BUSINESS INTELLIGENCE PROJECT

Presented by:
Andrea D'Amicis 869008
Gabriele Sormani 869217

OVERVIEW

01	02	03
Dataset	Preprocessing	ChromaDB
04	05	06
Query	Classification	Explainability

DATASET

Sentiment Analysis for Financial News (FinancialPhraseBank).

- Context: Sentiments of financial news headlines from a retail investor's perspective.
- **4846** Sentences: Financial news headlines in English.
- Application: Suitable for sentiment analysis and NLP in financial domains.

13%
Negative

59%
Neutral

28%
Positive

kaggle

Financial News

Sentiment Analysis for Financial News

Dataset contains two columns, Sentiment and News Headline

k kaggle.com

PREPROCESSING

01 **POS Mapping**: Converts POS tags to WordNet-compatible formats for better lemmatization.

02 **Text Cleaning**: Removes non-alphabetic characters and converts text to lowercase.

03 **Stopword Removal**: Filters out common words to retain meaningful terms.

04 **Tokenization & Lemmatization**: Breaks text into words and reduces them to their base forms.

05 **Output**: Produces cleaned, standardized text ready for NLP or machine learning tasks.

EMBEDDING

Model:

- **DistilRoBERTa financial sentiment** : A **lighter**, and **faster** version of RoBERTa, retaining most of its performance.
  - 82M parameters (compared to 125M parameters for RoBERTa-base)
  - Accuracy: 0.9823
- **Fine-tuned**: Fined tuned on the financial_phrasebank dataset and specifically adapted for sentiment classification in financial news (positive, neutral, negative).

Key Features:

- **SentenceTransformer**: Generates contextual sentence embeddings providing fixed-size vectors capturing sentence meaning (768 dimensions).
- **Domain-Specific Tuning**: Trained on financial news data for sentiment detection in financial contexts.

Benefits:

- **Accuracy**: Detects financial-specific sentiments effectively.
- **Efficiency**: Lightweight, fast, and scalable.anding.

CHROMA DB

ChromaDB is a vector database designed for storing and managing large-scale embeddings. It allows efficient retrieval and querying of text data along with their associated vector representations for tasks like similarity search.

Client Initialization

Persistent client was setup with a specified directory for storage on cloud. The Collection was created to perform database population.

Text & Metadata

The raw text and associated metadata of sentiment and cleaned text are added into collection.

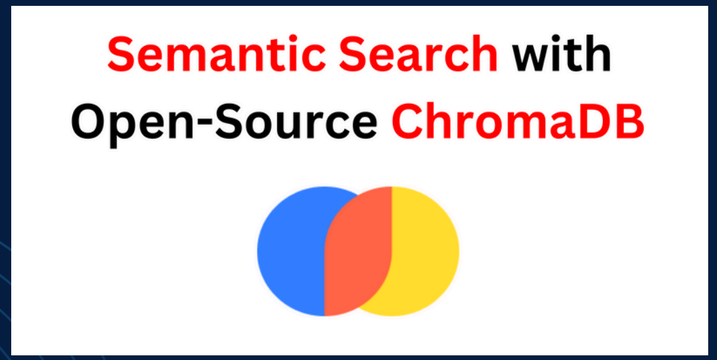
Embeddings

Related sentence vector embeddings are stored in the collection for each text entry in the appropriate field.

CHROMA DB QUERIES

The following 7 natural language queries were formulated and, for each of them, the 3 most relevant documents were obtained with their distance measure and associated sentiment:

1. "Silicon Valley's impact on global technology development"
2. "Positive solutions in renewable energy"
3. "Negative controversies in the automotive industry"
4. "Political tensions in Asia-Pacific region"
5. "Positive news in the stock market"
6. "Issues regarding cybersecurity threats in the financial sector"
7. "Celebrity endorsements positively influence brand equity"



output

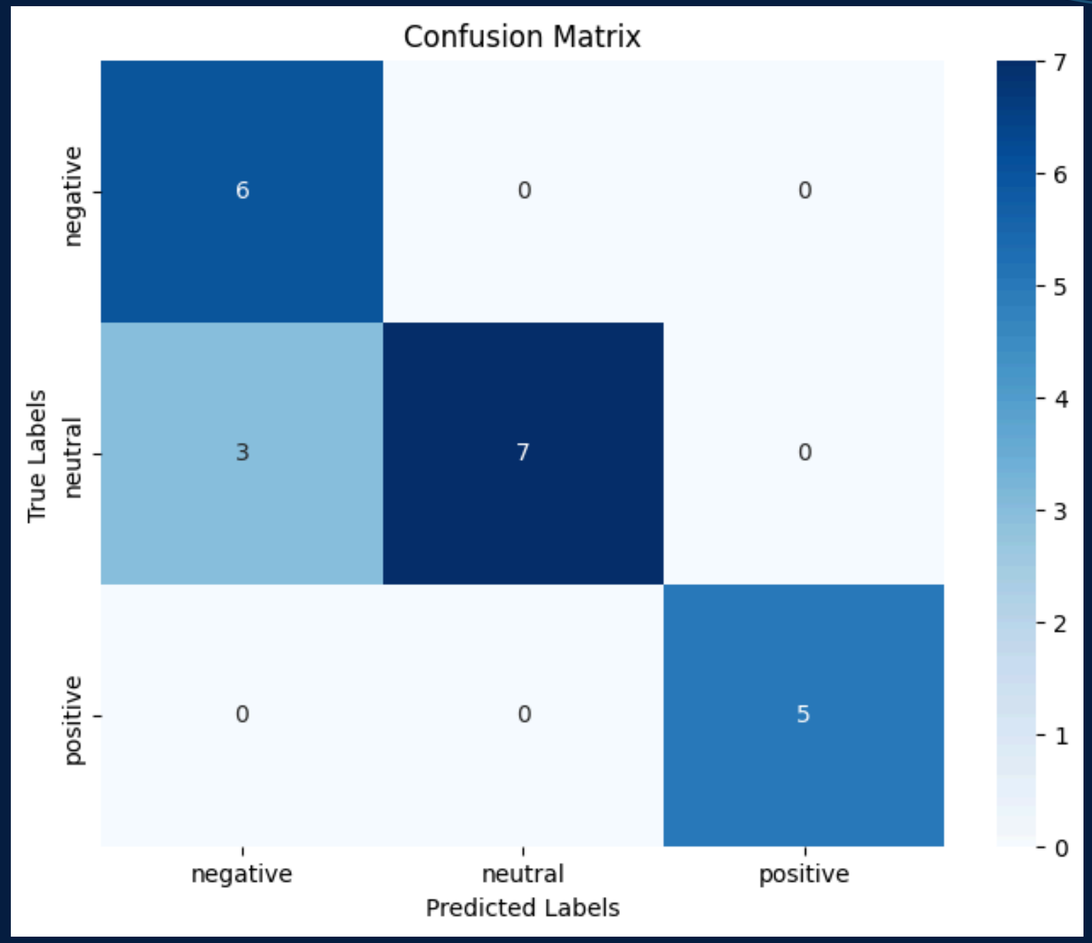
```
Query: Positive news in the stock market
--------------------------------------------------

Document 1:
Text: Technical indicators for the stock are bullish and S&P gives NOK a positive 4 STARS out of 5 buy ranking .
Sentiment: positive
Cleaned Text: technical indicator stock bullish p give nok positive star buy rank
Distance: 81.5416
--------------------------------------------------

Document 2:
Text: The price of raw material aluminium went up at the end of 2005 , but the company considers its outlook for 2006 favourable .
Sentiment: positive
Cleaned Text: price raw material aluminium go end company considers outlook favourable
Distance: 86.7795
--------------------------------------------------

Document 3:
Text: Cash flow from operating activities is estimated to be positive .
Sentiment: positive
Cleaned Text: cash flow operating activity estimate positive
Distance: 88.1726
```

SENTIMENT CLASSIFICATION



Then, for each document obtained from the queries, we classified the sentiment in real time using the fine-tuned DistilRoberta-financial-sentiment previously described. Finally, we evaluated the performance of the model based on the results obtained with confusion matrix and classifcation report.

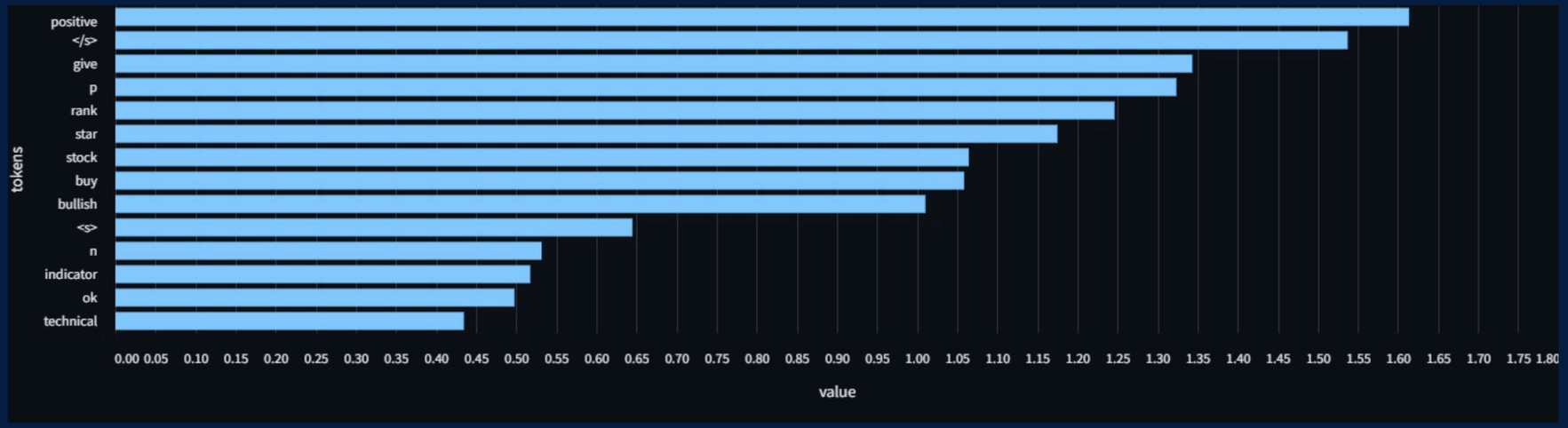
Classification Report

	precision	recall	f1-score	support
positive	0.67	1.00	0.80	6
neutral	1.00	0.70	0.82	10
negative	1.00	1.00	1.00	5
accuracy			0.86	21

https://github.com/cdpierse/transformers-interpret

EXPLAINABILITY

IA barplot was created to visualize which tokens the model focused on the most during the analysis of each sentence based on attention values.



Using the transformer explainer, it is possible to visualize which words have the greatest impact on sentiment prediction, providing an interpretation of the model's decision-making process by seeing which ones influence the predicted sentiment class.

Legend: Negative impact Neutral impact Positive impact

Prediction Score	Attribution Label	Attribution Score	Word Importance
(0.08)	negative	1.15	#s technical indicator stock bullish p give n ok positive star buy rank #/s
(0.05)	neutral	-2.21	#s technical indicator stock bullish p give n ok positive star buy rank #/s
(1.00)	positive	2.31	#s technical indicator stock bullish p give n ok positive star buy rank #/s

Thank's For Watching

Now let’s see the final dashboard!