Scalable Detection of chatGPT-generated Text

Anders Fongen
Norwegian Defence University College (FHS)
Lillehammer, Norway
email: anders@fongen.no

***Abstract*—Text generated by services based on Artificial Intelligence (AI) should optionally be identified for the sake of exam validity, and intellectual honesty in general. This paper reports from an ongoing study on detection techniques to identify text generated by the AI-based chatGPT service, where text documents are represented by high-dimensional feature vectors on which calculations are employed to reveal semantic similarities.**

***Keywords*—chatGPT; feature vectors; similarity calculation**

I. Introduction

Rightfully, the performance of recent AI-based text generating tools have raised concern in the education community. Teachers fear that these tools will reduce the validity of student examinations since the quality of the generated text is sufficient to pass an exam without additional intellectual efforts by the student.

Several attempts to develop methods to identify AI-generated text are currently being made [1]–[3]. The contribution of this paper is to employ the well-known *feature vector* document representation and to create similar representation of document collections, sometimes called *centroids* [4]. More details on the algorithms used will be presented in Section II.

The research question of this paper is: *Can a feature vector representation of a document collection be used to identify the AI-generated documents in the collection?*

The AI-based text generation service invoked for the purpose of this paper is *chatGPT 3.5* [5], which has attracted a lot of attention due to its versatility. The chatGPT service not only generates non-fictional text, but also poems, fiction, propaganda, and even programming code.

The so-called *Vector Space Model* [4] employed in this experiment has proven successful to distinguish documents on the semantic content, i.e., their topic. In this experiment, the difference in *writing style* is probably the best distinguishing property. It is not clear, however, how the Vector Space Model is suitable for that task.

The remainder of this manuscript is organized as follows: in Section II, we describe the experiment design and the results found. Section III provides a brief survey of related work over the last few years. In Section IV, a discussion on the results compared to other related work is presented. The paper finally provides some conclusive remarks and suggests further research in Section V.

II. Experiment design

In this section, the design of the experiment, as well as the algorithm used for investigation, will be presented.

A. Training and evaluation set

The chosen approach involves a *supervised learning* algorithm, where a *training set* of documents is used to calibrate the testing algorithm, and an *evaluation set* which is used to determine its efficacy. Both these sets are relatively small, since the documents need to be generated one by one through queries sent to text generating services. The documents contain a mix of text generated by chatGPT and text found on the Web as identified by the Google search engine.

In order to generate text related to a range of topics with different vocabulary and writing style, three distinct questions were formulated:

- Why should abortion be illegal?
- Describe the tactical advantages of the F-35 fighter airplane.
- Describe the poetry of William Blake.

Each question was entered into the Google search engine, and the 10 topmost results were retrieved and saved to disk. 5 of these went into the training set, and 5 into the evaluation set.

Likewise, the questions were entered into chatGPT, but in order to obtain a collection of related documents, four additional questions were asked for each, e.g., for the second question:

- Why is the F-35 lightning so expensive?
- Is the F-35 a safe airplane?
- How many countries have bought the F-35 fighter jet?
- Which fighter jets can compare to the F-35?

The response for the four additional question were collected into another training set for chatGPT output.

For each of the three main questions listed above, we obtained in this way a training set for Google search engine output consisting of 5 documents, and a training set for chatGPT output in the form of 4 short documents. The evaluation set for chatGPT output was one single document and 5 documents gathered through the Google search engine.

B. Feature vector representation

Text documents may be represented by a feature vector, where each vector element represents a specific term/feature in the document, and the element value is, e.g., the frequency of a specific term, often the count of the specific word. The feature vector represents a point in a high dimensional space, and the relation between such points has been shown to effectively estimate the semantic relation between the represented documents.

 ISBN: 978-1-68558-045-2

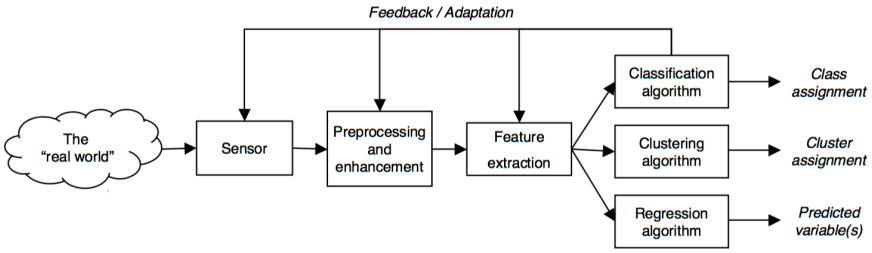

Figure 1. The Feature Vector construction process, and the role of the stemming and stopword processing.

To reduce the dimensionality of the feature vector, the number of included terms are reduced by the use of a *stopword list*. A stopword list contains words commonly used in text of any category and topic, and is considered to be insignificant for the construction of a feature vector. For the experiment conducted for this manuscript, the stopword list has 250 entries.

For the same reason, different grammatical forms of a term (plural, singular, present, past etc.) are merged into the same spelling through a *stemming* process. The most prominent implementation of this task is the *Porter stemmer* [6], which is used in this experiment. The construction process for a feature vector is shown in Figure 1.

Even though the feature vector is constructed using statistics on a lexical level where sentence structure are disregarded, they have been shown to adequately represent also semantic properties of a document, and feature vectors with short distance or angle in the high dimensional space most often represent documents with semantically related content.

Document collections can be represented in the same way by constructing a feature vector over the concatenation of the member documents. A feature vector representing a collection is also known as a *centroid* [4], indicating that it represents the "centre of gravity" of the document collection.

A centroid constructed over a training set is useful for the identification of documents belonging to the same "semantic class" as the training set. The cosine of the angle between the feature vector of the document and the centroid indicates the document's similarity to the training set.

C. Operations on feature vectors

Both the distance between points in high dimensional space and the angle between the corresponding vectors can be used to measure document similarities. The angle between vectors is not dependent on the vector length, and the cosine of the angle will have a convenient range of values, 1 for totally similar documents, and 0 for "orthogonal" (dissimilar) documents.

The cosine of the angle between two vectors $\vec{a}$ and $\vec{b}$ can be calculated with the following formula:

$$\cos(\vec{a},\vec{b}) = \frac{\vec{a}\times\vec{b}}{\|\vec{a}\|\|\vec{b}\|} = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}} \qquad (1)$$

TABLE I
COSINE OF ANGLES BETWEEN CENTROID VECTORS FOR EACH QUESTION

Q nr.	Cos(angle)
1	0.628
2	0.697
3	0.668
All	0.648

TABLE II
COSINE OF ANGLES BETWEEN CENTROID VECTORS FROM DIFFERENT TRAINING SETS

Q nr.	Cos(angle)
1 and 2	0.101
2 and 3	0.119
1 and 3	0.091

D. Similarity calculations

As already mentioned, the goal of the feature vector representations of training sets and evaluation samples, and the calculation of angles between them, is to estimate semantic relations between evaluation samples and training sets. This approach has proven successful for document *classification*, where he centroids represent "class prototypes" and documents are assigned the class associated with the centroid with the smallest angle from its feature vector. The angle between the centroids themselves indicate the confidence level of the classification process, as well as the measured difference between the candidate classes.

In the following presentation of the calculation results, one graph for each question is shown where the documents in the evaluation set are represented as dots in a cartesian plane. The axes represent the centroids from the training sets, and dots on the diagonal line have the same similarity between the two training sets. The blue circles show the documents retrieved through the Google search engine, while the red square shows the document retrieved from chatGPT. The graphs are shown Figures 2-4.

While the cartesian projection suggests that the two centroids constructed from the training sets are orthogonal, this is not the case. The cosine of the angles between the two centroid vectors for the three main test questions are shown in Table I and indicate that the centroids share some similarities, which reduces the confidence in the identification process.

 ISBN: 978-1-68558-045-2

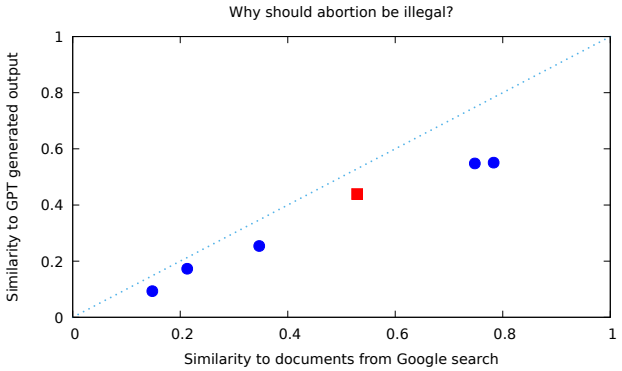

Figure 2. Detection performance of question 1

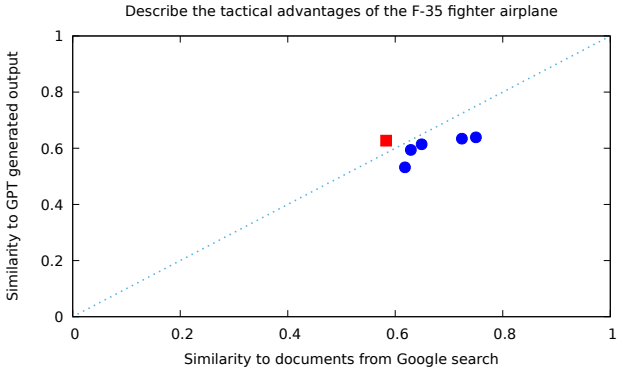

Figure 3. Detection performance of question 2

III. Related work

The interest in discriminating between machine-generated and human-written text dates back before 2014. Concerns for fake news, biased reviews, and presently, concern for examination validity, has created a certain volume of reports. Some of these will be mentioned here.

The most recent addition to AI-based text generating services is chatGPT, where the service named GPTZero [7] offers web access to a discriminator with reportedly high precision.

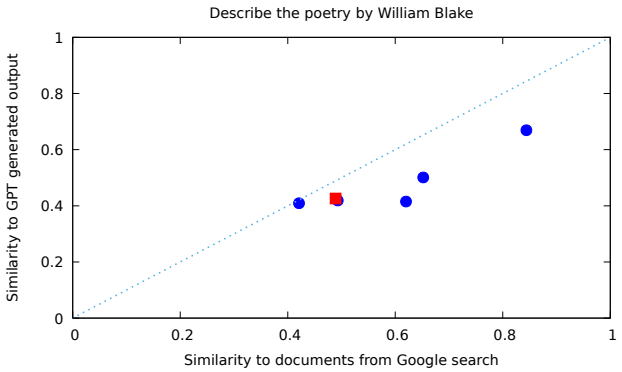

Figure 4. Detection performance of question 3

No peer-reviewed report on this service has been found, but a good technical report in [8] explains the design during an interview with its creator, Edward Tian. In the article, he explains that the tool was designed over a weekend and is far less complex than chatGPT. GPTZero calculate properties called *perplexity* and *burstiness*, whose values represent the irregularities in sentence structure which indicate text written by humans (so far).

No detection tools using AI has been found, and the discriminator designs use supervised or unsupervised learning of machine-learning (e.g., neural networks) or feature vector based algorithms.

Shijaku and Canhasi [9] build feature vectors with a different feature selection strategy than the presented experiment, but with familiar operations on the vectors. Their training set consists of student-written texts, and their experimental reports are promising.

Gallé et al. [1] argues that supervised learning is an unrealistic protection mechanism, since an adversary has access to both the training set and the algorithm and can formulate his/her text accordingly. They propose unsupervised learning based on N-gram distribution. ChatGPT output is not evaluated in [1] since the experiment predates chatGPT.

Gehrmann et al. [2] design their discriminator as a visual helping tool for human detection of machine-generated text. Their algorithm uses statistical properties of single words to present the text in a colorized form in which the human detector can recognize certain patterns that suggest the nature of the text (machine or human created). Their experimental evaluation was done with a panel of humans who were asked to identify the nature of documents with/without visual hints.

It is worth noting that the dating of these efforts spans over three years, and the technology for machine-generating text is evolving at a rapid pace. They are evaluated on different text sources and their results must be compared with some reservation.

IV. Discussion

Under certain circumstances, there may be a large volume of text which needs to be analyzed for machine generation in real time. None of the reported efforts presented in Section III consider *computational demands or scalability properties*, which we believe will be a future requirement. The effort shown in the presented paper have chosen a feature selection strategy which is fast and proved to identify semantic similarities with good precision.

The results shown in Figures 2-4 indicates that the use of feature vectors with word frequencies as the only feature is not suited for detection of text generated by chatGPT. There is no straight line in the graphs that can separate the blue dots from the red. We should not disregard the chosen algorithm without pointing at other factors that can affect the final results.

A. Similar training sets

The chatGPT is trained on the same document body as the training set used in this paper, i.e., documents found on the

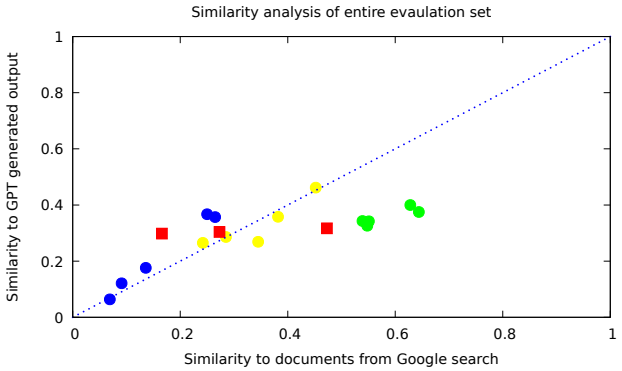

Figure 5. Detection performance for entire evaluation set

Internet. The documents selected for the training set were the topmost results from a Google search engine query, likely to be written by journalists and technical writers in a conformant style with few personal traits. Table I also shows that the two training sets for each question is indeed quite similar. Using, e.g., documents written by students as a training set, the similarities may be smaller, but a document body with these properties were not available for use in this experiment. For the record, Table II shows the similarities between training sets made from output from Google search engine related to the three test questions, which are nearly orthogonal.

B. Feature selection

The efforts presented in [1] argue that N-gram analysis yields better discrimination properties than the use of single words. An N-gram analysis will also significantly increase the dimensionality of the feature vectors and the computational workload necessary for their processing.

C. Volume of training and evaluation sets

In this early phase of investigation, a limited number of documents were used for training and evaluation, and output from chatGPT also tends to be relatively brief. Common knowledge in this area is that a larger body of text for training and evaluation provides better results and a smaller confidence interval. Future phases of investigation will spend more time on the establishment of larger text bodies.

D. Use of distinct topics

The presented experiment has compared documents within the same topic, implicitly introduced through the three test questions. Related work has not made this distinction, but used document collections not related to a specific topic.

In order to compare our results to other works, e.g., those presented in [9], the training set for the three questions were joined to make a new pair of centroids representing all training documents from the two sources. A calculation of the similarities of the evaluation documents is presented in Figure 5. The cosine of the angle between the two centroids is shown on the bottom row of Table II. The results are quite similar to what is presented in Figures 2-4, where no straight line can separate documents from the two sources (red vs. blue/green/yellow). Still the documents belonging to the test questions can be found as clusters along the diagonals, indicating their semantic relations, shown in blue, green and yellow, respectively.

V. Conclusion

. The presented efforts aimed to distinguish machine-generated from human-written text based on the well proven vector space model otherwise used for text classification and information retrieval. The results were not encouraging, indicating that the two text categories exhibit nearly the same lexical properties for each of the test questions given.

Possible reasons for the disappointing results are analyzed in Section IV, where the generation of the training and evaluation sets are likely candidates for future investigations. A larger body of text for training and evaluation, written by a diverse selection of human writers may result in centroids more distinct than what was obtained in the chosen experiment design, and better spreading of results in the Figures 2-4. Also, similar tests using Bing Chat and Google Bard (still not available in Norway in May 2023) are also part of future research plans.

AI-technology for text generation is evolving fast, and will generate text increasingly difficult to distinguish from human -written text, if the designers wish to. Reports mentioned in Section III dates as far back as 2019, and may not produce the same results on chatGPT-generated output. The work presented in [9] is more recent and was evaluated on chatGPT-generated output, and provides a basis for comparisons with the presented experiment, although their training set is quite different.

References

[1] M. Gallé, J. Rozen, G. Kruszewski, and H. Elsahar, "Unsupervised and distributional detection of machine-generated text," https://arxiv.org/abs/2111.02878, 2021, [Online; accessed 16-May-2023].

[2] S. Gehrmann, H. Strobelt, and A. M. Rush, "Gltr: Statistical detection and visualization of generated text," https://arxiv.org/abs/1906.04043, 2019, [Online; accessed 30-Jan-2023].

[3] D. Ippolito, D. Duckworth, C. Callison-Burch, and D. Eck, "Automatic detection of generated text is easiest when humans are fooled," https://arxiv.org/abs/1911.00650, 2020, [Online; accessed 30-Jan-2023].

[4] G. Salton, A. Wong, and C.-S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[5] openai.com, "chatGPT Terms of Use," https://openai.com/policies/terms-of-use, [Online; accessed 16-May-2023].

[6] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.

[7] E. Tian, "GPTZero," https://gptzero.me/, [Online; accessed 16-May-2023].

[8] E. Ofgang, "What is GPTZero? The ChatGPT Detection Tool Explained By Its Creator," *Tech & Learning*, Jan 2023, [Accessed 16-May-2023]. [Online]. Available: https://www.techlearning.com/news/what-is-gptzero-the-chatgpt-detection-tool-explained

[9] R. Shijaku and E. Canhasi, "ChatGPT Generated Text Detection," 10.13140/RG.2.2.21317.52960, Jan 2023, [Accessed 16-May-2023].

 ISBN: 978-1-68558-045-2