Unified Dataset for Fake, Spam, and Legit Data

Description
This dataset is designed for classifying information into three categories: fake, spam, and legit. It combi

Dataset Structure
The dataset includes the following files:
- `data.csv`: The main dataset file containing all entries.
- `README.md`: This file, providing details about the dataset.

Columns in `data.csv`
- `text`: The text content of the message, email, or news article.
- `category`: The label indicating whether the entry is `fake`, `spam`, or `legit`.

Source Attribution
This dataset is a compilation of multiple sources:

1. **YouTube Spam and Ham Data**:
   - Source: [YouTube Spam Collection](https://doi.org/10.24432/C58885)
   - Description: Contains spam and legitimate comments from YouTube videos.
   - Used: 1005 spam entries and 951 legitimate entries.

2. **Email Spam Dataset**:
   - Source: [Email Spam Collection](https://www.kaggle.com/datasets/venky73/spam-mails-dataset)
   - Description: Collection of spam and legitimate emails.
   - Used: 936 spam entries and 2363 legitimate entries.

3. **SMS Spam Dataset**:
   - Source: [S https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset)
   - Description: Dataset of spam and legitimate SMS messages.
   - Used: 747 spam entries and 4825 legitimate entries.

4. **WELFake Dataset**:
   - Source: [WELFake Fake News Dataset](https://www.kaggle.com/datasets/vcclab/welfake-dataset)
   - Description: Contains fake and legitimate news articles.
   - Used: 2641 fake entries and 1359 legitimate entries.

5. **GossipCop Dataset**:
   - Source: [GossipCop Fake News Dataset](Retrieved from https://github.com/KaiDMML/FakeNewsNe
   - Description: Dataset of fake and legitimate news articles from GossipCop.
   - Used: 238 fake entries and 762 legitimate entries.

Data Collection and Processing
The above datasets were combined and processed to create a comprehensive dataset for classifying inf

Usage Instructions
To use this dataset:
1. Download the `data.csv` file.
2. Load the dataset into your preferred data analysis or machine learning tool (e.g., Pandas, scikit-learn)
3. Use the `text` column for your input data and the `category` column for labels.

License
This dataset is released into the public domain under the [Creative Commons Zero (CC0) license](https:/

Acknowledgments
We would like to thank the creators of the original datasets for their valuable contributions.

nes data from multiple sources, including YouTube comments, email, SMS messages, and news articles,

t)

ormation into three categories: fake, spam, and legit. Duplicate entries were removed, and the data wa

).

to provide a diverse and comprehensive collection. The dataset can be used for training machine learni

s cleaned and preprocessed to ensure consistency and quality.

vithout any restrictions. Please note that the original datasets used to create this compilation may have

ng models for text classification and is ideal for research in fake news detection, spam filtering, and text

their own licensing conditions, and it is your responsibility to ensure compliance with those terms wher

analytics.

ı using this dataset.