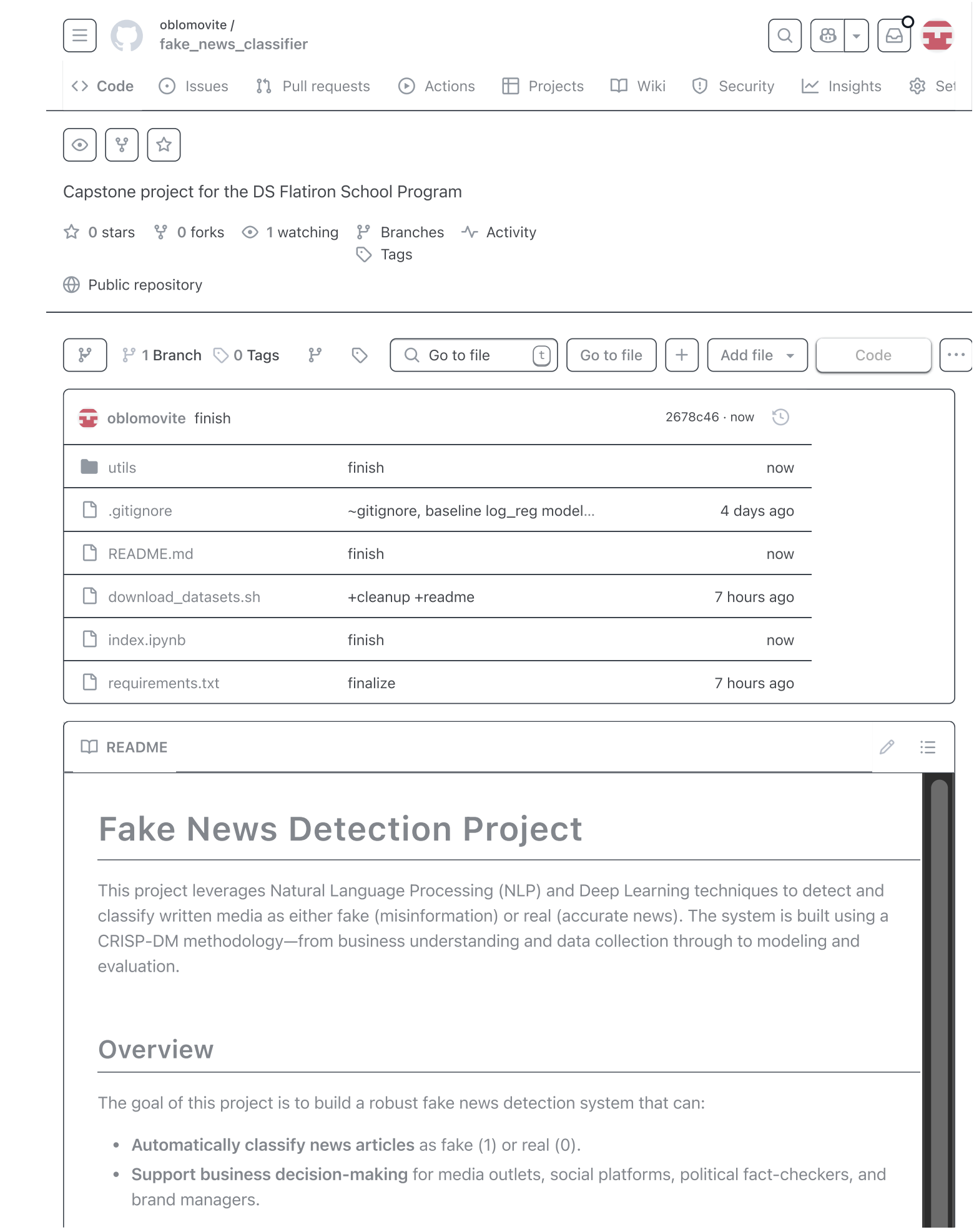
oblomovite /
fake_news_classifier
Code
Issues
Pull requests
Actions
Projects
Wiki
Security
Insights
Capstone project for the DS Flatiron School Program
0 stars
0 forks
1 watching
Branches
Activity
Tags
Public repository
1 Branch
0 Tags
Go to file
t
Go to file
Add file
Code
oblomovite finish
2678c46 · now
utils finish now
.gitignore ~gitignore, baseline log_reg model... 4 days ago
README.md finish now
download_datasets.sh +cleanup +readme 7 hours ago
index.ipynb finish now
requirements.txt finalize 7 hours ago
README
Fake News Detection Project
This project leverages Natural Language Processing (NLP) and Deep Learning techniques to detect and classify written media as either fake (misinformation) or real (accurate news). The system is built using a CRISP-DM methodology—from business understanding and data collection through to modeling and evaluation.
Overview
The goal of this project is to build a robust fake news detection system that can:
Automatically classify news articles as fake (1) or real (0).
Support business decision-making for media outlets, social platforms, political fact-checkers, and brand managers.

- Utilize both traditional ML (TF-IDF + Logistic Regression) and advanced deep learning (LSTM with Word2Vec embeddings) techniques.

Business Understanding

Misinformation is an exponentially growing problem in the digital age, with fake news spreading rapidly across social media platforms and news aggregators. This project aims to provide a solution to this problem by developing a fake news detection system that can automatically classify news articles as fake or real.

Business Use Cases

1. **News Aggregators & Media Outlets:**
   Automatically flag misinformation to improve credibility and audience trust.
2. **Social Media Monitoring:**
   Detect and mitigate the spread of fake news on social platforms.
3. **Political Fact-Checking:**
   Rapidly identify misleading political claims for quicker verification.
4. **Brand Reputation Management:**
   Monitor online content to protect brands from false and damaging information.
5. **Government & Regulatory Agencies:**
   Track misinformation trends to support informed policy decisions.

Data Understanding

The project uses two main datasets:

- Fake News Classification Dataset
- Fake and Real News Dataset

The model uses pre-trained word embeddings from:

- Google News Vectors
- Glove

I used the information provided by this article to guide me through the project.

The datasets contain news articles from various sources, along with metadata such as the article's URL, date, and text. The data is labeled as either fake or real, allowing for supervised learning model training.

Data Preparation

The data preparation process involves cleaning, preprocessing, and transforming the raw data into a format suitable for model training:

- **Text Cleaning:** Remove special characters, URLs, and non-alphanumeric characters.
- **Text Tokenization:** Split text into individual words or tokens.
- **Text Vectorization:** Convert text tokens into numerical vectors using TF-IDF or Word2Vec embeddings.

Exploratory Data Analysis (EDA)

This project includes an EDA and Visualization section to explore the dataset's characteristics. The notebook checkpoints the results of the EDA here.

Modeling

- **Baseline Model:** TF-IDF + Logistic Regression.
- **Advanced Model:** LSTM using pre-trained Word2Vec (Google News Vector) embeddings.

Project Structure

```
news-classification/
├── datasets/  # downloaded datasets
│ ├── fake-and-real-news/
│ └── fake-news-classification/
├── images/ # EDA images
├── metrics/ # evaluation results
├── index.ipynb # main notebook
├── download_datasets.sh # dataset download script
├── README.md
└── requirements.txt
```

Conclusion

The performance of the baseline model (TF-IDF + Logistic Regression) was exceptional, easy to train and interpret. The advanced model (LSTM + Word2Vec) performed similiarly well but required considerably more computational resources and time to train. The choice of model depends on the specific use case and available resources.

Languages