CLASSIFYING REDDIT DISCUSSIONS TO IDENTIFY TRENDING TECHNOLOGIES AND GADGETS

Problem Statement

- **Objective:** Understand which technologies and gadgets generate the most discussion on Reddit.
- **Goal:** Develop a machine learning model to classify posts from technology and gadgets.
- **Challenge:** Limited labeled data makes accurate classification difficult.

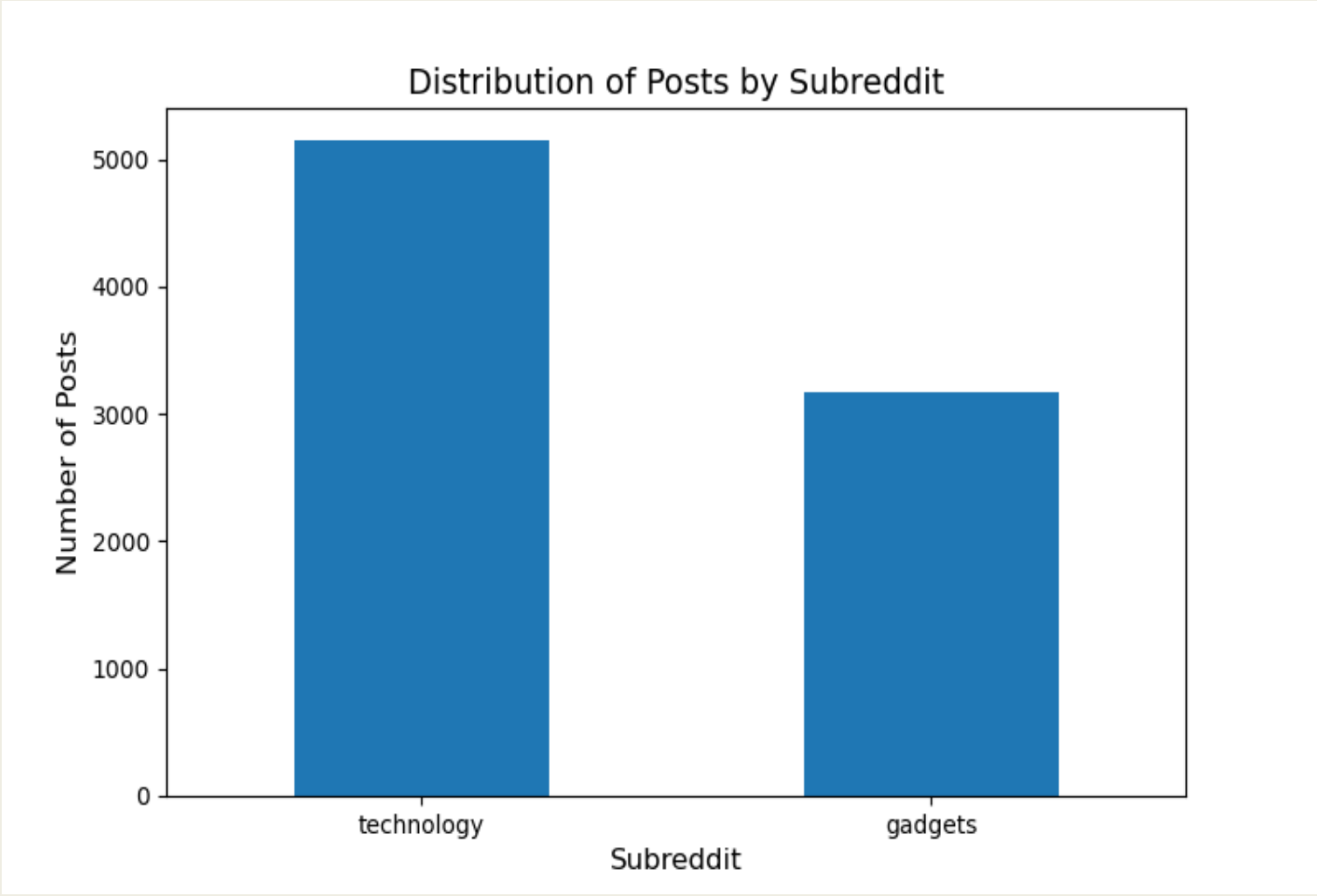
Data Collection

- **Data Source:**
  We used the Reddit API to collect posts from two subreddits:
  - **Technology**
  - **Gadgets**

- **Total Posts:**
  - **8306 posts** were gathered across both subreddits.
  - ***Features:*** *Post ID, Title, Content, Timestamp, Subreddit, Comments*

Post Distribution

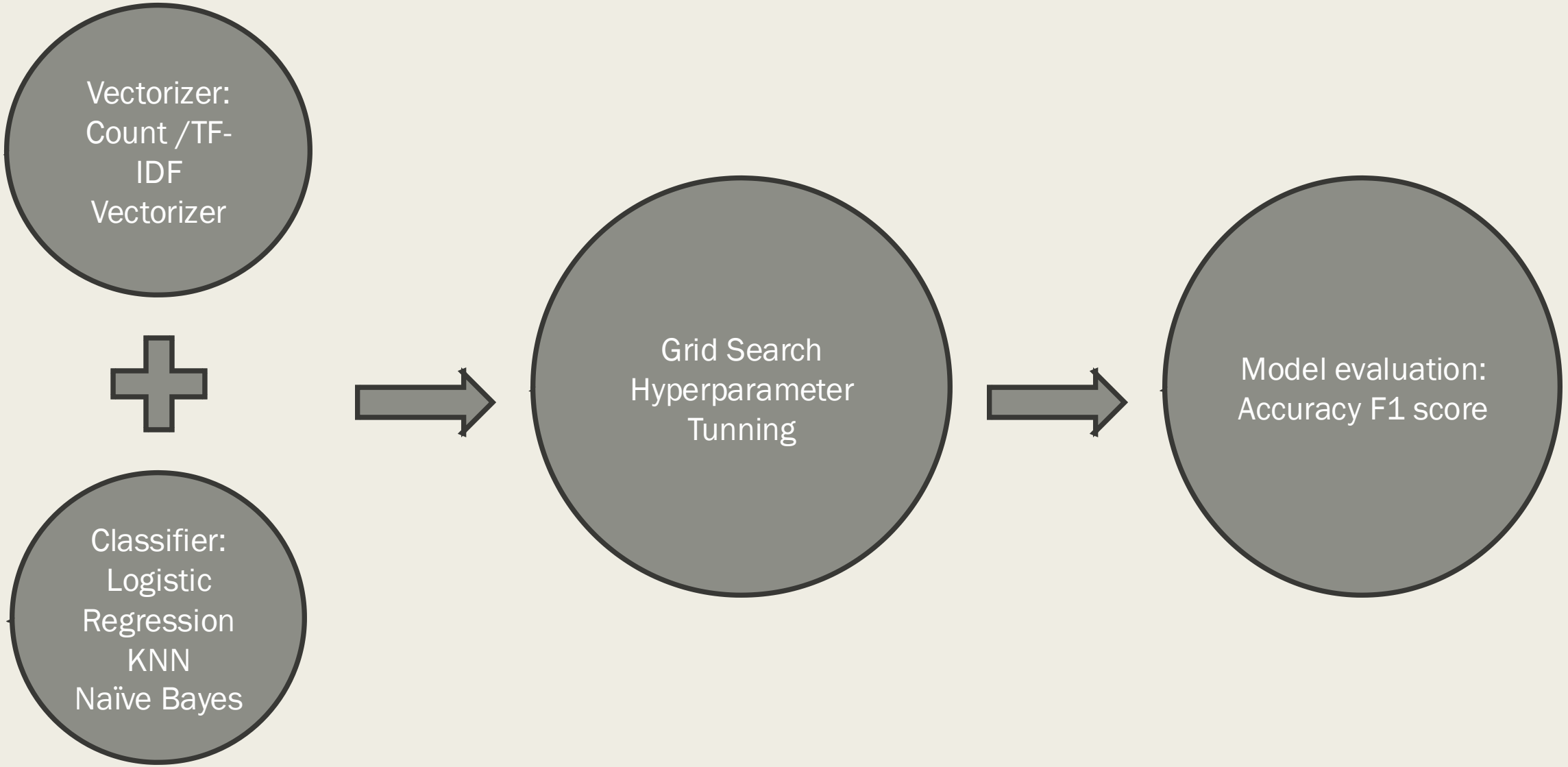
- **Total Posts Collected:8306**
  - *Technology: 5143*
  - *Gadgets : 3163*



Data Processing

- **Combined :post titles and comments.**
- **Removed:** URLs, HTML tags, emoticons, special characters.
- **Standardized:** Lowercased text, removed punctuation and numbers.
- **Tokenized & Lemmatized:** Tokenize and lemmatize words.
- **Filtered Stop Words:** Removed common words

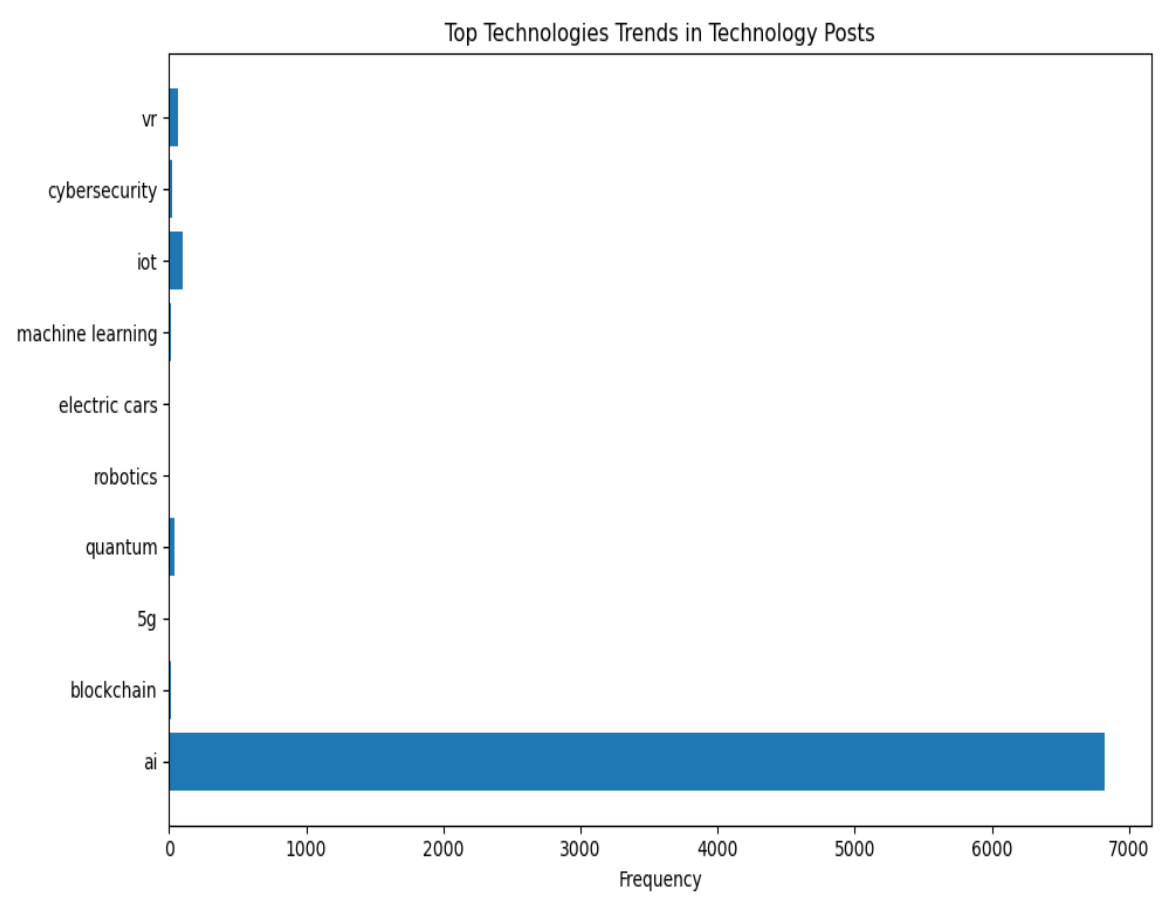
Modeling Selection

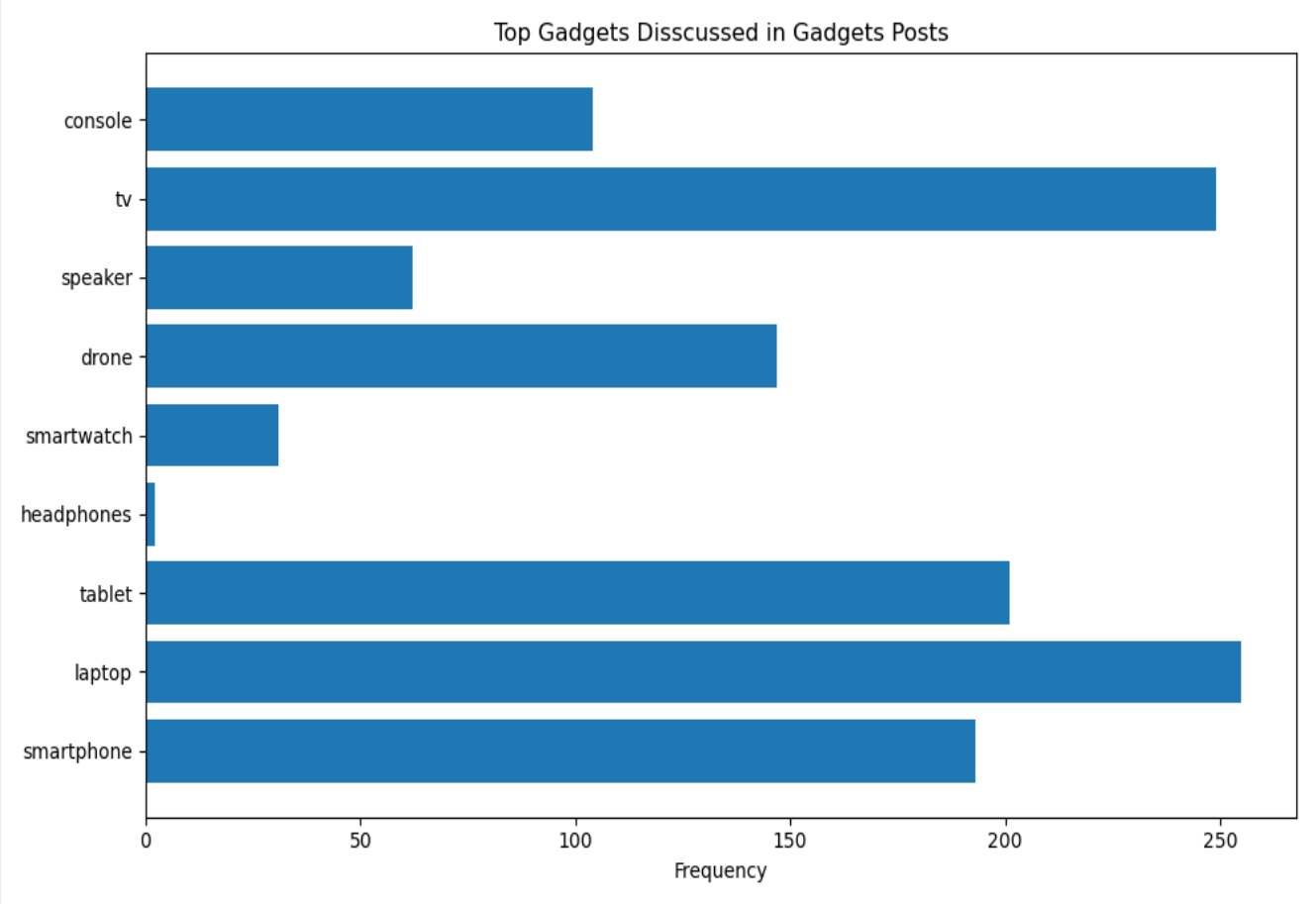


Model Performance Comparison

Metric	Testing Accuracy	Specificity	Recall	F1-score Technology	F1-score Gadgets
Logistic Regression with Count Vectorizer	0.84	0.75	0.89	0.87	0.78
KNN with TF-IDF Vectorization	0.84	**0.83**	0.85	0.87	**0.80**
Naive Bayes with TF-IDF Vectorizer	**0.85**	0.77	**0.91**	**0.88**	**0.80**

Most Discussed Technologies and Gadgets





Conclusion

- **Best Model:**
  Naive Bayes with TF-IDF vectorizer achieved the highest accuracy and F1 score for classifying Reddit posts.

- **Key Insights:**
  - **Accurate Classification of Posts :** High distinction between technology and gadget posts.
  - **Identification of Emerging Trends :** AI (technology); smartphones, laptops(gadgets).
  - **Scalability :** Can track trends across other subreddits.