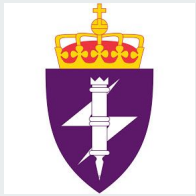
# Scalable Detection of chatGPT-generated text

**dr. Anders Fongen,** June 2023
*Norwegian Defence University College, Cyber Defence Academy, Lillehammer*
email: anders@fongen.no

FUTURE COMPUTING 2023, Saint-Laurent-du-Var, France



# Presenteres bio

## Anders Fongen

- Associate Professor, Norwegian Defence University College
- Field of research: Distributed Systems, Networking security
- PhD in Distributed Systems, Univ. of Sunderland, UK, 2004
- Career history
  - 7 years in military engineering education (Associate Professor)
  - 10 years in defence research (Chief Scientist)
  - 8 years in civilian college (Associate Professor)
  - 11 years in oil industry
  - 6 years in electronics industry

# Objective of this research effort

AI-generated text mistaken as human generated text represents serious problems:

- During formal assessment of a person's theoretical competence (exams)
- When establishing originality or intellectual property rights
- When trust in the author's impartiality and integrity is essential

Therefore, a process to distinguish human and AI authorship is of great importance

REMEMBER: Work-in-progress on a moving target!

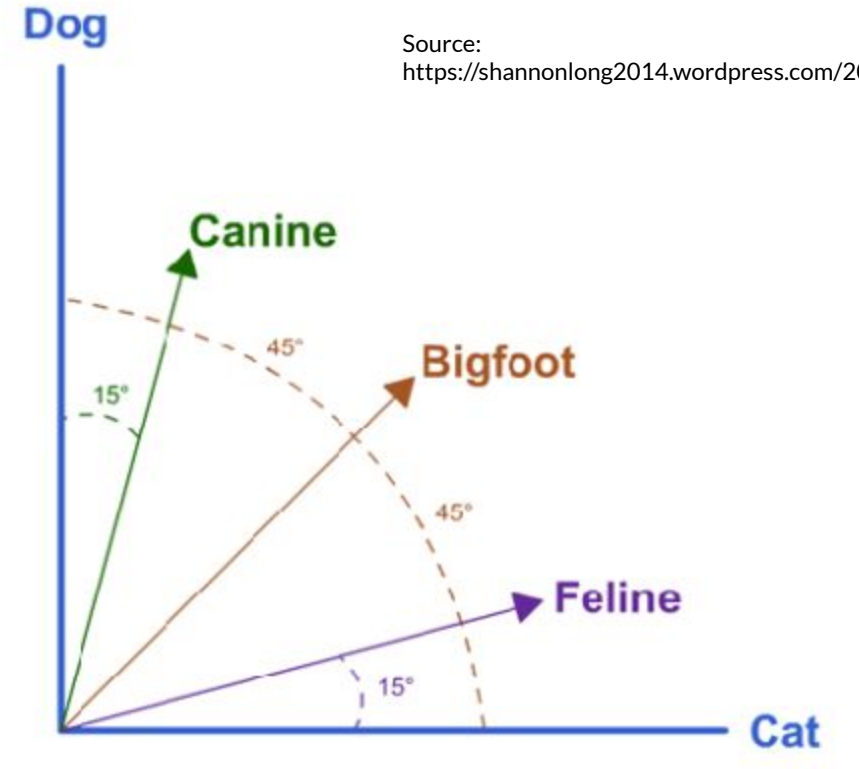
# Ways to distinguish text

- Language based analysis (syntax-oriented)
- Lexical analysis (statistics-oriented)
  - Simple, fast and scalable algorithms
  - Is shown to successfully reveal semantic relations between texts
  - Well suited for **text classification**

- **Vector space model** is employed for analysis
  - A text object is represented by a high dimensional *term vector*.
  - A training collection of related text objects is also represented by a term vector
  - The semantic relation between two text objects (or with the training collection) is estimated by the *cosine* between their term vectors (0..1)

# Simplistic Term Vector Model

(only two topics exist - "dog" and "cat" - and all words are measured by their relationship to these two words)




Source: https://shannonlong2014.wordpress.com/2014/06/01/ir-models-vector-space-model/

# Experimental design

1. Collect a number of documents on related topics from
    a. chat PT (centroid-A)
    b. from the top 5 found by a google search (centroid-B)
2. Process them to create a term vector from each
    a. the term vectors will not “represent” chat GPT/humans on that topic

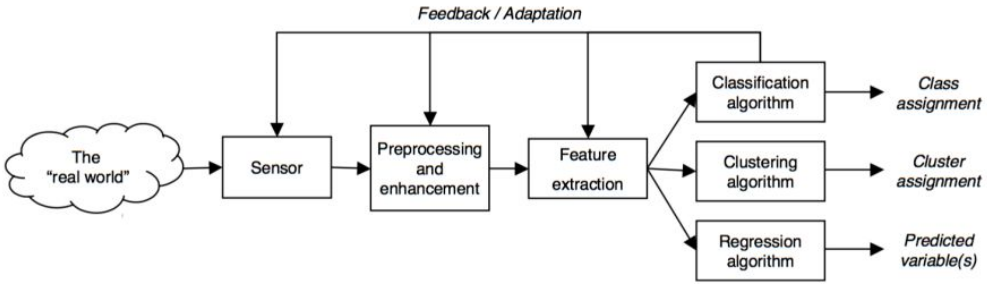
Used topics: “Why should abortion be illegal”,
“Describe the tactical advantages of the F-35 fighter airplane”
“Describe the poetry of William Blake”

Topics were chosen for the sake of *diversity* in content and writing style

# Term vector processing

Each element in the vector represent a language term, and the value of that element is the frequency/count of that term in the document (-collection). Also applied:

- Stopword removal
- Stemming



# Document classification as human/chat PT

1. Create a term vector of the document in question (doc)
2. Calculate cosine(doc,centroid-A) and cosine(doc,centroid-B)
   a. use these values as (x,y) in a cartesian plot
3. How far away from the diagonal line in the diagram? *(Not very far, regrettably)*

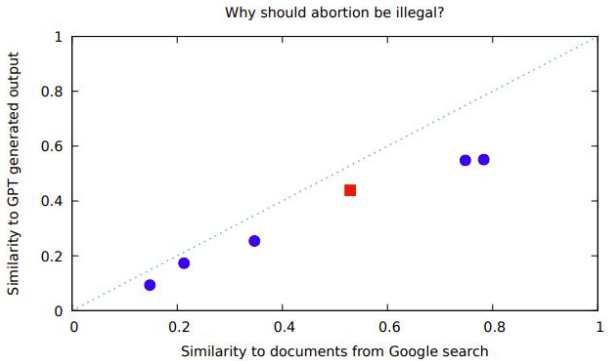
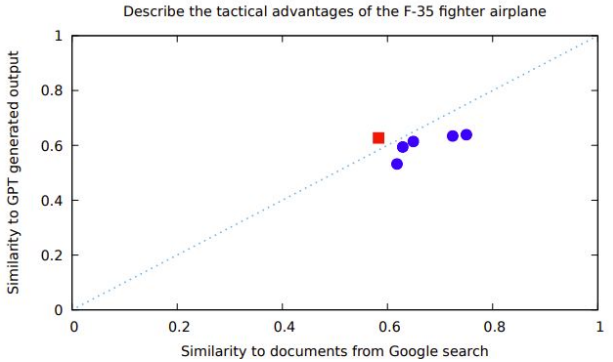



Figure 2. Detection performance of question 1



# Classification “cross-topic”

Combining all training text into a single pair of term vectors, gave these results:

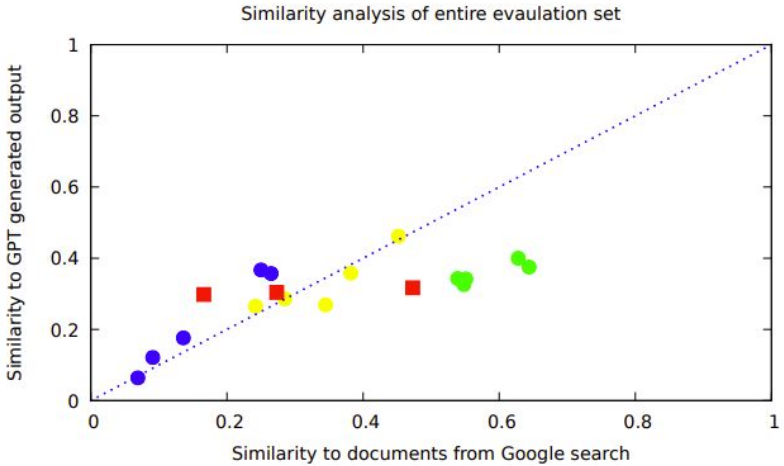



Figure 5. Detection performance for entire evaluation set

*No straight line can separate the red squares from the rest.*

# Possible reasons for poor classification results

1. Small volume of training documents
2. Poor selection of internet text
    a. Mostly written by professional writers (assumed)
    b. chat GPT is trained on the same document collection
    c. English text written by Norwegian students was not available
3. The chosen algorithm (counting 1-grams) not useful
    a. Although this allows fast processing and good scalability

*Further investigation needed, while AI is improving their sophistication*

# Conclusion

The problem: **How to distinguish text originated by AI or by humans?**

- 1-gram analysis with limited training collection is *not successful*
- Larger training set unlikely to make a difference
- Different training set (not used for training AI) may do
  - e.g., assignment deliverables written by students
- Deeper analysis (n-gram, language model) has been shown to give better results
- Again: AI is under development and their sophistication is improving

Thank you for your attention, any questions?