Project Deliverable D1.2

Sentiment Analysis for Financial News

Team Information

Team Name: Finance bros
Members:

- **Makeev Roman** - r.makeev@innopolis.university
- **Martyshov Maxim** - m.martyshov@innopolis.university
- **Smirnov Elisey** - el.smirnov@innopolis.university

Repository Link

https://github.com/kezouke/Sentiment_Analysis_for_Financial_News

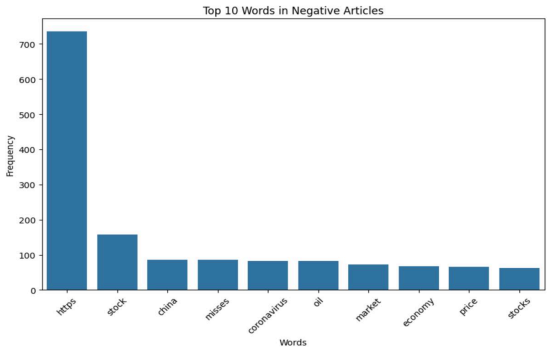
Progress Overview

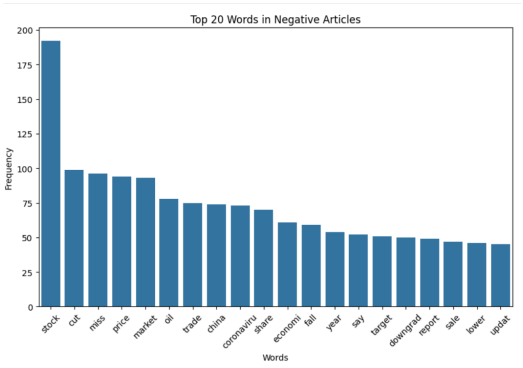
1. **Data Versioning with DVC**
   We successfully implemented **DVC** to manage data versioning and storage. This setup allows us to store different versions of our dataset efficiently. Run src/scripts/dvc_init.sh for initializing DVC.

2. **Initial Data Validation with Great Expectations**
   We integrated **Great Expectations** for data validation to ensure the quality and consistency of our dataset. We created checkpoints that validate our data against pre-defined expectations. You can see all data requirements in notebooks/initial_expectations.ipynb

```
Validation success: True
Expectation expect_column_to_exist: SUCCESS
Expectation expect_column_values_to_not_be_null: SUCCESS
Expectation expect_column_values_to_be_of_type: SUCCESS
Expectation expect_column_value_lengths_to_be_between: SUCCESS
Expectation expect_column_values_to_be_unique: SUCCESS
Expectation expect_column_to_exist: SUCCESS
Expectation expect_column_values_to_not_be_null: SUCCESS
Expectation expect_column_values_to_be_of_type: SUCCESS
Expectation expect_column_values_to_be_in_set: SUCCESS
Expectation expect_table_row_count_to_be_between: SUCCESS
Expectation expect_table_columns_to_match_ordered_list: SUCCESS
```

3. **Feature Engineering**
   We performed **feature engineering** by preprocessing the text data. This included tokenization, stop word removal, and stemming. All steps are described in [notebooks/feature_engineering.ipynb](). These preprocessing steps were crucial in preparing the dataset for training. For example we were able to collect and select the most appropriate words for our model. On the following graphs Top 10 words in Negative Articles before and after preprocessing





4. **Model Training and Experimentation**
   We trained a model using **BERT** from the *transformers* library on our dataset. After 10 epochs, we achieved an accuracy of **82.8%**. Below are the key training metrics for epoch 9. All other details in [notebooks/model_experiments.ipynb]()

```
Epoch 9: train: 100%|████████| 106/106 [00:42<00:00,  2.51it/s, loss=0.0219]
Epoch 9: val: 100%|████████| 52/52 [00:07<00:00,  7.34it/s, loss=0.836, acc=0.828]
```

Work Distribution:

Roman Makeev - **Model Training and Experimentation**
Elisey Smirnov - **Initial Data Validation with Great Expectation**
Maxim Martyshov - **Data Versioning with DVC, Feature Engineering**

Plan for the Next Weeks

1. **Hyperparameter Tuning**
   We will fine-tune our model by experimenting with different hyperparameters (learning rate, batch size, etc.) to further improve accuracy and generalization.
2. **Model Evaluation and Validation**
   We plan to implement cross-validation and assess our model performance on additional metrics such as F1 score, precision, and recall.
3. **Model Deployment**
   Our next major milestone will be deploying the trained model through an API using **Flask** and Docker, allowing for real-time sentiment classification of financial news articles.