UNCOVER: Identifying AI Generated News Articles by Linguistic Analysis and Visualization

Lucas Liebe[1][a], Jannis Baum[1][b], Tilman Schütze[1][c],
Tim Cech[2][d], Willy Scheibel[1][e], and Jürgen Döllner[1][f]

[1]*University of Potsdam, Digital Engineering Faculty, Hasso Plattner Institute, Germany*
[2]*University of Potsdam, Digital Engineering Faculty, Germany*
{lucas.liebe, jannis.baum, tilman.schuetze}@student.hpi.uni-potsdam.de,
willy.scheibel@hpi.uni-potsdam.de, {tcech, doellner}@uni-potsdam.de

Abstract: Text synthesis tools are becoming increasingly popular and better at mimicking human language. In trust-sensitive decisions, such as plagiarism and fraud detection, identifying AI-generated texts poses larger difficulties: decisions need to be made explainable to ensure trust and accountability. To support users in identifying AI-generated texts, we propose the tool UNCOVER. The tool analyses texts through three explainable linguistic approaches: Stylometric writing style analysis, topic modeling, and entity recognition. The result of the tool is a prediction and visualization of the analysis. We evaluate the tool on news articles by means of accuracy of the prediction and an expert study with 13 participants. The final prediction is based on classification of stylometric and evolving topic analysis. It achieved an accuracy of 70.4 % and a weighted F1-score of 85.6 %. The participants preferred to base their assessment on the prediction and the topic graph. In contrast, they found the entity recognition to be an ineffective indicator. Moreover, five participants highlighted the explainable aspects of UNCOVER and overall the participants achieved 69 % accuracy. Eight participants expressed interest to continue using UNCOVER for identifying AI-generated texts.

1 INTRODUCTION

In recent years, artificial intelligence has become able to generate texts that are similar to those written by humans. While readers could easily recognize if a computer wrote a text just a few years ago, today's systems are getting increasingly better at producing convincing content leading to new challenges (Floridi and Chiriatti, 2020). One of the most impactful systems is ChatGPT[1] by OpenAI due to the great public attention it has gained (Lund and Wang, 2023). This release sparked many discussions in media about finding a way of identifying generated texts in the areas of: (1) Application of text generators in many

daily use systems like translators[2], (2) Evaluation of Students' writing skills[3] and (3) Validity of news sources[4]. This lead to regulators considering to halt AI development, until strategies for how to deal with such technologies are developed.[5] To tackle the issue, OpenAI released a prototype for a black box tool to identify texts from multiple generative models, which successfully identifies 26 % of AI authors (Jan Hendrik Kirchner, 2023).

Aside from working with the relatively low accuracy, end-users should not be required to put their trust in a black box solution which they have no control over (Rudin, 2019). A well explainable tool can build trust, by providing a deep understanding. Moreover, for such sensitive decisions humans need to be part of

---

[a] https://orcid.org/0009-0004-9252-4764
[b] https://orcid.org/0009-0008-8874-0279
[c] https://orcid.org/0009-0007-3321-9489
[d] https://orcid.org/0000-0001-8688-2419
[e] https://orcid.org/0000-0002-7885-9857
[f] https://orcid.org/0000-0002-8981-8583
[1] https://openai.com/blog/chatgpt

[2] https://www.forbes.com/sites/bernardmarr/2023/03/01/the-best-examples-of-what-you-can-do-with-chatgpt/
[3] https://abcnews.go.com/Health/wireStory/explainer-chatgpt-schools-blocking-96269407
[4] https://www.wired.com/story/ai-write-disinformation-dupe-human-readers/
[5] https://www.forbes.com/sites/jackkelly/2023/06/05/artificial-intelligence-is-getting-regulated/

a trustworthy solution to enable introspection. Visually intuitive and convincing explanations will make services in this domain more accessible than lengthy textual explanations.

This work applies explainable linguistic analysis to the task of identifying AI-generated text to offer an in-depth linguistic comparison of AI-generated and human-written texts. To achieve this, we introduce UNCOVER, which employs stylometric approaches, topic modeling, and entity recognition to analyze the linguistic features of news articles. We apply Stylometry, as a concept that is already successfully used to differentiate human authors, by implementing best practices in this field in an explainable way. For the topic modeling approach of UNCOVER, we propose the *"Topic Evolution Model" (TEM)*, that we derived from the "Topic Flow Model" by Churchill et al. (2018). In addition to re-implementing the original, we made various adjustments to their model and developed a visualization for the resulting topic graph. TEM resolves the requirement of large numbers of documents in each temporal period and is optimized to work with overall small corpora, instead of just small documents. The component featuring entity recognition mostly consists on coreference resolution and its visualization.

To evaluate UNCOVER, we conduct an expert study. 13 participants evaluated the tool on multiple usability aspects. Further, we introduce and test a novel AI-news data set for public benchmarking. The data set consists of training and evaluation data to compute metrics of accuracy. This data set is, to the best of our knowledge, the first publicly available, medium-sized-text data set featuring AI-generated news articles.

First, this work reviews related work for the identification of AI-generated text in subsection 2.1 and linguistic approaches to text analysis in subsection 2.2. The proposed tool consists of multiple components: a stylometric component, discussed in subsection 3.1, topic modeling, discussed in subsection 3.2, a prediction based on both of these components, explained in subsection 3.3, and entity recognition, found in subsection 3.4. We introduce our self-generated dataset in subsection 4.1 and the conducted expert study in subsection 4.2. Each of the four components – stylometry, topic modeling, prediction and entity recognition – is evaluated in subsection 4.3, subsection 4.4, subsection 4.5, and subsection 4.6 respectively. Explainability, limitations, threats to the validity, and possible negative impacts on society are discussed in section 5. We conclude this paper in section 6.

2 RELATED WORK

The tool UNCOVER builds upon work in the areas of (1) identification of AI-generated texts and (2) linguistic features, coherence analysis and authorship attribution.

2.1 Identifying AI Generated Text

OpenAI presented a service applicable to various generation models, where they achieved classification performance of 26 % AI texts correctly classified as such (Jan Hendrik Kirchner, 2023). Many other commercial classification tools are available online achieving good results, for instance gowinston.ai, contentatscale.ai, and gptzero.me. Other research approaches, like the Giant Language Model Test Room (GLTR) (Gehrmann et al., 2019), GROVER (Zellers et al., 2019), and a fine-tuned Bidirectional Encoder Representations from Transformers (BERT) (Ippolito et al., 2019) are trained to recognize the output of a singular model. GLRT provides a human-in-the-loop solution to help users make informed decisions instead of providing a prediction itself (Gehrmann et al., 2019). BERT models achieve the best accuracy of the mentioned approaches (Ippolito et al., 2019). However, this method has not been re-implemented and only evaluated on GPT-2, newer or multiple AI generators have not been tested, making it unclear how this approach would perform today. UNCOVER aims to correctly classify *Large Language Models (LLMs)* in general making the task more complex. AI-generated texts also attracted attention in the generation of fake news. Further research aims to detect such news through data mining techniques (Shu et al., 2017; Zhang and Ghorbani, 2020). However, this approach is not sufficient to UNCOVER since we aim to identify AI-generated texts independently of the facts.

2.2 Linguistic Text Analysis

By analyzing and schematically describing the contents of a text, linguistic text analysis can improve Natural Language Understanding (Zhang and Wang, 2022). We have identified three approaches relevant to our use case.

Stylometry. Human authors can be differentiated by the statistical distribution of "Style Markers" (Houvardas and Stamatatos, 2006). "Style markers" are n-grams of textual features that can consist of characters, words, or part-of-speech tags. Houvardas et al. found that 3- (tri), 4-, and 5-grams contain the most information for successful author iden-

tification and further highlighted trigrams for shorter texts (Houvardas and Stamatatos, 2006). Posadas-Durán et al. (2017) introduced an algorithm for extracting syntactic n-grams from sentences using a dependency tree. These were found to perform better than most other n-grams, with character n-grams in second place (Ríos-Toledo et al., 2022). However, the proposed measures were not tested on AI generators. UNCOVER uses character and syntactic trigrams for its analysis.

Topic Modeling. The variety of topic models is ever-growing, and many of the more recent models employ less explainable AI methods such as neural networks (Churchill and Singh, 2022). Models based on these methods are not applicable to UNCOVER due to its constraint of being explainable. Another branch of topic modeling research has focused on creating graph-based methods, which are more explainable by nature. A more recent graph-based models is Topic Flow Model, which is used to produce semantic graphs that describe how topics change throughout defined temporal periods of the text corpus (Churchill et al., 2018). While the Topic Flow Model was created to work with short documents within the corpus' periods, the number of documents in each period needs to be large for it to produce relevant output.

Entity Recognition. *Named Entities (NE)* have been found to positively influence the performance of Machine Learning Systems that require context information (Zhang and Wang, 2022). The Stanford Named Entity Recognizer uses multiple machine learning sequence models and rule-based components to label 12 different NE classes (Finkel et al., 2005). Entity Grids are a way of representing a text to capture the location in which NE occur and can be a measure of coherence (Mohiuddin et al., 2018). One method of achieving this is called *coreference resolution*, that finds all expressions in a text that refer to the same entity (Clark and Manning, 2016). Due to the generation of Language Models word-by-word, we expect to find and observed abnormalities in AI-generated text.

3 APPROACH

UNCOVER uses the linguistic approaches of stylometry, topic modeling, and entity recognition. The final tool and its results are presented in a web interface found through the projects GitHub repository[6].
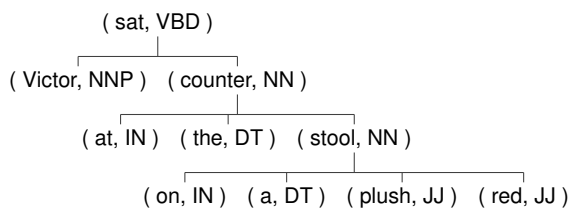
---

[6] https://github.com/hpicgs/unCover

Figure 1: Syntactic dependency tree for "Victor sat at the counter on a plush red stool.".

3.1 Stylometry

UNCOVER employs character and syntactic trigrams. The trigrams are obtained from the text as it is, i.e. without preprocessing, since techniques such as character normalization or stemming would disrupt character trigrams (Chen and Manning, 2014). The distribution of syntactic trigrams is extracted from dependency trees generated by the Stanford NLP Dependency Parser (Chen and Manning, 2014) (Figure 1). Counting all discovered character triples creates the character trigram distribution. Both distributions are trimmed to only include the top 100 most common trigrams. One logistic regression model is trained for each author and distribution. For classification, models are split into Humans and AI and the highest score of each group is compared against a threshold. If both groups or no group hits the threshold, the stylometry component cannot decide on the author. This case is represented as an "unsure" result. If only one of the groups passes the threshold, the text is classified to be written by that group.

3.2 Topic Modeling

Coherent articles are texts where the covered topics are introduced, change slightly to cover different aspects, and then get replaced by a different topic, or evolve into a connected and advanced theme. UNCOVER analyzes and describes this specific trait. The patterns found when analyzing the topics are then used to differentiate human-written and AI-generated texts. To illustrate how this is achieved by UNCOVER, we will first introduce *Topic Evolution Model (TEM)*, then go into detail on how it is integrated into the tool, and finally explain how texts are classified based on TEM's output.

3.2.1 Topic Evolution Model

UNCOVER's most important requirement for topic modeling is being able to generate an explainable overview of the change of topics in a single news article. Churchill et al. (2018) have introduced *Topic Flow Model (TFM)* as a graph-based model to analyze changes to topics over time in a large corpus con-

sisting of multiple temporal periods, each containing many documents. Based on this work, we have developed *Topic Evolution Model (TEM)* to fit our requirements of working with significantly smaller corpora such as a single news article, where single paragraphs make up the periods, each consisting of individual sentences as documents. In the following, we will illustrate TEM's main differences to TFM.

First Period. Churchill et al. (2018) introduce *nutrition* (*nut*), and *energy* values for words in the corpus, given by the following equations

$$nut(w)_p = (1-c) + c \cdot \frac{tf(w)_p}{tf(w_p^*)_p} \qquad (1)$$

$$energy(w)_p = \sum_{t=1}^{p} \frac{1}{i}(nut(w)_t^2 - nut(w)_{t-i}^2), \qquad (2)$$

where w and p are the given word and period, w_p^* is the most common word in p, $tf(w)_p$ is the term frequency of w in p, and $c \in [0, 1]$ is a tuning parameter. By definition, the *energy* (Equation 2) of all words in the first period in TFM is equal to 0. This may lead to all words falling through the energy threshold, meaning no emerging words are found. TEM instead sets the energy of words in the first period equal to their squared nutrition to allow for the existence of emerging words in the first period.

Flood Words. TFM classifies all words that appear in at least half of all documents in a period as flood words (Churchill et al., 2018), an a posteriori alternative of stop words that should be ignored in the analysis. Therefore, all words are flood words when a period only has two documents. To enable processing paragraphs with only two sentences, TEM classifies words as flood only when they appear in more than half of all documents.

Correlation. For all pairs of words in a period, TFM applies the following formula for term correlation $c_{k,z}$ of word k to word z at time t

$$x_{k,z}^t := \frac{n_{\{k,z\}}/(n_{\{k\}} - n_{\{k,z\}})}{(n_{\{z\}} - n_{\{k,z\}})/(|D_t| - n_{\{z\}} - n_{\{k\}} + n_{\{k,z\}})} \qquad (3)$$

$$c_{k,z}^t = \log\left(x_{k,z}^t\right) \cdot \left| \frac{n_{\{k,z\}}}{n_{\{k\}}} - \frac{n_{\{z\}} - n_{\{k,z\}}}{|D_t| - n_{\{k\}}} \right|, \qquad (4)$$

where n_A is the number of documents all words in A co-occur in, and $|D_t|$ is the number of documents in period t (Churchill et al., 2018).
This formula has an edge case where division by 0 occurs when two terms in a period only co-occur and
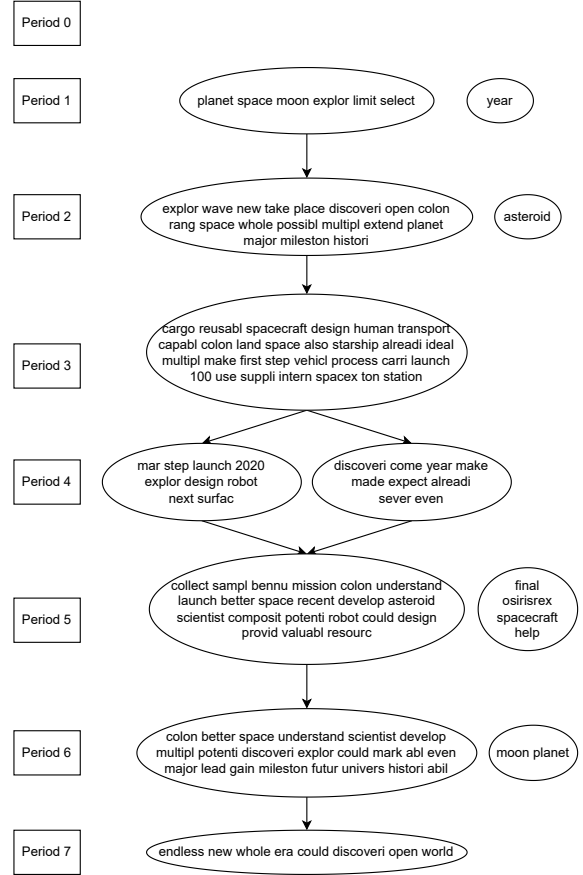

Figure 2: Example of an AI-generated Topic Evolution graph by GPT-3.

never occur in a document by themselves. The way this is handled by TFM is unknown. When the number of documents in a period is low, this edge case is common. TEM solves this by recursively merging strictly co-occurring terms into the same node in the semantic graph. This way, the correlation between strictly co-occurring terms never has to be evaluated. All terms of the node are accounted for in the formula, to consider nodes with multiple terms during topic distance evaluation.

Topic Distance. TFM matches newly discovered topics to existing themes soley based on the recurrence of a single leader term that describes the theme (Churchill et al., 2018). When the number of documents in multiple periods is small, so is the total number of terms in each of these periods. This makes it likely that a leader term doesn't reoccur in a topic, even when the theme persists. To be able to recognize existing themes despite the leader term's potential absence, TEM reemploys the measurement of topic distance by comparing all new emerging topics with the predecessor period's topics. The topic distance

between a pair of topics is given by Churchill et al. (2018) as follows

$$td_{t_1,t_2} = \frac{\min(|t_1 \setminus t_2|, |t_2 \setminus t_1|)}{|t_1 \cap t_2|}, \tag{5}$$

with t_1 and t_2 being the sets of words in the two topics.

TEM can process corpora containing periods with as low as two documents, and therefore meets UN-COVER's requirements by allowing it to process individual articles. TEM generates a list of periods. Each period contains a list of topics, which in turn have a list of words and a theme identifier. Discovered topics that are sufficiently similar to an existing topic receive the same theme identifier as the existing topic, which describes the topic evolution.

3.2.2 Integration into UNCOVER

To analyze a text with TEM, UNCOVER first splits the text into paragraphs by considering newline characters, and each paragraph into sentences by considering sentence termination characters such as periods or question marks. Should a paragraph consist of only a single sentence, it is merged with its predecessor. This ensures that each paragraph is a made up of at least two sentences to make them meet the requirements for periods of documents in TEM's input. Before the text is passed into TEM, all non-alphanumerical characters and stop-words are removed, all letters are transformed into lowercase, and all words are stemmed. TEM's output is finally visualized with a vertical, directed graph. The lists of words for each topic in a period are placed in horizontally aligned nodes, and these rows are aligned from top to bottom. The nodes of topics with a theme identifier that has occurred in the predecessor period are connected with a directed edge to the respective predecessor topic. Figure 2 shows one example of such topic evolution graphs.

3.2.3 Classification

Aside from serving as a user-directed visualization, the discussed graph is also used for automatic classification with a multinomial logistic regression on the following *Topic Evolution connectivity metrics (TEcm)*:

1. The absolute value of 1 minus the ratio of the number of distinct theme identifiers, and the total number of topics

2. The ratio of the number of topics with the most common theme identifier, and the total number of topics

3. The number of periods that have at least one incoming edge, i.e. a topic with predecessor topic, divided by the total number of periods minus 1

4. The ratio of the longest chain of connected periods, and the total number of periods

All of the connectivity metrics represent a different interpretation of *connectedness* of the graph between 0 and 1, with 0 being the least, and 1 being the most connected.

3.3 Final Prediction

The prediction of stylometry and the prediction of *Topic Evolution connectivity metrics (TEcm)* are combined together into a single final output through a basic decision tree. To begin with, we output the predictions if both components agree. If the stylometry result is uncertain or TEcm's confidence level is over 80 %, we immediately output the TEcm classification. If stylometry predicted the text to be AI-generated, we output this decision if TEcm's confidence level is below 70 %. However, if the text is predicted to be human-written by stylometry, we decrease the threshold to 60 % to minimize misclassifying human authors. Finally, if the confidence exceeds the threshold, we present an uncertain outcome. Figure 3 shows one example for a final prediction of an AI-generated text presented to the user after running the full analysis.

3.4 Entity Recognition

A coherently written text should introduce and mention entities in an orderly way that does not confuse its readers. Therefore, especially the change and occurrence of different entities may offer a relevant indicator for an AI generator's weaknesses. Stanford's CoreNLP Parser offers the ability to recognize and track entities in multiple sentences (Finkel et al., 2005). For entity recognition, UNCOVER uses the parser with its default parameters. For coreference resolution, CoreNLP offers various models, out of which UNCOVER uses the most accurate neural network. The output is visually represented as a vertical stacked bar chart, where each sentence is displayed together with a stack of bars (see Figure 4). Each bar
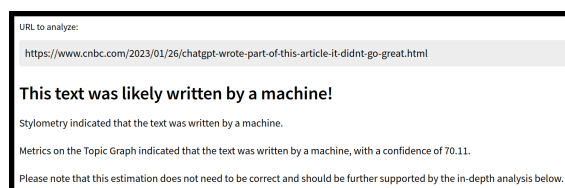

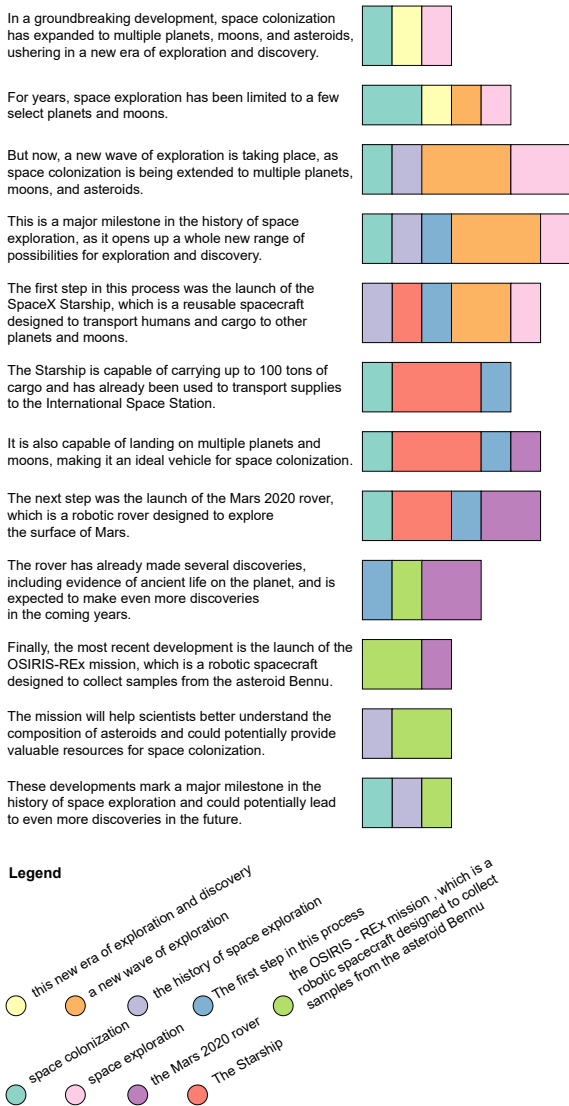Figure 3: Screenshot covering an example classification.

Figure 4: An example of an entity occurrences diagram. The visualization may be challenging to comprehend, because the used parser is not perfect.

represents an entity mentioned in a rolling window that consists of the respective sentence, its predecessor, and its successor. The bars are sized according to the number of times the entity was mentioned in the corresponding rolling window.

4 EVALUATION

We evaluate all three components of UNCOVER regarding accuracy and explainability. To represent the difference of effect of the answers, the weighted F1-score is calculated with half penalty for unsure answers and a full penalty for opposite answers.

4.1 Dataset

Our final datasets consist of scraped news articles written by human authors and generated news articles sourced between March and Mai 2023. In total, the training dataset contains 2837 news articles by five different authors scraped from the Guardian[7] as a big trust-worthy news outlet. It also comprises 2400 articles that were created using five distinct LLM queries, with 600 articles generated for each query. To write news articles on this topic, GPT-2 (Radford et al., 2019) was given news titles as a prompt. Similarly, we utilized GPT-3 (Brown et al., 2020) twice - first, we tasked it in the same manner as GPT-2, and second, we provided the full-text news article and asked GPT-3 to rewrite it as a news article. This approach allowed us to evaluate possible divergences in the results with queries containing different amounts of context. To match its input style, GROVER (Zellers et al., 2019) was given full news articles and titles, along with the original URL and other mock data. The original news articles were scraped from Google News using search queries of similar topics as the Guardian articles.

To ensure the accuracy of our components in real-world scenarios, we devised a separate dataset for testing purposes. We included 200 articles each from human authors, GPT-2, GPT-3, and GROVER. The GPT-3 articles comprised 100 titles-only and 100 full-text generated articles. For this test dataset, we chose ten general news queries like "environmental concern", scraped the top 15 results of Google News, and generated articles as described for the training dataset. This time, pages that could not be scraped with our system or had excessively long articles were skipped entirely to ensure all classers contained the same 200 sources.
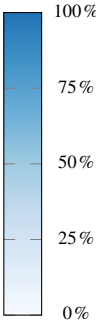
4.2 Expert Study

As another mean of evaluation, an expert study was set up to evaluate UNCOVER's practical use and explainability with 13 participants from an academic machine learning or Natural Language Processing background. They were provided with ten unique random texts from the dataset and the corresponding UNCOVER output and asked to classify the texts while using the tool. The participants were questioned about the use and understandability of UNCOVER and it's individual visualizations. On average, the participants classified texts correctly in 69 % of cases. As to the usefulness of UNCOVER, half of the participants stated that UNCOVER's output changed

---

Table 1: Assessment results from the classification of news articles using character trigrams (a), syntactic trigrams (b), combined stylometry (c), topic evolution connectivity metrics (d), and final prediction (e).

	(a) Character Trigrams			(b) Syntactic Trigrams			(c) Combined Stylometry			(d) TEcm			(e) Final Metric		
	Machine	Human	Unsure	Machine	Human	Unsure	Machine	Human	Unsure	Machine	Human	Unsure	Machine	Human	Unsure
GPT-2	80%	6%	14%	67%	6%	27%	86%	6%	8%	84%	16%	0%	91%	1%	8%
GPT-3	58%	18%	24%	70%	8%	22%	69%	12%	19%	81%	19%	0%	85%	4%	11%
GROVER	24%	37%	39%	76%	7%	17%	53%	14%	33%	55%	45%	0%	62%	21%	17%
Human	14%	54%	32%	57%	19%	24%	32%	33%	35%	52%	48%	0%	35%	45%	20%

their mind about the origin of the text two or more times out of ten and most of them would want to use the tool again next time. Before being questioned about it or informed of the goal, five participants highlighted the explainable aspects of the tool. More detailed descriptions of the result can be found in the following subsections or in the Appendix subsection A.1

4.3 Stylometry

For the training of a model, the training dataset was split into 80 % training and 20 % validation data. The final accuracy of the stylometry regression on the validation data was 73 % with less than 5 % of human-written texts predicted as AI-generated. This changed on the test dataset, where the model combining both trigrams achieves a total accuracy of 59.3 % and weighted F1-score of 80.66 % (see Table 1c). Syntactic trigrams alone had an accuracy of 57.6 % (see Table 1b) and performed slightly better than character trigrams with an accuracy of 53.4 % (see Table 1a). However, since syntactic trigrams performed much worse on texts labeled as human-written, they can not be considered overall better. A result of the expert study was that the stylometry approach, as a main part of UNCOVER's prediction, had the highest influence on the participants decisions when classifying texts. Since a logistic regression model was used for classification and the same 100 trigrams were used as features for every single trained regression, the classifications are still explainable to technical users, as the model contains relevance values for each trigram.

4.4 Topic Modeling

For testing, we conducted Mann-Whitney tests (Nachar, 2008) on all pairs between GPT-2 (Radford et al., 2019), GROVER (Zellers et al., 2019), GPT-3 (Brown et al., 2020)

(two groups with different prompt styles), and humans (three different authors). All of these tests proved a significantly different mean between the groups with $p < 5\%$, with the exception of connectivity metric 3 between GPT-2 and humans, and connectivity metric 2 between GROVER and humans. Figure 5 shows density plots comparing the connectivity metrics for human-written texts (solid line) to those generated by different AI models (dashed lines).

The logistic regression classification based on these TEcm achieves an overall accuracy of 67.4 % and weighted F1-score of 72.77 %. A more detailed analysis of the results can be seen in Table 1d. On the validation dataset we achieved an average 77.98 % accuracy using 5-fold-cross-validation. In comparison to the Stylometry on both datasets TEcm performs slightly better. However, it also does not manage to generalize better loosing accuracy on the test dataset aswell. The smaller F1-score and many wrong predictions on human authors are mainly caused by the lack of certainty in TEcm classification.

Figure 2 shows an example of a Topic Evolution graph. This graph predominantly consists of a single topic evolving throughout the majority of temporal periods (article paragraphs). We have found the Topic Evolution graphs to give clear insights into how themes develop over the course of the analyzed articles. Graphs for human-written articles tend to have multiple distinct sections of a few periods in length that are internally very connected. In contrast, graphs for AI-generated articles often contain a single evolving topic that spans most of the article and the majority of nodes. This finding is supported by the conducted expert study, with participants judging the graphs to have a median "understandable" (4/6) clarity and having an average higher influence on the participants' decision than the entity diagram with a median of "strong" (5/6).

(a) connectivity metric 1.

(b) connectivity metric 2.

(c) connectivity metric 3.
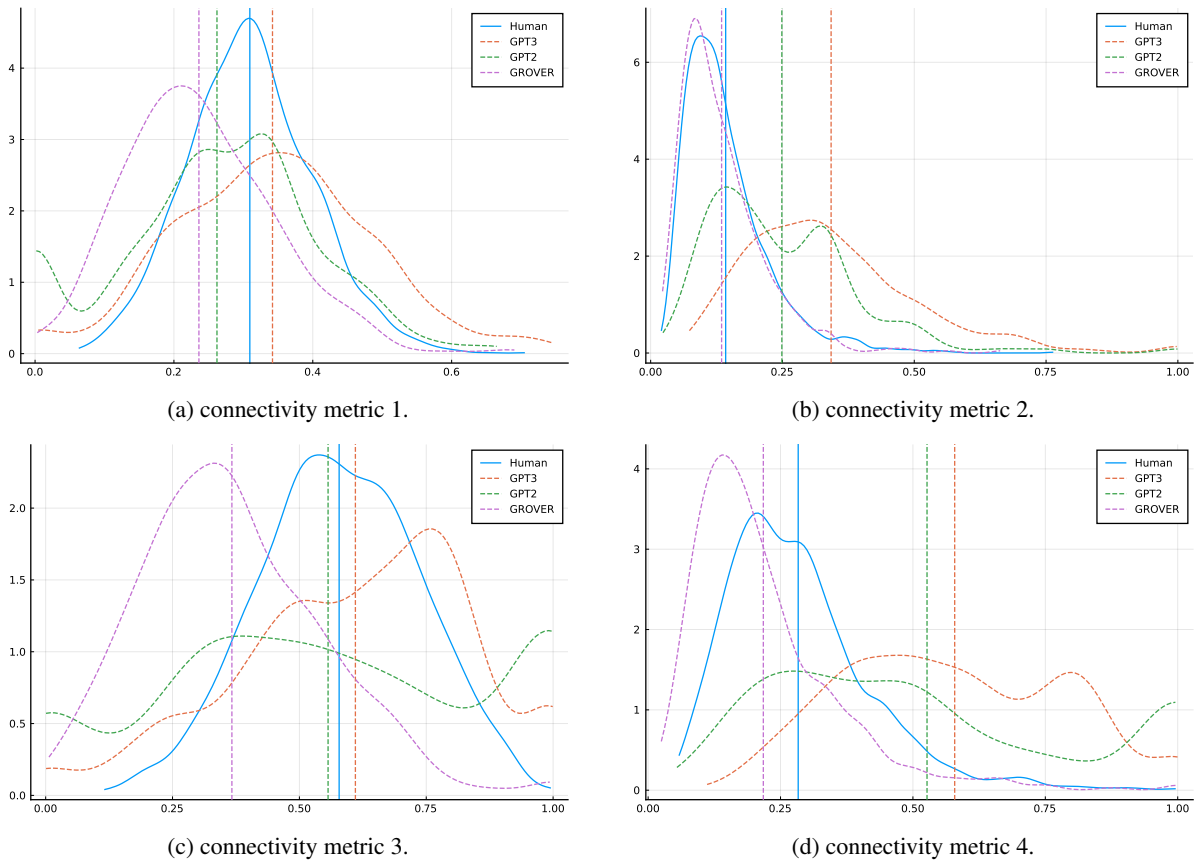
(d) connectivity metric 4.

Figure 5: Density plot for the different *Topic Evolution connectivity metrics (TEcm)* comparing humans (solid line) to different AI generators (dashed lines) with mean values shown as vertical lines.

4.5 Final Prediction

The combined result of TEcm and Stylometry regression achieved an accuracy of 70.4 % and weighted F1-score of 85.65 %. 14.2 % of the classifications are unsure and 15.8 % are assigned to the wrong label. In Table 1e the results are presented in more detail. The participants in the expert study described this aspect as the best component of UNCOVER, because it contains the most direct indication on the author of a text. Four participants said that they would use the tool in the future only because of the prediction. The output of UNCOVER aligned with the intuition of eight participants on self-chosen articles and led to three participants questioning their intuition.

4.6 Entity Diagram

From our experience, this approach was not successful at distinguishing AI-generated text from human-written text. We found that many misclassifications occurred due to inaccuracies of the Stanford NLP parser. This behavior leads to an explosion of newly introduced entities in texts, making it harder to find patterns for human authors and Language Models. While the Named Entity Recognizer itself is only based on explainable machine learning models, the used coreference resolution model is a neural network. Therefore it is the only part in UNCOVER's components to be considered non-explainable. Figure 4 shows UNCOVER's visualization of entity occurrence.

UNCOVER's visualization of the Stanford NLP parser's output is overall hard to oversee, because of the number of sentences in an article and the number of entities used. This was also noticed by the participants in the expert study. Eleven of 13 subjects noted that the diagram does not produce perceivable differences in entity occurrence patterns based on whether the text was written by a human or generated by AI. The Entity Diagram also turned out to be the worst rated component in the tool on clarity and helpfulness. While, in the context of our study, the total number of entities used per sentence may give some indication of complexity, it is not possible to reliably distinguish AI-generated texts from human-generated

ones. Even though the participants disliked this component for the task of classification, three of them said they would want to use it for other tasks. For instance, one subject said that it is incredibly helpful to identify parts of the text that are interesting to him.

5 DISCUSSION

UNCOVER can differentiate human from machine authors. It achieves a high classification accuracy on multiple state-of-the-art generation models, while not requiring a pre-trained *Large Language Model (LLM)*. For instance OpenAI's classifier achieved 26 % correctly classified AI texts (Jan Hendrik Kirchner, 2023), while UNCOVER achieves 70.4 %. In addition, through the provided visualization on topic evolution, a user can better analyze structures inside a text and make an informed decision.

As we think it is most undesirable to make false claims about human authors, we set up the training to reduce the occurrence of incorrectly classified human texts which is represented in our weighted F1-score of 85.65 %. However, the highest error rate is found among human authors in the final evaluation. This is at least partly due to the fact that the human-written texts in the validation data set consisted of randomly crawled news articles. In comparison, the models used were trained to specifically recognize texts written by five authors from the same publisher, indicating a similar writing style.

When we compare the classification performance of different models, we can see that they differ in their ability to resemble human-written texts based on the concept used. For instance, GPT-2 can be more readily identified when using character trigrams, whereas GPT-3 unexpectedly is identified better than GPT-2 when using syntactic trigrams. This highlights the need for AI-detection methods to incorporate multiple text analysis concepts and cover a wide range of aspects. In the evaluation of GPT-3 articles, we combined the two queries to generate them into one accuracy, as no differences in their data were found in our experiments.

Through the expert study, *Natural Entity Recognition (NER)* has not been proven to benefit the goal of UNCOVER but still offers further insights into the text. Other researchers also pointed out, that NER needs to be improved as a separate component to achieve more consistent results (Zhang and Wang, 2022). The expert study showed, that the tool overall builds trust with users, making them more secure in their decisions.

Achieved Explainability. The primary objective of UNCOVER was to provide clear explanations for a system's decisions, using visual aids that enable users to make informed choices. While one visualization has been found to be useful in an expert study, more and better visualizations need to be developed to improve explainability. Nevertheless, we have achieved complete technical explainability of all crucial components. The decision-making process of UNCOVER is based on a decision tree that employs explainable metrics. Experts familiar with logistic regression can interpret the trained models and understand the tool's decision based on this. The topic modeling metrics can be derived from the topic graph visualization. However, due to the vast number of parameters and the complexity of graph algorithms, the decision-making process cannot be fully comprehended by an average user.

Limitations. UNCOVER currently only works for English texts and its approaches might lose accuracy quickly, because of the rapid advances made by LLMs. New model releases, like GPT-4 (OpenAI, 2023), or fine-tuned systems that are developed to break our analysis are an ever-existing threat. We already documented differences between classification results of the used models and might experience worse performance on other models that have not been included in training as well. When evaluating the performance on human-authored texts we can already see this effect that trained authors achieve much better accuracy. On the same subject, the tool was only evaluated on self-generated news from three LLMs and needs further evaluation on more data to evaluate how generalized the performance is.

Threats to Validity. A potential drawback of the reported findings is that the accuracy calculations were based on data generated through simplistic prompts. We compensated this effect by providing two different prompts to GPT-3 in the generation of the dataset and therefore including multiple levels of complexity. Newer models can take specific task queries to follow a specific story line which would make the topic modeling classification more difficult. Similarly, style transfer is a concept where LLM mimic a certain writing style (de Rivero et al., 2021), which would make the detection via stylometry impossible if applied successfully. Further, we assumed that news outlet texts on the top of Google News are written by humans when creating test data, which may not be accurate anymore and impact our human classification accuracy. The tool also was only evaluated on English texts, specifically news articles, which carry specific

language traits, which can differ a lot between languages and text types. Therefore, the evaluations are not generalizable on a larger scale. The reported performance has also not been compared to other tools available online, making it difficult to judge its efficiency.

The conducted study has limited abstraction potential since only 15 participants took part. Further, it only questioned participants studying towards a degree in computer science at the same institution with prior experience in artificial intelligence. Therefore the results of the study are sampling biased and not generalized. Also, most participants self-applied to the study because they are interested in the area of Natural Language Processing, however these effects where necessary to accept to achieve a larger number of participants. Another factor is the bias found in UNCOVER's visualizations, as they have been created to visualize effects that we observed. By evaluating them in an expert study we found that people build similar intuitions, which could have been influenced by the style of presenting the visualizations.

Negative Societal Impact. The accuracy of the identifier is not high enough to trust the output but may influence people's opinions. This effect can lead to problems when deploying such services. The proposed tool can potentially bring injustice to authors with writing patterns similar to generative models. For instance, non-native authors could struggle to write as coherently as native writers. These authors could struggle by facing wrongful social judgment of being framed for not writing their publications by themselves. In the same way, authors that use generators might get exposed by such tools.

6 CONCLUSIONS

We presented UNCOVER, a tool that uses concepts of linguistic text analysis to distinguish between human-written and AI-generated news articles. The concepts considered are stylometry, topic modeling, and entity recognition. For topic modeling, we introduced *Topic Evolution Model (TEM)*. A final classification and two visualizations are shown to the user of UNCOVER inside a web interface.

We evaluated the tool on news articles by means of accuracy of the prediction and an expert study with 13 participants. Stylometry was found to overall be able to identify AI-authors. TEM is very successful in describing topics and their development, while also being a good measure for theme coherence. In the study, participants preferred to base their assessment

on the prediction and the topic graph, while rating the entity recognition as the least effective indicator. Eight participants expressed interest in continuing to use UNCOVER for identifying AI-generated texts and five participants highlighted the explainable aspects.

Because we observed common inconsistencies in how entities occur in AI-generated texts, we think entity recognition can become helpful if developed further. Therefore, we will look into an improvement in coreference resolution and different visualizations of this component. During the development of the tool, new LLMs already have been released. UNCOVER should be evaluated with models like GPT-4 (OpenAI, 2023) and Google's PaLM 2 (Anil et al., 2023). Besides the evaluation on different models, we only evaluated the performance on news articles. In the future, the accuracy should be evaluated on differently sized and differently structured texts. Other linguistic approaches, like sentiment analysis could be tested and added to the tool. Finally, a larger user study should be considered to complement the positive findings of this work.

To ensure reproducibility of the tool and results, the code is published open-source[8] together with our generated news dataset.

ACKNOWLEDGEMENTS

REFERENCES

Anil, R., Dai, A. M., Firat, O., Johnson, M., et al. (2023). PaLM 2 technical report. *arXiv CoRR*, cs.CL. arXiv:2305.10403.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In *Proc. Neural Information Processing Systems*, NeurIPS '20, pages 1877–1901. Curran Associates, Inc.

Chen, D. and Manning, C. D. (2014). A fast and accurate dependency parser using neural networks. In *Proc.*

---

[8]https://github.com/hpicgs/unCover

SIGDAT Conference on Empirical Methods in Natural Language Processing, EMNLP '14, pages 740–750. ACL.

Churchill, R. and Singh, L. (2022). The evolution of topic modeling. *ACM Computing Surveys*, 54(10):215:1–35.

Churchill, R., Singh, L., and Kirov, C. (2018). A temporal topic model for noisy mediums. In *Proc. 22nd Pacific-Asia Conference on Knowledge Discovery and Data Mining*, PAKDD '18, pages 42–53. Springer.

Clark, K. and Manning, C. D. (2016). Deep reinforcement learning for mention-ranking coreference models. In *Proc. SIGDAT Conference on Empirical Methods on Natural Language Processing*, EMNLP '16, pages 2256–2262. ACL.

de Rivero, M., Tirado, C., and Ugarte, W. (2021). FormalStyler: GPT based model for formal style transfer based on formality and meaning preservation. In *Proc. 13th International Conference on Knowledge Discovery and Information Retrieval*, KDIR '21, pages 48–56. SciTePress.

Finkel, J. R., Grenager, T., and Manning, C. D. (2005). Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proc. 43rd Annual Meeting of the Association for Computational Linguistics*, ACL '05, pages 363–370. ACL.

Floridi, L. and Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Springer Minds and Machines*, 30:681–694.

Gehrmann, S., Strobelt, H., and Rush, A. M. (2019). GLTR: Statistical detection and visualization of generated text. In *Proc. 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, ACL '19, pages 111–116. ACL.

Houvardas, J. and Stamatatos, E. (2006). N-gram feature selection for authorship identification. In *Proc. 12th International Conference on Artificial Intelligence: Methodols, Systems, and Applications*, AIMSA '06, pages 77–86. Springer.

Ippolito, D., Duckworth, D., Callison-Burch, C., and Eck, D. (2019). Automatic detection of generated text is easiest when humans are fooled. In *Proc. 58th Annual Meeting of the Association for Computational Linguistics*, ACL '19, pages 1808–1822. Association for Computational Linguistics.

Jan Hendrik Kirchner, Lama Ahmad, S. A. . J. L. (2023). OpenAI AI classifier. https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text. last Accessed: 2023-03.

Lund, B. D. and Wang, T. (2023). Chatting about ChatGPT: how may AI and GPT impact academia and libraries? *Library Hi Tech News*, 40(3):26–29.

Mohiuddin, T., Joty, S., and Nguyen, D. T. (2018). Coherence modeling of asynchronous conversations: A neural entity grid approach. In *Proc. 56th Annual Meeting of the Association for Computational Linguistics – Volume 1: Long Papers*, ACL '18, pages 558–568. ACL.

Nachar, N. (2008). The Mann-Whitney U: A test for assessing whether two independent samples come from the same distribution. *Tutorials in Quantitative Methods for Psychology*, 4(1):13–20.

OpenAI (2023). GPT-4 technical report. *arXiv CoRR*, cs.CL. arXiv:2303.08774v3.

Posadas-Durán, J.-P., Sidorov, G., Gómez-Adorno, H., Batyrshin, I., Mirasol-Mélendez, E., Posadas-Durán, G., and Chanona-Hernández, L. (2017). Algorithm for extraction of subtrees of a sentence dependency parse tree. *Acta Polytechnica Hungarica*, 14(3):79–98.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. Technical report, OpenAI.

Ríos-Toledo, G., Posadas-Durán, J. P. F., Sidorov, G., and Castro-Sánchez, N. A. (2022). Detection of changes in literary writing style using n-grams as style markers and supervised machine learning. *PLOS ONE*, 17(7):1–24.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36.

Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., and Choi, Y. (2019). Defending against neural fake news. NeurIPS '19. Curran Associates, Inc.

Zhang, C. and Wang, J. (2022). Tag-Set-Sequence learning for generating question-answer pairs. In *Proc. 14th International Conference on Knowledge Discovery and Information Retrieval*, KDIR '22, pages 138–147. SciTePress.

Zhang, X. and Ghorbani, A. A. (2020). An overview of online fake news: Characterization, detection, and discussion. *Elsevier Information Processing & Management*, 57(2):102025:1–26.
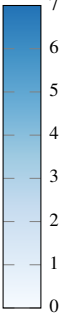
A APPENDIX

A.1 Expert Study Design

The expert study with 13 participants was conducted using Google Forms[9] to collect the given answers and guide the participants through the process. The participants were observed by an instructor in person during the time of the study to capture direct comments on the components. Before the study, participants were told that the efficiency of the tool was to be evaluated and the background in explainable AI was not mentioned explicitly. During the session, instructors minimized their communication with the participants, ex-

---

[9]https://docs.google.com/forms/d/1-dLyWXKx01stPUcldRJ35dovfQmpuRyxcdtIqSj6bjk/prefill

Table .2: Answers to "How strong did each component influence your decision?" (a) and "Please rate the visualizations based on their clarity." (b).

(a)

	not at all	very little	little	strong	stronger	very greatly
Prediction	0	2	2	4	5	0
Topic Graph	1	1	2	7	2	0
Entity Diagram	2	2	7	2	0	0

(b)

	no clue	confusing	difficult	understandable	easy	very easy
Topic Graph	0	0	1	5	6	1
Entity Diagram	0	2	6	4	1	0

cept for a question and answer session in the middle of the session. One 14th expert was used to evaluate the design of the study before collecting the final results. After this first execution, we adopted some unclear instructions, added more information, and changed the layout of the examples.

Detailed Course of the Study The first step asks participants for their participant ID to reconstruct the results. The ID could not match to participants' names but the test data in the main section. Then participants indicate their experience with machine learning and Natural Language Processing on a scale from 1 (unexperienced) to 6 (expert). For machine learning, we observed one participant entering 2, three participants entering 4 and 5 each, and six participants entering 3. In regards to Natural Language Processing, one participant entered 1, three participants entered 3 and 4 each, and six participants entered 2.

In the second step, UNCOVER's components are introduced in brief texts to test how easily participants can understand the tool. The introduction is written to explain how the components work and what information they offer in three Texts with less than 130 words each and contains two pictures. To leave the participants unbiased, common patterns and other beneficial information, based on our experience, to separate human and AI-generated texts are left out. Afterward, participants were questioned on their understanding of the components. Entity Diagram and overall Prediction both achieved one "perfect" vote and six votes each for "good" and "better". Topic Graph received nine "better" and four "good" votes.

The third step is time to ask comprehension questions to the instructors, six did not use this opportunity and continued further on their own.

The main part of the study is taking place in the fourth step, where participants have to evaluate ten unique texts that were randomly chosen from our dataset. Five participants talked about good explainability while working on this task by themselves. Most participants correctly classified seven of the given texts, rarely choosing unsure, and achieving a total accuracy of 69%. Then they explained their decision process and what they looked out for. The participants used the text a lot, paying attention to similar sentence structures, synonyms and choice of words, and punctuation. Only five participants named components of UNCOVER but claimed to discover patterns and find them helpful to classify the examples. Next, each participant analyzed one text, that they chose themselves and gave to us before the start of the study. In eight cases the self-chosen text was analyzed according to the participants own perspective on the text author.

In the fifth and final step of the study we asked participants how often they changed their opinion on a text based on UNCOVER. Six participants said two times, three said one time, two said four times, and zero and seven times were answered by one participant each. In the next question, participants indicated how much each component influenced their decision. The results are shown in Table .2. It also shows the results of the ratings on the clarity of the visualizations. The last question in the questionnaire asked the participants how likely they would want to use this tool on their own on a scale from 0 (never) to 6 (always). To this, one participant answered 2, five participants chose 5, four participants chose 3, and three selected 4. Three participants who gave a score of 3 or below stated that they believe their own estimation of a text author is sufficient.

After the study, we held a debriefing with the participants to give them the opportunity to ask further questions about the tool. This part showed that the participants had a great interest in the functionality of the tool and enjoyed to use it.