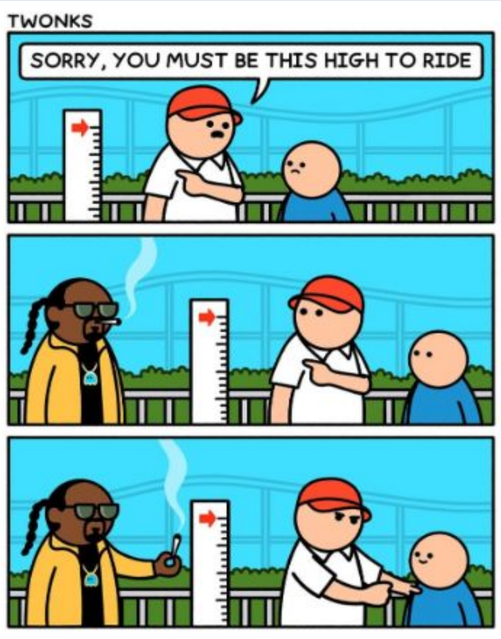
CERIAS

The Center for Education and Research in Information Assurance and Security

Text embeddings for efficient search in digital investigations

Koeken, Adil; Nandwana, Samay; Shah, Avantika; Lima, Vinicius; Karabiyik, Umit.

Overview



With the increasing storage capacities of modern devices, digital investigators face significant challenges in efficiently locating relevant information on incriminating devices. It is not uncommon for a single smartphone to contain over 100GB of data, making manual searches impractical. The problem becomes even more complex when the needed information is not easily discoverable through keyword searches or when investigators do not know which keywords to use. This makes digital investigations cumbersome and time-consuming, highlighting the need for more efficient search methods. For text-based data, one promising solution is the use of text embeddings (or vectors) instead of traditional keyword searches. In this study, we simulated a database containing text related to drug-related discussions, where specific drug names were not explicitly mentioned. This reflects real-world scenarios, such as conversations among drug dealers who use coded language (slangs) instead of technical terms. Our research demonstrates how embeddings can be leveraged to identify drug-related content without relying on predefined drug keywords. We applied the LLM2Vec technique using different models and instruction formats, achieving strong accuracy metrics. Our findings serve as a proof of concept that text embeddings can significantly enhance search efficiency in digital investigations compared to conventional keyword-based methods.

Data Acquisition

To detect drug slang on social media, we required a dataset with sample drug slangs to be embedded. Since no comprehensive dataset existed, we built our own using a hybrid approach:

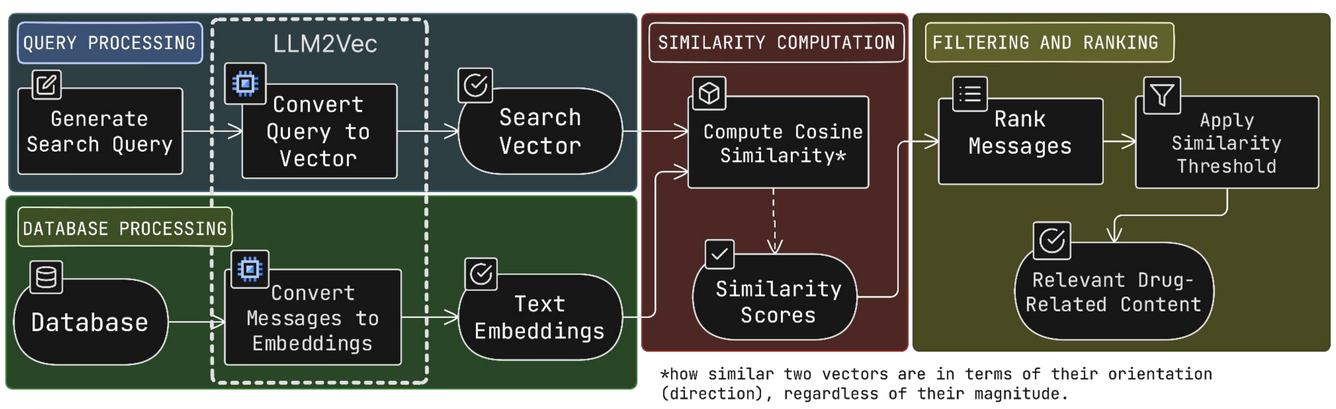
LLM-Assisted Generation: We prompted CLAUDE and GPT4o to simulate realistic tweets containing a variety of drug slang terms, based on public domain slang databases. Since many slang terms may be uncommon, there might not be existing data for a lot of the sample slang terms.

Real-World Data from scraping tweets: We augmented the synthetic LLM generated data by using a customizable actor from APIFY. This allowed us to use various parameters to filter down tweets and get real live data that used in slangs

This combined methodology ensured that we had both authentic social media language and sufficient representation of drug-related slang for robust evaluation.

After obtaining all of our data through the sources, we categorized into NON_DRUG, IMPLICIT_DRUG, and EXPLICIT_DRUG, to semantically label the sentences.

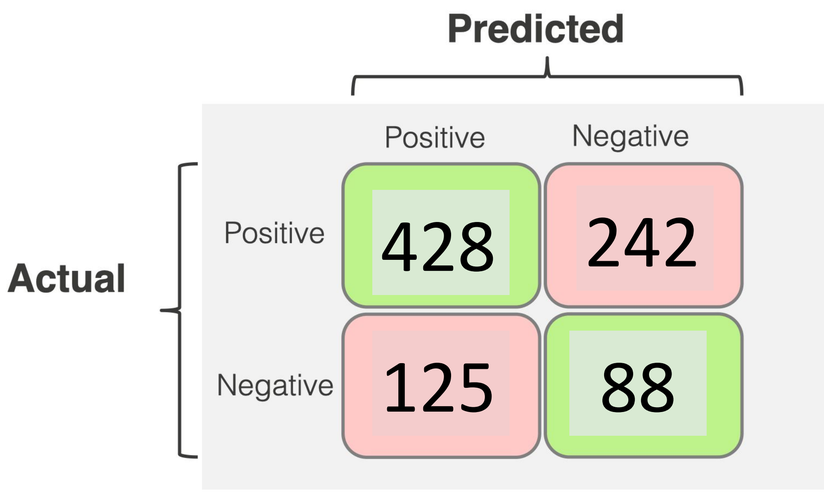
Methodology



*how similar two vectors are in terms of their orientation (direction), regardless of their magnitude.

Results

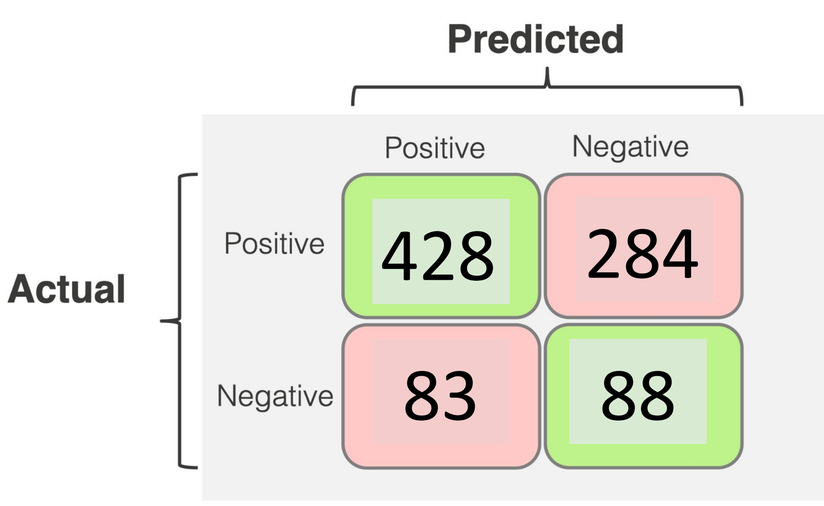
Without instruction



Accuracy: 75,8%

F1 Score: 80,0%

With instruction



Accuracy: 80,6%

F1 Score: 83,3%

Examples

Claude: "the slopes were amazing this weekend, cant wait to go skiing again with the crew 👃 " IMPLICIT_DRUG

GPT4o: "A quick toot never hurt nobody, right?" IMPLICIT_DRUG

Twitter: "Dr. Barber discusses the challenges and strategies of weed science, including resistance manageme..." NON-DRUG

PURDUE UNIVERSITY

CERIAS