitative metrics for comparison of the topic extraction capabilities of the models. Initial results indicate a disparity in the performance of topic models and LLMs, with LLMs demonstrating human intuitive topic extraction, but exhibiting only 15% of similarity in exact topic words and 23% of similarity in word embeddings compared to topic models.*

Keywords: *topic modelling, prompt engineering, ChatGPT, Bard, twitter*

INTRODUCTION

The task of navigating and extracting knowledge from the vast textual data in today's world is complex. As communication technology and hardware continue to advance, the volume of text data increases daily. Rapid comprehension of this data can yield significant economic benefits and assist in strategic decision-making. One of the possible tools for automatically extracting insights from textual data is topic modeling. This machine learning algorithm organizes sentences or documents with similar thematic content, labeling them with topic words, such as grouping sentences related to the stock market. Models such as BERTopic (Grootendorst 2022, p. 1) are effective; however, their implementation demands considerable time and a meticulously prepared data set for reliable results. Additionally, quantitatively evaluating topic modeling can be challenging due to its inherent subjectivity. Lately, large language models (LLMs) like ChatGPT and Bard have evolved into versatile tools for various tasks. This work aims to create quantitative metrics to compare the effectiveness of these LLMs in topic extraction tasks against topic modeling algorithms on social media content.

Topic modelling

Topic modeling (Blei 2012, p. 1; Abdelrazek et al. 2022, pp. 1–2) is a machine learning algorithm that discovers abstract themes or topics in text data. It is especially useful for finding hidden semantic structures, and organizing large volumes of text data. The challenge of topic modeling is to determine the optimal number of topics to extract from the data, which can be addressed using techniques such as coherence score evaluation or topic visualization. The resulting topics are typically represented with a set of topic words that are extracted from the text document of a certain topic and semantically best describe that topic. Topic modeling has been successfully applied in various fields such as natural language processing, social media analysis, and economics (Abdelrazek et al. 2022, p. 13). This work focuses on recent topic modelling algorithms, BERTopic and Top2vec (Angelov 2020, pp. 1–2), which are based on word and document embeddings (Mikolov et al. 2013, p. 1) (Le & Mikolov 2014, p. 1). Such embeddings represent words and documents in a high-dimensional numerical space and are able to capture semantic properties.

Large language models

Large Language Models are self-supervised machine learning algorithms that are trained on a very large amount of textual data (Bender et al., 2021, pp. 2–6): LLMs are trained on hundreds of billions of text tokens by predicting the next token given all the preceding tokens of a text. They are capable of understanding and generating human-like text and can be used for a broad spectrum of linguistic tasks. LLMs are not only able to generate text but also show understanding through tasks like translation, summarization, and answering questions. ChatGPT (Bahrini et al., 2023, pp. 1–6; OpenAI 2023, p. 1) and Google Bard (Google AI 2023) are prominent examples of LLMs. These models are built upon generative pre-trained transformer (GPT) architectures, which were trained on a vast corpus of textual data gathered from all over the internet. Users can communicate with the models in a browser chat application via messages, known as prompts, and utilize them in various fields. There is also an application programming interface (API) available that allows developers to access the model's capabilities and integrate them into their own applications.

Related work

Topic modeling has been successfully applied to Twitter data in the context of climate change (Uthirapathy & Sandanam 2023, pp. 1–2). In this study, the publicly available climate change data from Twitter was processed using the Latent Dirichlet Allocation (LDA) topic modeling method (Ramage et al. 2009, p. 1) and then BERT uncased model (Devlin et al. 2018, p. 1) has been applied to classify sentiment of the data. The models demonstrated a precision of 91.35%, a recall of 89.65%, and an F1-Score of 93.50%. Moreover, the information extraction capabilities of ChatGPT have been evaluated on 14 different textual datasets and compared with existing baseline models such as BERT and RoBERTA (Li et al. 2023, pp. 1–3). ChatGPT demonstrated weak performance in supervised tasks (where labels are pre-defined), but excelled in unsupervised tasks, including those similar to topic modeling, suggesting its potential in such applications. The combination of established topic modeling machine learning algorithms and LLMs has been explored using a psychiatric clinical notes dataset (Rijcken et al. 2023, pp. 1–2). The resulting topic clouds from the LDA algorithm were used to generate human-readable summaries of the topics with ChatGPT. These summaries were found useful by human domain experts in half of the cases, indicating a beneficial blend of automated and human decisions. Explicit topic modeling using GPT2 and GPT3 has been explored with the DBPedia ontology classification dataset (Wang et al. 2023). The authors conclude that, given a few demonstrations as prompts for a GPT, the large language model can learn in context and transfer the knowledge to new data. Topic modelling with the GPT2 and GPT3 achieved an average accuracy of 70%.

RESEARCH METHODOLOGY

Dataset

The tweet dataset used for this study is a subset of a collection of historical tweets about the top NASDAQ companies between years 2015 and 2020 (Doğan et al. 2020), made available on a public data science data & code platform, Kaggle (Kaggle 2023). The dataset consists of more than 3 million tweets. Each tweet record in the dataset holds attributes of the tweet's unique identifier, the tweet author, date of tweet creation, textual content contained in the tweet, corresponding NASDAQ company and social engagement indicators such as number of comments, likes, and retweets. For this work, the 720,000 tweets regarding the company Amazon were selected from the dataset, so that the application of topic modeling could be conducted within a compact, coherent domain. Due to the Twitter developer content distribution policy, the contents of the tweets cannot be displayed in this study.

Topic model training

The dataset is preprocessed by removing all stop words, hyperlinks and numbers. Then by using Top2Vec and BERTopic algorithms the topic models are trained with minimal vocabulary occurrence of 1, meaning that every word, even if it occurs once in the dataset is taken into account. All other parameters of the algorithms are set to default value. Moreover, during the training process custom word and document embeddings are trained for every word and tweet of the dataset.

Prompt engineering

The following prompt is given as input to LLMs, ChatGPT 4 and Google Bard, beforehand:

“Given tweets separated by a new line in the next input, for each tweet output topics this tweet belongs to. Use only single words as topics. Separate the topics with”.

Because the topic model training occurs with single words, only single words are required as topic outputs from an LLM. Then, a random subset of 100 tweets from the dataset is selected and used as a prompt, with each tweet separated by a new line. The selection of the subset is then repeated 10 times for a total of 1000 tweets. The example inputs and outputs of ChatGPT 4 and Bard, as of 20.11.2023, for the first three input tweets are shown below (the tweet contents are not of real tweets, but are semantically similar):

User input Prompt:

$AMZN will fall down 500 points in the next days
$AMZN is causing disturbances in transportation industry
#Amazon Leases More Planes For Air Cargo Network

LLM output:

Stocks;Trending;Finance
Amazon;Industry;Transportation;Disruption
Amazon;Aircraft;Logistics

RESULTS

First, two quantitative metrics, UCI Coherence Score and topic diversity, are applied to the resulting BERTopic and Top2vec topic models. UCI Coherence Score is a metric that measures the degree of semantic similarity between high scoring words within each discovered topic (Röder et al. 2015, p. 2). It is computed by multiple steps:

1. For each topic, the top *N* topic words are selected. Then pairs of words (segments) from these top words are created.

2. For each pair of words w_i and w_j of top N topic words, the following probabilities are calculated.

- *$P(w_i)$:* The probability of w_i appearing in the documents.
- *$P(w_j)$:* The probability of w_j appearing in the documents.
- *$P(w_i, w_j)$:* The joint probability of w_i and w_j appearing together in a sliding window of texts.

3. Coherence Score Calculation: For each pair of w_i and w_j, score is calculated using Normalized Pointwise Mutual Information (NPMI) in the equation 1, with prechosen constants ϵ and γ (here 1).

$$NPMI(w_i, w_j)^{\gamma} = \left(\frac{\log\frac{P(w_i,w_j)+\epsilon}{P(w_i)P(w_j)}}{-\log\left(P(w_i,w_j)+\epsilon\right)}\right)^{\gamma} \tag{1}$$

4. The coherence score for a topic is aggregated as the average of the NPMI scores for all the word pairs in that topic. The total UCI coherence for the model is calculated as mean of the coherence scores of all topics.

Topic diversity is a metric for evaluating the quality of topic models (Murakami & Chakraborty 2022 pp. 10–18): The topic diversity of a topic model is the proportion of unique words in the top-N words across all calculated topics, shown in the equation 2.

$$Topic\ diversity_N = \frac{(Number\ of\ Unique\ Top{-}N\ words)}{(Total\ Number\ of\ Top{-}N\ words)} \tag{2}$$

The resulting topic models are evaluated using the UCI coherence score and topic diversity metrics, as shown in **Error! Reference source not found.** The number of discovered topics is very high; the majority were discovered by the BERTopic model. The coherence scores and topic diversity scores of the both models are average, but sufficient for the large number of discovered topics. The BERTopic model performs better than the Top2Vec model.

Table 1. Evaluation of topic models

Model	**Top2vec**	**BERTopic**
Number of topics	7561	8567
CV Coherence Score	0.397	0.445
Topic diversity$_{10}$	0.435	0.542

To compare the topic extraction results of LLMs and topic models, multiple steps are undertaken. First, the extracted topics of LLMs for the tweets are manually observed to determine if any of the tweets contain semantically unintuitive topics. **All** topics extracted by ChatGPT and Bard are considered **intuitive**.

Subsequently, LLM topic extraction and the topic models are compared using multiple self-defined metrics. The first metric, Exact_words, as described in Metric 1, aims to compare the topic words found by the LLM with the exact topic words of the topic models. The metric, Word_embeddings, as described in Metric Metric 2, compares the word embeddings of the topic words obtained during the training process of the topic models. The last metric

Average_word_embeddings , described in Metric 3, compares the word embeddings, but not of each individual topic word, it instead considers the average word embedding of all the topic words within a topic. The output of all three metrics ranges from 0 to 1, with 1 representing the perfect value. The results of all three metrics, for different k values ranging from 1 to 10, are presented in Table 2.

The exact topic words from LLMs, when compared using *Exact_words*, are rarely included in the topic words of the topic models, with the highest value observed in the comparison between ChatGPT and the BERTopic model. The metric *Word_embeddings* indicates that, for the 10 nearest topics to the topic word identified by ChatGPT, the corresponding topic identified by BERTopic is present among these 10 topics in approximately 20% of the cases. The last metric, *Average_word_embeddings*, indicates that in 23% of the cases, the correct topic identified by ChatGPT and BERTopic is present among the 10 nearest topics. The BERTopic model outperforms the Top2Vec model in all of the cases, which corresponds to its higher coherence and topic diversity scores. Additionally, the topic extraction by ChatGPT outperforms that of Bard in all cases. However, the performance of all metrics is moderately low, and no strong connection between the metric performance and the UCI Coherence Score or topic diversity can be established.

Metric 1. Exact_words

1. Take the list of topic words *tw* of tweet, which were output for this tweet by the LLM.
2. For every word *tw*$_{\mathrm{i}}$ in *tw*, check if this word is among the 10 topic words of the topic identified for this tweet by the topic model. If so, label *tw*$_i$ as *topic_similar*$_i$, otherwise, label it as *topic_not_ similar*$_i$.
3. Calculate the *topic_ similar* $_{tw}$, percentage value of total words that were labeled *topic_similar*$_i$ in *tw*, by dividing the sum of number of *topic_similar*$_i$ labels in *tw* by the length of *tw*.
4. Calculate the average percentage value *topic_ similar* $_{LLM}$ of all tweet samples by dividing the sum of *topic_ similar* $_{tw}$ by the number of tweet samples.

Metric 2. Word_embeddings

1. Take the list of topic words *tw* of tweet, which were output for this tweet by the LLM.
2. For every word *tw*$_i$ in *tw* find the first *k* most similar topics to the *tw*$_{i,}$ ranked by using the cosine similarity between the word embedding of *tw*$_{i,}$ and the word embedding of every topic word of each topic of the model. Store the first *k* topics of the most similar word embeddings to *tw*$_i$ as list of topic identifiers *tid*$_i$.
3. Check if topic identifier, which was calculated for this tweet by the topic model is included, in *tid*$_i$. If so, label *tw*$_i$ as *topic_similar*$_i$, otherwise, label it as *topic_not_ similar*$_i$.
4. Calculate the *topic_ similar* $_{tw}$, percentage value of total words that were labeled *topic_similar*$_i$ in *tw*, by dividing the sum of number of *topic_similar*$_i$ labels in *tw* by the length of *tw*.
5. Calculate the average percentage value *topic_ similar* $_{LLM}$ of all tweet samples by dividing the sum of *topic_ similar* $_{tw}$ by the number of tweet samples.

Metric 3. Average_word_embeddings

1. Take the list of topic words *tw* of tweet, which were output for this tweet by the LLM.

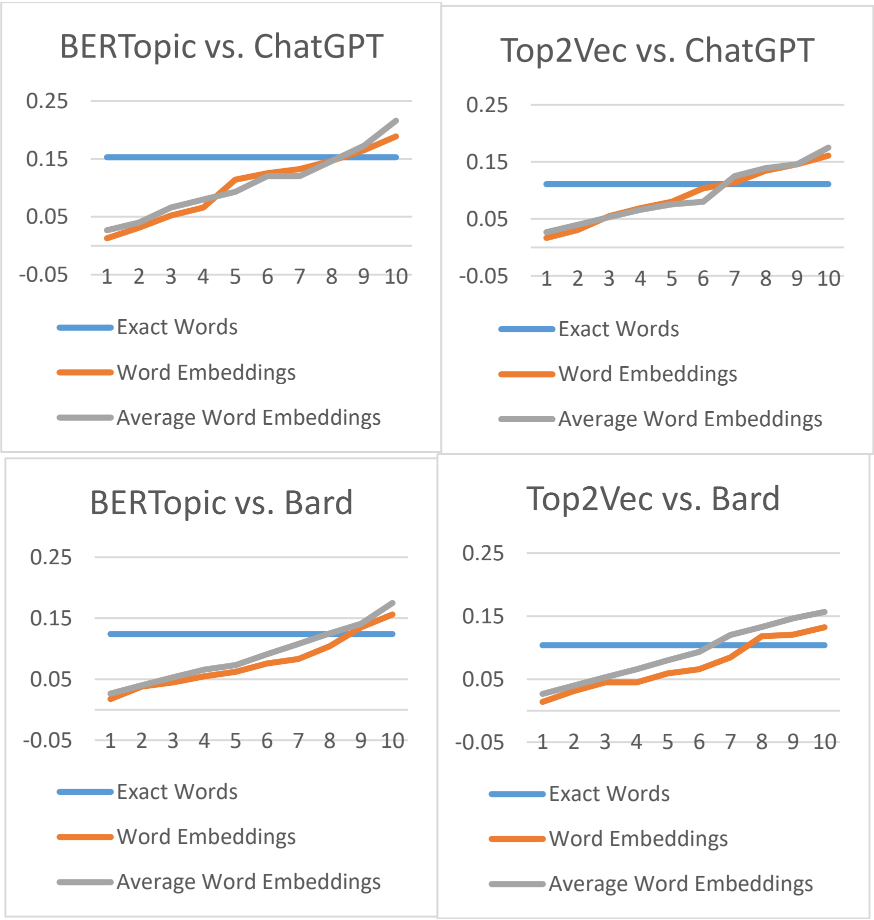
2. For every word tw_i in *tw* find the first *k* most similar topics to the tw_i, ranked by using the cosine similarity between word embedding of $tw_{i,}$ of the average word embedding of all topic words of each topic of the model. Store the first *k* topics of the most similar average word embeddings to tw_i as list of topic identifiers tid_i.

3. Check if topic identifier, which was calculated for this tweet in the topic model, is included in tid_i. If so, label tw_i as $topic\_similar_i$, otherwise, label it as $topic\_not\_ similar_i$.

4. Calculate the $topic\_ similar_{tw}$, percentage value of total words that were labeled $topic\_similar_i$ in tw, by dividing the sum of number of $topic\_similar_i$ labels in *tw* by the length of *tw*.

5. Calculate the average percentage value $topic\_ similar_{LLM}$ of all tweet samples by dividing the sum of $topic\_ similar_{tw}$ by the number of tweet samples.

Table 2. Comparison results



CONCLUSIONS/DISCUSSION

The results from self-defined metrics highlight that comparing topic extraction outcomes from LLMs with those from topic models on social media data is not straightforward, primarily due to the observed low performance of the metrics. No strong connection can be observed between the self-defined metrics and the UCI Coherence Score or topic diversity. However, from a human perspective, the topics identified by LLMs appear rational and intuitive. Nevertheless, the metrics provide a quick overview of the prevalence of topic-specific words in topic models and demonstrate a trend: higher metric performance is associated with greater topic coherence and diversity in the models.

To improve the effectiveness and flexibility of the proposed metrics, there are several strategies that can be considered. Firstly, it is important to experiment with a wider range of prompt inputs. While the focus of this work has been primarily on topic extraction, prompts that not only facilitate topic extraction but also enable topic categorization could be explored. Such changes in the prompts align with the objectives of topic models, which are also designed to categorize similar topics together to create an organized and meaningful analysis of the data.

Another strategy to explore is the use of pre-labeled topic models that are specific to social media contexts. These models, which have topics manually assigned, could provide a more structured framework for the analysis. However, such models are rare and highly specialized. Additionally, it could be beneficial to explore other topic modeling algorithms besides the recent document embedding-based techniques utilized in this work. Other topic modeling algorithms, like LDA, should not be overlooked.

It is important to point out that topic models are subjective models.

The quality and effectiveness of these models can vary significantly across different text domains, making it challenging to establish baselines or draw direct comparisons between various models. This variability suggests that training multiple models on the same social media datasets could yield different results with the proposed metrics. Furthermore, the evolving nature of LLMs plays a significant role in this context. The continuous advancements and changes in LLMs and their user interfaces can significantly influence how they process user prompts, which, in turn, impacts the performance of topic extraction. These dynamics of LLMs need to be considered when evaluating their effectiveness in topic extraction tasks, especially in comparison to topic models.

REFERENCES

Abdelrazek, A. et al. (2022). Topic modeling algorithms and applications: A survey. Information Systems, Issue 112.

Angelov, D. (2020). Top2Vec: Distributed Representations of Topics. arXiv, Issue arXiv:2008.09470.

Bahrini, A. et al. (2023). ChatGPT: Applications, Opportunities, and Threats. arXiv, Issue arXiv:2304.09103.

Bender, E. M., T. **Gebru,** A. **McMillan-Major** & S. **Shmitchell** (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big. Virtual Event Canada, Association for Computing Machinery, p. 610–623.

Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM. Communications of the ACM, Issue 55(4), pp. 77–84.

Devlin, J., M. **Chang,** K. **Lee,** & K. **Toutanova** (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Seattle, Association for Computational Linguistics, pp. 1–11.

Doğan, M. et al. (2020). Speculator and Influencer Evaluation in Stock Market by Using Social Media. Altanta, IEEE, pp. 4559–4566.

Gillis, N. (2014). The why and how of nonnegative matrix factorization. arXiv, Issue arXiv:1401.5226.

Google AI (2023). Google AI. [Online] Available at: https://ai.google/static/documents/google-about-bard.pdf[Accessed 01 12 2023].

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv, Issue arXiv:2203.05794.

Kaggle (2023). Tweets about the Top Companies from 2015 to 2020. [Online] Available at: https://www.kaggle.com/datasets/omermetinn/tweets-about-the-top-companies-from-2015-to-2020 [Accessed 20 01 2024].

Le, Q. & T. **Mikolov** (2014). Distributed Representations of Sentences and Documents. arXiv, Issue arXiv:1405.4053.

Li, B. et al. (2023). Evaluating ChatGPT's Information Extraction Capabilities: An Assessment of Performance, Explainability, Calibration, and Faithfulness. ArXiv, Issue arXiv:2304.11633.

Mikolov, T., K. **Chen,** G. **Corrado** & J. **Dean** (2013). Efficient Estimation of Word Representations in Vector Space. arXiv, Issue arXiv:1301.3781.

Murakami, R. & B. **Chakraborty** (2022). Investigating the Efficient Use of Word Embedding with Neural-Topic Models for Interpretable Topics from Short Texts. Sensors, Issue 22(3). OpenAI, 2023. OpenAI. [Online] Available at: https://openai.com/blog/introducing-chatgpt-and-whisper-apis?ref=blog.orgspace.io [Accessed 28 07 2023].

Ramage, D., D. **Hall,** R. **Nallapati,** C. D. **Manning** (2009). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora.. Proceedings of the Conference on Empirical Methods in Natural Language Processing, Issue Association for Computational Linguistics, pp. 248–256.

Rijcken, E. et al. (2023). Towards Interpreting Topic Models with ChatGPT'. Daegu, IFSA.

Röder, M., A. **Both** & A. **Hinneburg** (2015). Exploring the Space of Topic Coherence Measures. Shanghai, Association for Computing Machinery.

Uthirapathy, S. & D. **Sandanam** (2023). Topic Modelling and Opinion Analysis On Climate Change Twitter Data Using LDA And BERT Model. Procedia Computer Science, Issue 218, pp. 908–917.

Wang, X., W. **Zhu** & W. Y. **Wang** (2023). Large Language Models Are Implicitly Topic Models: Explaining and Finding Good Demonstrations for In-Context Learning. arXiv, Issue arXiv:2301.11916.

СРАВНИТЕЛЕН АНАЛИЗ НА ТЕМАТИЧНОТО МОДЕЛИРАНЕ И ГОЛЕМИТЕ ЕЗИКОВИ МОДЕЛИ ПРИ ИЗВЛИЧАНЕТО НА ИНФОРМАЦИЯ ОТ СЪДЪРЖАНИЕТО НА СОЦИАЛНИТЕ МЕДИИ

Резюме: *Експоненциалното нарастване на текстовите данни, обусловено от комуникационните технологии, представлява предизвикателство за извличане на ценна информация. Това изследване се фокусира върху оценката на ефективността на алгоритмите за тематично моделиране BERTopic и Top2Vec в сравнение с големи езикови модели (LLM), особено ChatGPT и Google Bard в контекста на анализа на социални медии. По-конкретно се изследва способността на тези модели да извличат теми от корпус от данни в Twitter, отнасящи се до компанията Amazon. Методологията включва предварителна обработка на подмножество от повече от 720 000 туита, последвана от обучение на тематични модели и разработване на подсказки за LLM. В изследването са разработени количествени показатели за сравнение на възможностите за извличане на теми на моделите. Първоначалните резултати показват несъответствие в представянето на тематичните модели и LLM, като LLM демонстрират интуитивно извличане на теми от човека, но показват само 15% сходство в точните думи на темата и 23% сходство във вложенията на думите в сравнение с тематичните модели.*

Ключови думи: *тематично моделиране, промпт инженеринг, ChatGPT, Bard, Twitter*

Vitali Chaiko, PhD candidate
University of Library Studies and Information Technologies
E-mail: chaikovitali.cv@gmail.com