Research Paper review – Reuters 21578

Name : jill Padariya Roll no : 92000133007

Dataset and framework:

1. Reuter 21578:

- A well-known document collection from Reuters news in 1987.

- fig 6. Contains multi-class and multi-label datasets with 90 categories and 10,788 documents.

- Split into training and testing sets with 7,769 and 3,019 documents, respectively.

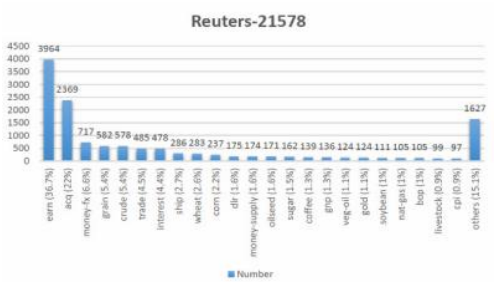


Figure 6. Category distribution of dataset Reuters21578

2. Re0:

- A subset derived from Reuter 21578.

- fig 1. Contains 13 categories with 1,504 documents.

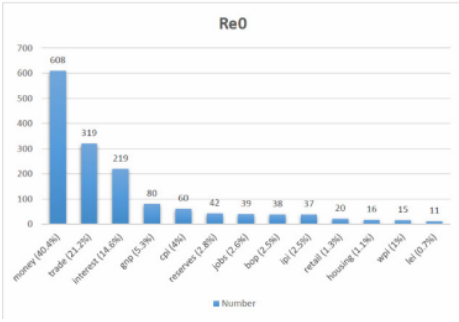


Figure 1. Category distribution of dataset Re0

3. Re1:

- Another subset of Reuter 21578.

- fig 2. Contains 25 categories with 1,657 documents.

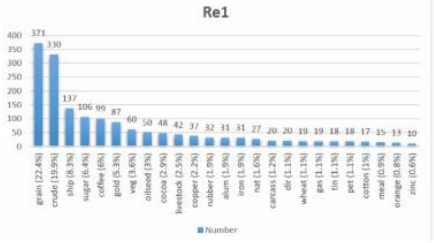


Figure 2. Category distribution of dataset Re1

4. Re52:

- A single-label subset of Reuter 21578.

- fig 5. Contains 52 categories and 9,130 documents.

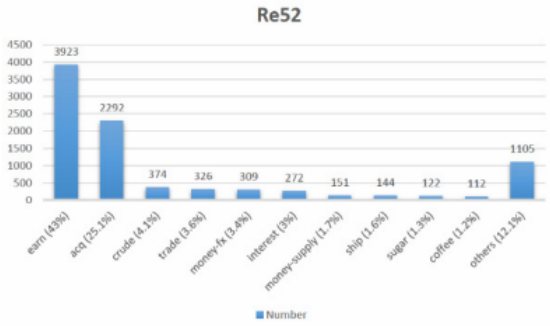


Figure 5. Category distribution of dataset Re52

5. k1a and k1b:

- Subsets of WebACE.

- Contain up to a total of 2,340 documents and 21,839 effective words.

6. RCV1 (Reuters Corpus Volume I):

- An extensive archive of over 800,000 manually categorized newswire stories from Reuters, Ltd.

- Comprises 103 categories and 804,414 documents.

7. Imbalanced Classes:

- All the datasets exhibit imbalanced class distributions, meaning that the number of instances in each class is not equal. This is an important consideration, as it reflects real-world scenarios where certain categories may be more prevalent than others.

8. Algorithm Evaluation:

- The study evaluates the performance of various machine learning algorithms and term weighting methods on these datasets. The inclusion of diverse datasets and class imbalances is crucial for assessing the robustness and generalizability of these algorithms.

feature selection method for reuter 21578:

Feature selection is the process of choosing relevant features for a classification model, commonly in text classification where features are words. The goal is to enhance model accuracy and efficiency by identifying informative terms for each class. Methods like Information Gain, Chi-Square, and Mutual Information

assess features based on statistical properties. The choice of method depends on dataset characteristics, acknowledging that different methods may excel for various text data types.

Method	PCA+logistic regression		SVM	
	macro-F1	micro-F1	macro-F1	micro-F1
One-hot encoding (Base)	6.39%	16.60%	8.86%	16.79%
TF	42.74% (+36.35%)	85.41% (+68.81%)	45.83% (+36.97%)	86.16% (+69.37%)
TFIDF	16.86% (+10.47%)	77.67% (+61.07%)	49.89% (+41.03%)	86.86% (+70.07%)
TFICF	46.84% (+40.45%)	85.46% (+68.86%)	42.10% (+33.24%)	87.00% (+70.21%)
TFChi	11.25% (+4.86%)	36.68% (+17.08%)	17.04% (+8.18%)	31.52% (+14.73%)
TFOdd	56.52% (+50.13%)	80.91% (+64.31%)	48.31% (+39.45%)	81.91% (+65.12%)
TFProb	10.11% (+3.72%)	62.21% (+45.61%)	50.71% (+41.85%)	74.86% (+58.07%)
TFRF	37.75% (+29.36%)	85.59% (+67.99%)	54.52% (+45.66%)	86.63% (+69.84%)
TFCRF	17.82% (+11.43%)	75.84% (+59.24%)	52.98% (+44.12%)	82.29% (+65.50%)

Result of re0, re1, re52:

In the Re0, Re1, and Re52 datasets, the study assessed four term weighting methods: TF, TFIDF, TFRF, and TFProb. Results varied across datasets and methods. Notably, TFRF and TFProb performed well in Re0 and Re1 under different classification models, while TFRF excelled in Re52 under SVM classification. The findings underscore the significance of selecting the right term weighting method for accurate text classification and offer insights for improving text classification system performance.

Method	PCA+logistic regression		SVM	
	macro-F1	micro-F1	macro-F1	micro-F1
One-hot encoding (Base)	76.93%	85.24%	76.26%	83.31%
TF	73.88% (-3.05%)	84.11% (-1.13%)	75.06% (+1.2%)	83.25% (-0.06%)
TFIDF	61.64% (-15.29%)	81.85% (-3.39%)	83.00% (+6.74%)	87.76% (+4.45%)
TFICF	66.60% (-10.33%)	78.66% (-6.58%)	73.83% (-2.43%)	78.92% (-4.39%)
TFChi	75.51% (-1.42%)	82.51% (-2.73%)	69.82% (-6.44%)	74.20% (-9.11%)
TFOdd	77.80% (+0.87%)	81.58% (-3.66%)	73.30% (-2.96%)	72.54% (-10.77%)
TFProb	39.95% (-40.98%)	92.80% (+7.56%)	49.51% (-26.75%)	74.93% (-8.38%)
TFRF	75.92% (-1.01%)	86.30% (+1.06%)	83.15% (+6.89%)	86.17% (+2.86%)
TFCRF	61.99% (-14.94%)	82.58% (-2.66%)	71.92% (-4.34%)	81.98% (-1.33%)

Method	PCA+logistic regression		SVM	
	macro-F1	micro-F1	macro-F1	micro-F1
One-hot encoding (Base)	69.72%	84.67%	70.43%	83.77%
TF	75.04% (+5.32%)	84.54% (+1.87%)	71.61% (+1.18%)	83.04% (-0.73%)
TFIDF	50.98% (-18.74%)	77.43% (-2.76%)	72.23% (+1.80%)	86.06% (+2.29%)
TFICF	78.01% (+8.29%)	86.96% (+2.29%)	75.18% (+4.75%)	83.16% (-0.61%)
TFChi	73.01% (+3.29%)	83.83% (-0.84%)	74.38% (+3.95%)	81.11% (-2.66%)
TFOdd	79.33% (+9.61%)	87.45% (+2.78%)	75.10% (+4.67%)	86.42% (+2.65%)
TFProb	36.61% (-33.11%)	68.13% (-16.54%)	70.72% (+0.29%)	83.34% (-0.43%)
TFRF	77.37% (+7.65%)	88.89% (+4.22%)	75.81% (+5.38%)	87.81% (+4.04%)
TFCRF	72.84% (+3.12%)	87.14% (+2.47%)	76.71% (+6.28%)	87.87% (+4.10%)

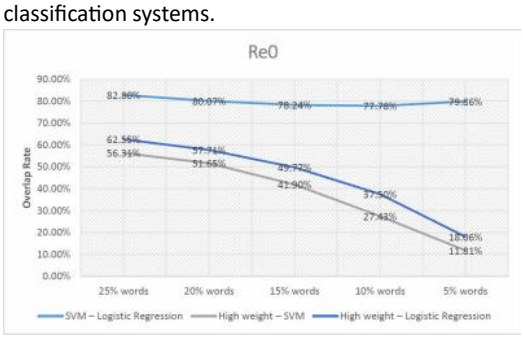
Method	PCA+logistic regression		SVM	
	macro-F1	micro-F1	macro-F1	micro-F1
One-hot encoding (Base)	11.07%	18.15%	11.39%	17.83%
TF	64.83% (+56.76%)	93.15% (+75.0%)	66.59% (+55.20%)	91.25% (+73.42%)
TFIDF	39.73% (+28.66%)	87.67% (+69.52%)	66.85% (+54.46%)	93.15% (+75.32%)
TFICF	74.16% (+63.09%)	94.20% (+76.05%)	67.40% (+56.01%)	89.81% (+71.98%)
TFChi	17.25% (+6.18%)	46.58% (+28.43%)	46.26% (+34.87%)	82.94% (+65.11%)
TFOdd	68.84% (+57.77%)	89.72% (+71.57%)	66.14% (+54.75%)	85.21% (+67.38%)
TFProb	14.80% (+3.73%)	76.46% (+58.31%)	49.10% (+37.62%)	83.89% (+66.06%)
TFRF	54.66% (+43.59%)	92.57% (+74.42%)	68.99% (+57.60%)	93.27% (+75.44%)
TFCRF	23.12% (+12.05%)	84.16% (+66.01%)	68.82% (+57.43%)	91.87% (+74.04%)

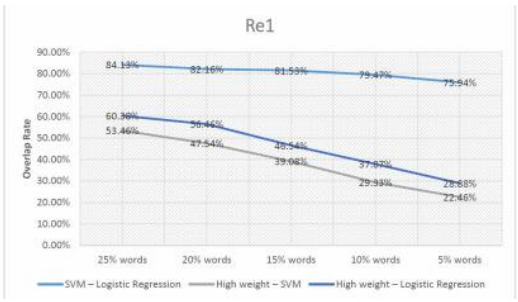
Overlap Rate of reuters 21578:

The study did not examine the overlap rate of the Reuters21578 dataset but examined it in other datasets. It found that the overlap between the different weighting methods can reach up to 70%, but between high-weight terms and important classification terms, the degree of overlap is, typically, very low less than 10% and the number of weights decreases at different word sizes. These insights can be valuable for researchers and practitioners aiming to improve textual classification systems.

Overlap Rate of re0, re1, re52:

The study examined weighting methods on the Re0, Re1, and Re52 datasets. It found that the overlap for different weighting methods can reach 70%, but the overlap between high-weight terms and important classification terms is, in general, very low less than 10% Differs across datasets by different word weighting methods There was effect, and the study highlighted the importance of choosing the right method to classify information accurately This insight is valuable to researchers and practitioners who they are aimed at improving text classification systems.





Classification report :

The code provides a comprehensive text classification implementation using the Reuters-21578 dataset. It covers resource download, preprocessing, and tokenization. The processed data is transformed into a DataFrame, split into training and testing sets, and subjected to TF-IDF vectorization. A Multinomial Naive Bayes classifier is trained and evaluated using scikit-learn's `classification_report`. While effective, adding comments and documentation for clarity is suggested. Parameter tuning and cross-validation could enhance classifier performance. Overall, it serves as a solid foundation for text classification research.

```
[nltk_data] Downloading package reuters to /root/nltk_data...
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
              precision    recall  f1-score   support

         acq       0.54      0.96      0.69       469
        alum       0.00      0.00      0.00         7
      barley       0.00      0.00      0.00         6
         bop       0.00      0.00      0.00        20
     carcass       0.00      0.00      0.00        15
  castor-oil       0.00      0.00      0.00         2
       cocoa       0.00      0.00      0.00        17
      coffee       0.00      0.00      0.00        25
```

Data Splitting for Model Evaluation: Training, Test, and Unused Sets :

The code provides a methodical approach to data splitting for machine learning model evaluation, ensuring reproducibility with a fixed seed. It effectively creates distinct training, test, and unused sets based on specified proportions, incorporating unique IDs and group names for transparency. The resulting DataFrames are well-structured, aiding clear separation of data. This code is valuable for researchers dealing with limited datasets, complementing text classification code for systematic data preparation. To enhance it, adding assertions for proportion validation and commenting on the chosen proportions would improve overall clarity.

```
                                          data    target    id  \
4593  bayer world group 1986 profit billion mark bil...      earn     0
3614  comput microfilm corp lt comi year net shr 23 ...      earn     2
510   spain foreign reserv rise februari spain forei...  reserves     6
7331  gerber lt grb buy gerber system lt gsti share ...       acq     7
5458  santa anita realti lt sar quarterli dividend q...      earn     8
...                                                 ...       ...   ...
5734  dst system inc lt dst regular payout set qtli ...      earn  10783
5191  ocean inc lt obci year net shr eight ct vs sev...      earn  10784
5390  drexel offici ha stake epsilon data lt epsi se...       acq  10785
860   pennsylvania real estat invest trust lt pei op...      earn  10786
7270  circon corp lt ccon 4th qtr shr loss two ct vs...      earn  10787

          group
4593  training_set
3614  training_set
510   training_set
7331  training_set
5458  training_set
...            ...
5734  training_set
5191  training_set
5390  training_set
860   training_set
7270  training_set

[7551 rows x 4 columns]
```

```
Test Set:
                                          data    target    id  \
8353  marcor lt maar expect fiscal year profit marco...      earn     1
8048  transamerica incom lt tai monthli dividend shr...      earn     4
8440  bank england said invit borrow 10 pct later to...  interest    16
8036  armtek lt arm unit leverag buyout complet cond...       acq    25
8811  fleet lt flt aim speed merger norstar fleet fi...       acq    28
...                                                 ...       ...   ...
7734  bank japan satisfi yen current rang bank japan...       gnp  10752
9167  intermark lt imi seek major pier 1 lt pir stak...       acq  10759
7849  diversifi industri lt dei 1st qtr oper net per...      earn  10762
8666  novar electron corp lt novr 4th qtr jan three ...      earn  10765
8322  hospit staf servic inc lt hssi 1st qtr feb 28 ...      earn  10778

        group
8353  test_set
8048  test_set
8440  test_set
8036  test_set
8811  test_set
...        ...
7734  test_set
9167  test_set
7849  test_set
8666  test_set
8322  test_set

[1618 rows x 4 columns]
```

```
Unused Set:
                                          data    target     id  \
10382  iran say ha better weapon silkworm iranian pri...      ship      3
10334  avatar hold inc lt avtr year net oper shr 32 c...      earn      5
9730   alli lt asu sell michigan unit alli supermarke...       acq      9
9281   twa lt twa court settlement usair lt u tran wo...       acq     17
9451   mcm corp lt mcmc set quarterli qtli div six ct...      earn     22
...                                                 ...       ...    ...
10555  recent oil demand pct year ago oil demand mean...     crude  10743
10627  burlington coat factori warehous corp lt bcf n...      earn  10745
9998   claremont tell sec seek 15 pct champion produc...       acq  10758
9274   india australia agre improv trade indian austr...    cotton  10764
10583  baker say stand pari currenc agreement treasur...  money-fx  10775

          group
10382  unused_set
10334  unused_set
9730   unused_set
9281   unused_set
9451   unused_set
...           ...
10555  unused_set
10627  unused_set
9998   unused_set
9274   unused_set
10583  unused_set

[1619 rows x 4 columns]
```

Optimizing LDA: Exploring Topic Numbers and Enhancing Interpretability :

The code enhances topic modeling by introducing coherence scores for evaluating Latent Dirichlet Allocation (LDA) topics. Using Gensim CoherenceModel, it calculates scores for various topic numbers, aiding researchers in selecting the optimal number of topics. The loop prints coherence scores and identifies the optimal number based on the highest score. This addition provides a crucial quantitative measure for topic selection. To improve, adding comments on the choice of 'c_v' coherence and considerations, along with a visualization of coherence scores against topic numbers, would enhance clarity and interpretation.

```
Number of Topics: 5, Coherence Score: 0.5024227698091698
Optimal Number of Topics: 5
```

Enhanced Topic Modeling Evaluation: Optimal Topics and Top-Level Categories :

The code enhances topic modeling by printing the optimal number of topics and predefined top-level categories, improving result interpretability. It utilizes Gensim's CoherenceModel for quality assessment, guiding researchers in model configuration selection. The explicit listing of top-level categories provides a clear reference to high-level themes. The code's simplicity and clarity make it a valuable tool for Reuters-21578 dataset topic modeling. To improve, adding comments explaining the rationale behind top-level categories would enhance understanding.

```
Optimal Number of Topics: 5
Top-Level Categories:
1. earn
2. acq
3. money-fx
4. grain
5. crude
6. trade
7. interest
8. ship
```

Category-Based Topic Modeling with LDA: Unveiling Themes in Text Data :

The code offers a comprehensive approach to topic modeling using Latent Dirichlet Allocation (LDA) on preprocessed text data. It employs Gensim and Pandas for essential functionalities, including tokenization and dictionary creation. The run_lda function enables modular experimentation with different topic numbers. The code efficiently applies LDA, calculates word counts for each category, and generates a categories_list for insights into topic distribution within the dataset. Its well-structured methodology makes it a valuable tool for category-based document analysis.

```
Categories:
acq
alum
barley
bop
carcass
castor-oil
cocoa
coconut
coconut-oil
coffee
copper
copra-cake
corn
```