Data Collection

Utilizing web scraping techniques, I extracted data from the Medium blog of author Will Kohersen. I focused on gathering articles that were freely available to maximize data collection. The tech stack comprised Selenium and Beautiful Soup. Selenium was employed to retrieve the HTML of the main page, while Beautiful Soup was used to scrape article links and tags. Subsequently, each link was iterated through using Selenium to extract various attributes such as title, blog content, upvotes, comment count, read time, publish date, and tags. The collected data was stored in CSV format.

Exploratory Data Analysis

Data Cleaning

- Converted variables like upvotes, comment count, and read time to integer type.
- Converted datetime to datetime object and extracted additional columns such as weekday, month, and year.
- Calculated the length of each blog and added it as a new column.
- Consolidated similar tags to address imbalance, resulting in a total of 33 tags.

Numeric Data Analysis

- Plotted distributions of upvotes, comment count, read time, and blog length, observing right skewness.
- Explored relationships between various attributes like upvotes, comment count, read time, and blog length using correlation analysis.
- Identified blogs with the highest upvotes, comment count, blog length, and read time.
- Analyzed the author's weekly and monthly writing patterns and interests over time.

Text Data Analysis

- Generated word clouds for each tag to identify recurring words.
- Employed TextBlob to analyze sentiment and subjectivity of each blog.
- Conducted unigram, bigram, and trigram analysis before and after removing stop words.
- Explored the author's use of different parts of speech.
- Utilized scattertext and spaCy to identify characteristic terms and associations within each tag.
- Employed Count Vectorizer and LDA for topic modeling to uncover hidden themes in the author's blog.

Models Used

Machine Learning Algorithms

- Employed Multinomial Naive Bayes, XGBoost, Random Forest, SVM, and hybrid SVM-XGBoost models for tag classification based on blogs.

Deep Learning Models

- Trained Word2Vec embeddings on the data and utilized LSTM, CNN, and combinations of LSTM & GRU for tag classification. However, due to the small dataset size, these models did not perform optimally.

Transformer

- Utilized TF Bert for classification, albeit with limited training due to resource constraints.

Model Evaluation

- Multinomial Naive Bayes emerged as the top-performing model with an accuracy of approximately 87.7%. Most ML models achieved accuracies above 80%.
- Deep learning models struggled to perform well, as anticipated given the small dataset size, achieving an accuracy of around 50%.
- The TF Bert transformer model yielded an accuracy of approximately 70%.

Conclusion

- The project successfully collected and analyzed data from the author's Medium blog.
- Various analyses provided insights into the content, writing patterns, and audience engagement.
- Model evaluations highlighted the effectiveness of traditional ML algorithms over deep learning approaches, likely due to the dataset's size limitations.