Subreddits Classification Model

Outline

1. Problem definition
2. Data Collection and Cleaning
3. Model Set up
4. Model evaluation and live testing
5. Conclusion

The Problem

- Practical problem:
  - Classify a text based on predefined reddit categories

- Data Science problem:
  - Digitize the text for natural language and machine learning processing

Data Collection, Cleaning and EDA

- The `subreddit` and `title` columns were complete
  - Start with the minimum: drop all other columns with missing data
- The index was reset
  - Data was brought in three batches
  - Duplication of indices when merged
- Baseline prediction
  - The majority class represents 62%

Model setting

- Logistic Regression Vs. Multinomial Naive Bayes
  - Gridsearch parameters:
    - Max_features
    - Min_df
    - Ngram_range
  - Other parameters:
    - stop_words
    - Lemmatization
    - Stratification

Model Evaluation

- Naive Bayes better with test data
  - Log Reg better with training data

- CountVectorizer outperformed TF-IDF

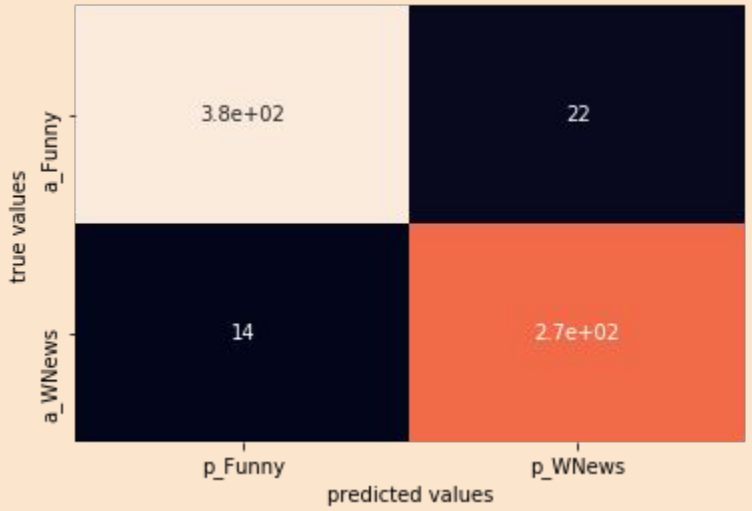
- Lemmatization did not improve test score

Model Evaluation

(in %, Train - Test)	LogReg	M.Naive Bayes
CountVect - reduced data	99.9 - 86.05	99.1 - 90.4
TF-IDF - reduced data	97.0 - 90.2	97.5 - 91.1
TF-IDF	96.2 - 89.5	97.4 - 92.7
CountVectorizer	99.1 - 90.0	98.8 - 94.7
cVec - stratify	97.8 - 91.8	99.8 - 93.0
Lemmatization	99.5 - 89.3	98.8 - 94.7

Model Evaluation: Confusion Matrix

	p_Funny	p_WNews
a_Funny	382	22
a_WNews	14	268



FP (Funny when predicted News) = 22

FN (News when predicted Funny = 14

Accuracy: 94.75, Sensitivity: 95% (low FN rate), Precision 92% (high FP rate, over predicting minority class)

Model testing

In[ ]: examples = ['school', 'school shooting', 'french', 'french guy', 'mother', 'young mother]

In[ ]: examples_preds = gs.predict(examples)

Out[ ]:

Example	Prediction
school	funny
school shooting	funny
french	worldnews
french guy	funny
mother	funny
young mother	worldnews

Possible improvements

- Increase the amount of data
- Try other models: KNN, SVM, or others
- Challenge the model with a third subreddit

Thank You !

Questions?