CRISP-DM FRAMEWORK

APPROACH TO THE SIX PHASES OF THE CRISP-DM FRAMEWORK

FABIOLLA MAYRINK & RAEL GUIMARÃES
CCT COLLEGE DUBLIN / MAY 2023

ABSTRACT

THE ADVANCEMENT OF TECHNOLOGY HAS GIVEN US VAST ACCESS TO INFORMATION, WHICH WE CAN ACCESS FROM ANYWHERE USING OUR PHONES OR COMPUTERS. THIS IS AN EXCELLENT TOOL FOR STAYING INFORMED ABOUT EVERYTHING HAPPENING AROUND THE WORLD AND IS ALSO USED BY COMPANIES TO MAKE STRATEGIC DECISIONS. HOWEVER, THE BIGGEST ISSUE IS THAT TECHNOLOGY IS BEING USED TO PROPAGATE FAKE NEWS, WHICH CAN MANIPULATE PEOPLE'S OPINIONS AND GREATLY IMPACT A COMPANY'S DECISION-MAKING PROCESS. TO ADDRESS THIS ISSUE, OUR FAKE NEWS DETECTION MODEL WILL HELP PEOPLE AND COMPANIES IDENTIFY WHETHER THE NEWS/INFORMATION IS REAL, ENABLING THEM TO MAKE BETTER DECISIONS BASED ON REAL NEWS AND AVOIDING BEING INFLUENCED BY FAKE NEWS. THE MODEL WILL USE NATURAL LANGUAGE PROCESSING (NLP) TO PREDICT WHETHER THE NEWS IS REAL OR FAKE BY ANALYZING PHRASE PATTERNS.

Business Understanding

The impact of fake news on society and business is significant, and a model to identify it is necessary. The model will use NLP to analyze text and classify it as real or fake. NLP combines computational linguistics rules with machine learning and deep learning models to find relationships between language constituents. The advantages of NLP are its ability to perform large-scale analysis, automate processes in real-time, and tailor the algorithm to meet industry-specific needs. Sharing articles without reading beyond the headline can cause even parody to be taken as truth, making it imperative to identify fake news.

Research Questions

1- Why is it important to use NLP techniques to develop a model for fake news detection?
2- What was the reason for using the oversampling technique in the dataset for the fake news detection model?
3- What is the deployment plan for the fake news detection model and who is the target audience?
4- What are the differences between the date ranges of the "True.csv" and "Fake.csv" datasets?
5- What was the process for choosing the best NLP model for implementation in fake news detection?

Data Preparation

The data was cleaned by removing duplicates and incomplete/malformed entries. The "True.csv" and "Fake.csv" datasets were merged and noisy data was removed. The date column was converted to datetime and plotted. The "text" and "label" columns were kept for model implementation, and the most used words were displayed for "real" and "fake" news. The data was imbalanced, so the oversample technique was used to balance it.

```
real_news_ds.duplicated().sum()
```

```
206
```

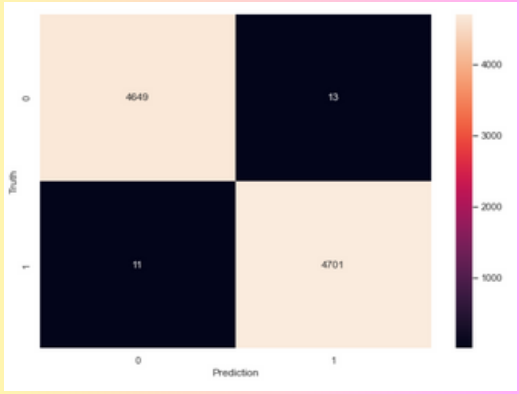
```
real_news_ds.drop_duplicates(inplace=True)
```

```
real_news_ds.duplicated().sum()
```

```
0
```

Deployment

The deployment plan for the fake news detection model includes creating a web application where users can input a news article or its URL and receive a prediction of whether it is true or fake. Clients will need to subscribe to use the service, and fees will vary based on the client's profile. The target audience includes companies and individuals who want to avoid being manipulated by false news. The plan is to start with a smaller-scale deployment and expand as demand increases. The team responsible for data collection and model maintenance will start small but expand if necessary. The plan includes constantly updating and training the model to avoid staleness.



Data Understanding

The "True.csv" dataset contains real news articles with four columns, and the "Fake.csv" dataset contains fake news articles with identical columns. The datasets were explored, and duplicates and malformed data were removed. The date range of the collected data for the "True.csv" dataset was approximately 1 year and 5 months, while the "Fake.csv" dataset's date range was approximately 2 years and 5 months.

Evaluation

The image shows that the Linear SVC evaluation results have shorter spread than the others. It means that its results are more concentrated, it does not spread much. It has a higher mean accuracy, which is above 99%. The Linear SVC (SVC) had its results slightly better than the others and we decided that it would be the model used for our implementation. Linear SVC has better results because it has better perform with unstructured and semi-unstructured data, such as text and image. It makes Linear SVC better than Logistic Regression on our context. Linear SVC has better results for linear and non-linear classification. It makes it better than Naive Bayes model.

```
train_accuracy: 99.971  train_precision: 99.957
test_accuracy: 99.669  test_precision: 99.618
```

Modelling

To choose the best NLP model for implementation, the data was split into 80% training and 20% testing sets. Five commonly used models were tested: SVM, Passive Aggressive Classifier, Multinomial Naive Bayes, Logistic Regression, and Decision Tree Classifier. Text was converted to numbers using Count Vectorizer and TF-IDF Transformer, and a pipeline object was used to automate the workflow. The models were tested using the pipeline and the performance was analyzed. The best model was chosen based on the score.

```
X = combined_news_ds['text']
y = combined_news_ds['label']
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state = 1)
```

Reference

Adachi, F.de P. (2021) Deploying a fake news detector web application with Google Cloud Run and flask, Medium. Towards Data Science. Available at: https://towardsdatascience.com/deploying-a-fake-news-detector-web-application-with-google-cloud-run-and-flask-eb750cce986d (Accessed: April 25, 2023).
Bisaillon, C. (2020) Fake and real news dataset, Kaggle. Available at: https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset (Accessed: April 25, 2023).
Borcan, M. (2020) TF-IDF explained and python sklearn implementation, Medium. Towards Data Science. Available at: https://towardsdatascience.com/tf-idf-explained-and-python-sklearn-implementation-b020c5e83275 (Accessed: April 25, 2023).
DeepLearning.AI (2023) *Natural language processing (NLP) - A complete guide, (NLP) [A Complete Guide]*. Available at: https://www.deeplearning.ai/resources/natural-language-processing/ (Accessed: April 10, 2023).
Ganesan, K. (2023) *What are stop words?*, *Kavita Ganesan, PhD*. Available at: https://kavita-ganesan.com/what-are-stop-words/#.ZEeRBXbMI6k (Accessed: April 25, 2023).
Haskins, J. (2023) *Fake news: What laws are designed to protect*, *LegalZoom*. Legalzoom.com. Available at: https://www.legalzoom.com/articles/fake-news-what-laws-are-designed-to-protect (Accessed: April 10, 2023).
IBM (2023) *What is natural language processing?*, *IBM*. Available at: https://www.ibm.com/topics/natural-language-processing (Accessed: April 25, 2023).
INNOQ (2023) *ML-ops.org, ML Ops: Machine Learning Operations*. Available at: https://ml-ops.org/content/mlops-principles#:~:text=Model%20staleness%20test.,of%20prediction%20in%20intelligent%20software. (Accessed: April 25, 2023).

Answers

1- Using NLP techniques for fake news detection is important because it allows for large-scale analysis, automated real-time processing, and customization of the algorithm to meet specific needs. NLP can identify patterns and relationships in language constituents that are difficult for humans to detect, improving the accuracy of detecting fake news.
2- The reason for using the oversampling technique was to balance the data as it was imbalanced, which can result in biased and inaccurate models. The oversampling technique increases the number of samples in the minority class, thereby balancing the dataset and improving the performance of the model in detecting fake news.
3- The deployment plan for the fake news detection model involves creating a web application for users to input news articles or URLs to receive a prediction of whether they are true or fake. The target audience is companies and individuals looking to avoid being manipulated by false news. The plan is to start small and expand as demand increases, with a subscription-based model and fees based on the client's profile. The team responsible for data collection and model maintenance will constantly update and train the model to prevent it from becoming stale.
4- The date range of the "True.csv" dataset is approximately 1 year and 5 months, while the "Fake.csv" dataset's date range is approximately 2 years and 5 months.
5- The data was split into 80% training and 20% testing sets, and five commonly used NLP models were tested: SVM, Passive Aggressive Classifier, Multinomial Naive Bayes, Logistic Regression, and Decision Tree Classifier. The text was converted to numbers using Count Vectorizer and TF-IDF Transformer, and a pipeline object was used to automate the workflow. The models were tested using the pipeline, and the best model was chosen based on the score.

GitHub

https://github.com/rael-guimaraes/Fake_news_prediction