Cyber Guard AI

Categorization of cyber-complaints into various predefined categories and subcategories.

The Idea

Vector embeddings are the most interesting product/concept that came to true life in the recent AI boom. A vector embedding, an array of a large number of real numbers, represents the position of the given text in a vector space. The dimension of this vector space is usually large, like 768 or 3072, i.e. an array of 768 or 3072 real numbers. The main selling point of this vector array is that it represents the position of the given text in the vector space contextually, semantically, and in terms of meaning. That means a similar text will also reside around the same position in the space.

This enables vector embeddings to be very useful for search, recommendations, classification, etc.

The idea of vector embeddings is old but the recent development in the field of AI particularly Deep Neural Networks has enabled vector embedding models to extract the meaning and context of the given text.

So, our method of solving text classification of user written complaints will extensively use the new age DNN vector embedding models.

The Method

We have written two methods to try to solve the challenge.

1. Embedding the text (user complaint) directly and indexing it in a vector database collection together with the category and sub-category metadata.
2. Summarising a batch of texts (user complaints) from a similar category and sub-category using an LLM and then indexing it in a different vector database collection together with the category and sub-category metadata.

3. Calculating average embedding of every category and subcategory pair and using these embeddings for predicting category and subcategory. It is the fastest method. Together with Top-K sampling the predictions were found to be converging to correct answers after K=2.

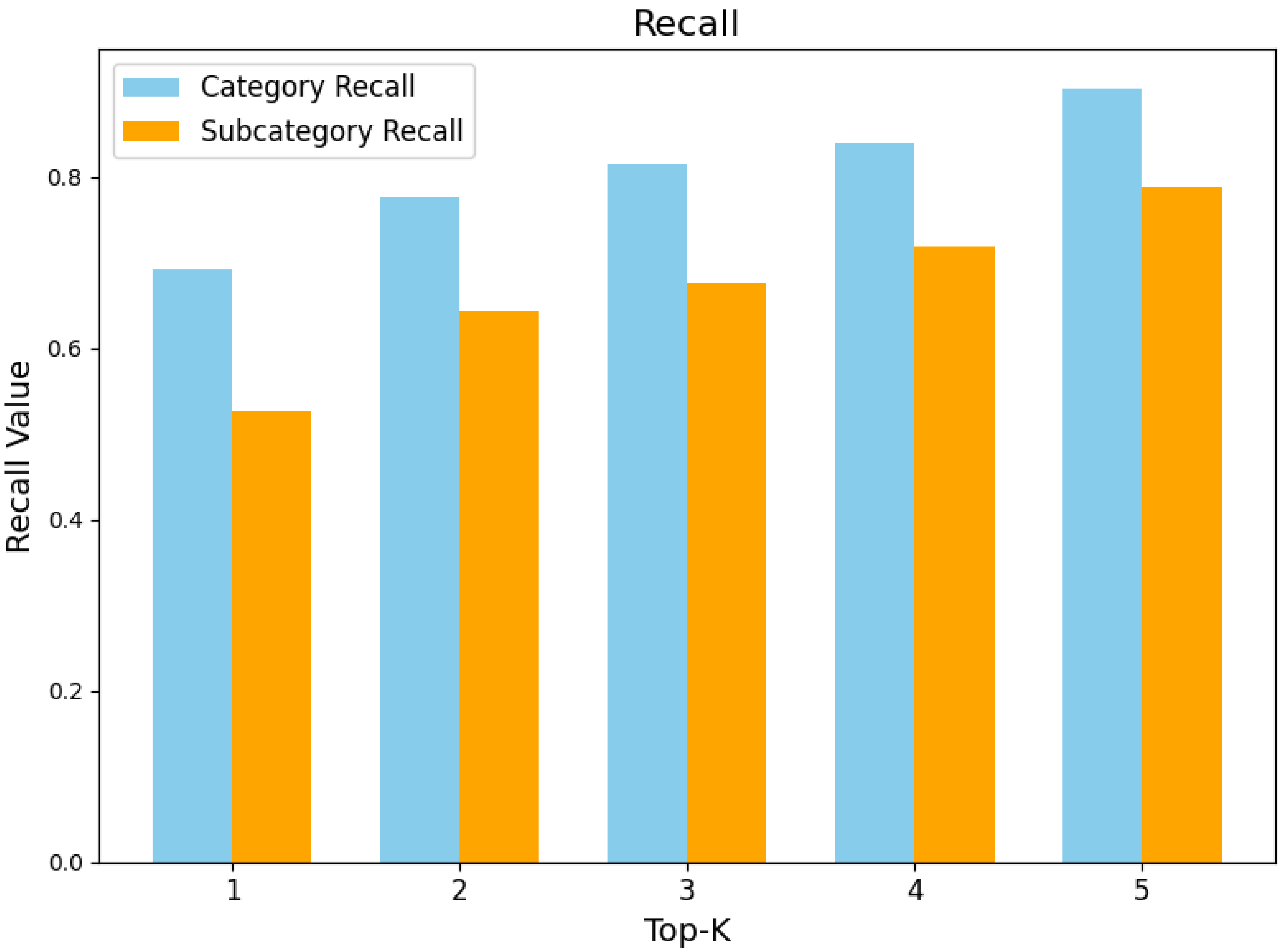
The next step is to perform a near vector search and retrieve the 10 nearest documents and use the metadata to guess which category and subcategory the input text (user complaint) might belong to.

Findings

It can be seen that in some cases when the predicted categories and subcategories do not match the expected values (when using test data), the predicted can either FIT or the predicted category and subcategories are a BETTER FIT than the expected values.
The time to predict the classification is always less than 400 milliseconds. In production, it can possibly reach to only 10s milliseconds.

Performance

Recall tests were performed using the average embeddings method for K = 1,2,3,4 and 5. The results were found to be "good". The minimum recall for category is ~0.7 with K = 1, then both the category and subcategory recall shows a steady increase with an increase in K.



Further Development

Using DNN vector embedding models has enabled a huge future development prospect. The following steps can be taken in the future for better results.

1. Fine tuning available embedding models like nvidia/NV-Embed-v2 and jinaai/jina-embeddings-v3.
2. More detailed summarization with manual human intervention.
3. Transforming the user complaints using prompt engineering methods for better feature extraction.