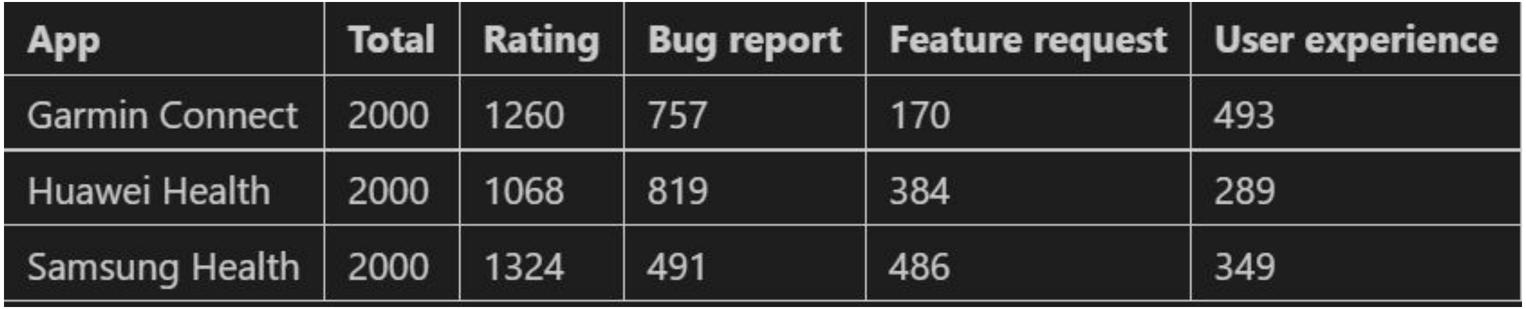
NLP - French user reviews classification dataset

2022/23 - MEIC

Data Provenance

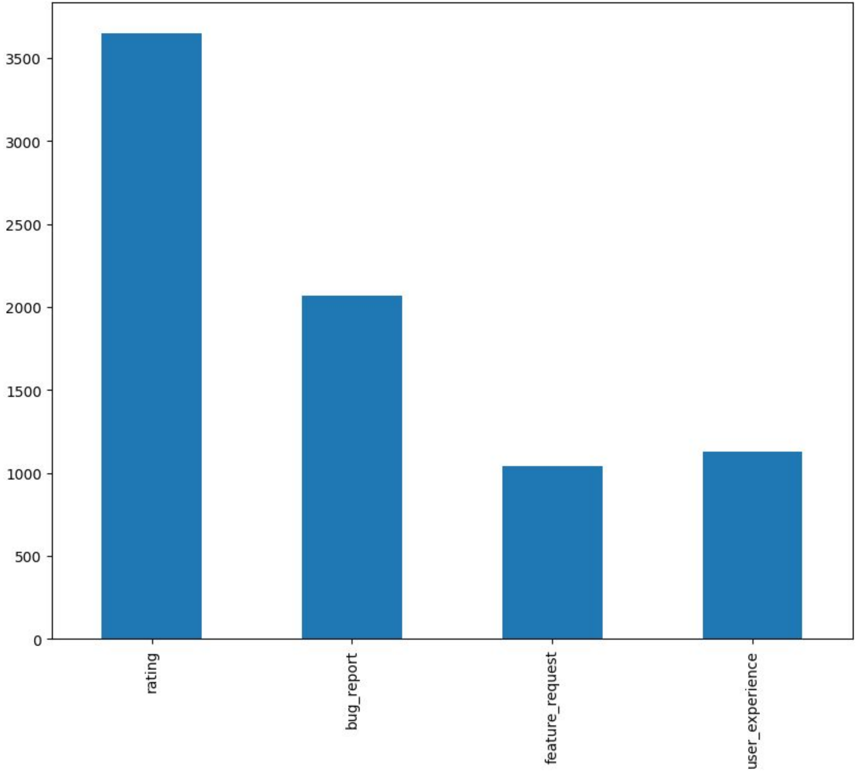
Nature of the problem

- Predict the type of review from from three applications on Google Play: Garmin Connect, Huawei Health and Samsung Health
- 6000 total entries
- Reviews only in french
- Multilabel classification problem



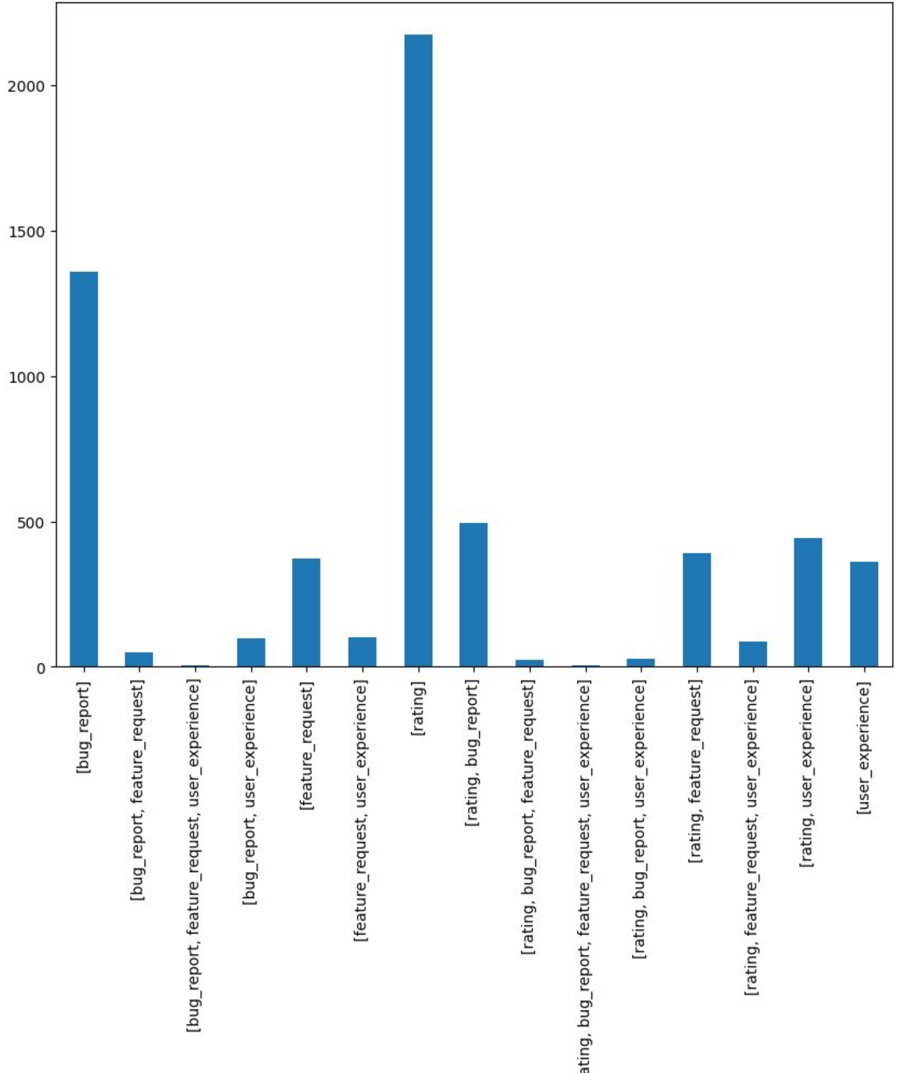
App	Total	Rating	Bug report	Feature request	User experience
Garmin Connect	2000	1260	757	170	493
Huawei Health	2000	1068	819	384	289
Samsung Health	2000	1324	491	486	349

Labels distribution

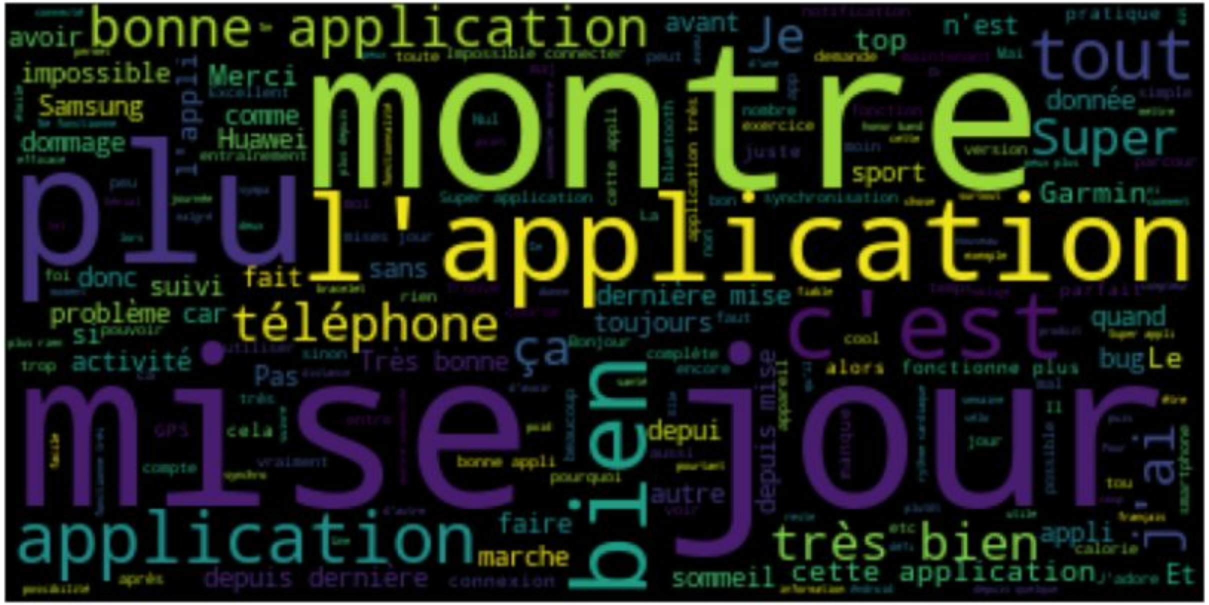


Data Exploration

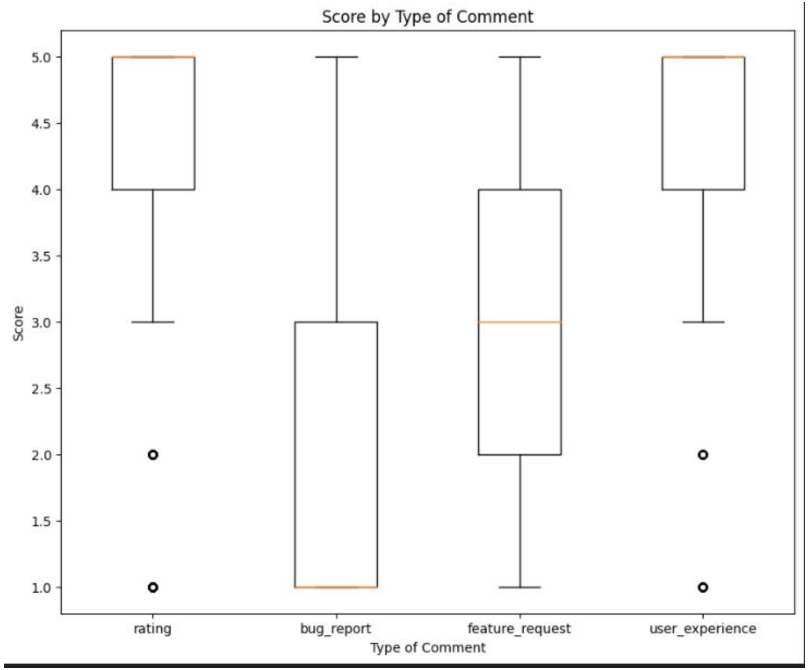
Class overlap



Word Cloud visualization



Score by Type of Comment



Machine Learning Approach

Approach 1: Problem Adaptation

Solving a multilabel classification problem requires some adaptation in order to make most regular models work on it. We tried the following approaches:

1. Binary Relevance
2. Classifier Chains
3. Label powerset

Approach 2: Using adapted algorithms

Use the algorithms that natively work on these kinds of problems, most of them provided by the scikit-multilearn library.

Pipeline steps

Pipeline steps

1. Choose pre-processor: sparse representations (TF-IDF)

or dense representation (word-embeddings)

2. Try each of the proposed problem approaches to multilabel classification.

3. Extract metrics

Step 1: Preprocessor fine tuning

Word2Vec : 2 models

In the cases for which we used Word2Vec for our preprocessing step, we tried two of them: a model trained by us on our own data, plus another trained on french wikipedia texts.

TF-IDF hyperparameter tuning with LogReg

Use GridSearchCV with a Logistic Regression in order to find optimal parameters for our TF-IDF Vectorizer.

Note on preprocessing steps

Before training the Word2Vec model, as stemming and lemmatization can alter the spelling of words and reduce their distinctiveness, leading to a loss of information and potentially affecting the quality of the learned word embeddings, we applied only very basic preprocessing.

Note on preprocessing steps

TF-IDF :

- Lowercasing
- Tokenization
- Removal of stopwords
- Stemming
- Lemmatization

Word2Vec :

- Lowercasing
- Removal of stopwords

Step 2: Apply each of the approaches

More hyperparameter tuning

Try different models with many combinations of parameters and use 5-fold stratified cross-validation.

Step 3: Extract metrics

Metrics extracted

Precision

Recall

F1-score - preferred metric to maximize, the metric we used to compare model performance essentially

Accuracy - really bad because of imbalance data

Results

Results - Adaptive Algorithms TF-IDF

Best Algorithm - Ridge Classifier; Accuracy - 0.57166

	Precision	Recall	F1-score
rating	0.810	0.801	0.806
bug report	0.820	0.816	0.818
feature request	0.756	0.567	0.648
user experience	0.594	0.350	0.440
micro average	0.787	0.710	0.746
macro average	0.745	0.634	0.678
weighted average	0.775	0.710	0.736
samples average	0.788	0.763	0.753

Results - Adaptive Algorithms Word2Vec

Best Algorithm - Ridge Classifier; Accuracy - 0.6111

	Precision	Recall	F1-score
rating	0.810	0.801	0.806
bug report	0.765	0.778	0.765
feature request	0.789	0.654	0.678
user experience	0.594	0.350	0.487
micro average	0.787	0.710	0.746
macro average	0.745	0.634	0.678
weighted average	0.775	0.710	0.736
samples average	0.788	0.763	0.753

Results - Binary relevance TF-IDF

Best Algorithm - Random forest classifier; Accuracy - 0.5375

	Precision	Recall	F1-score
rating	0.811	0.805	0.808
bug report	0.768	0.785	0.776
feature request	0.647	0.159	0.255
user experience	0.729	0.155	0.255
micro average	0.787	0.621	0.694
macro average	0.739	0.476	0.524
weighted average	0.766	0.621	0.648
samples average	0.733	0.669	0.684

Results - Binary relevance Word2Vec

Best Algorithm - Random Forest Classifier; Accuracy - 0.54567

	Precision	Recall	F1-score
rating	0.796	0.804	0.800
bug report	0.760	0.765	0.762
feature request	0.653	0.154	0.249
user experience	0.718	0.124	0.211
micro average	0.776	0.611	0.683
macro average	0.732	0.462	0.506
weighted average	0.757	0.611	0.633
samples average	0.722	0.661	0.676

Results - Classifier Chains TF-IDF

Best Algorithm - SVC; Accuracy - 0.625833

	Precision	Recall	F1-score
rating	0.828	0.825	0.826
bug report	0.799	0.789	0.794
feature request	0.753	0.601	0.668
user experience	0.624	0.345	0.444
micro average	0.793	0.717	0.753
macro average	0.751	0.640	0.683
weighted average	0.781	0.717	0.742
samples average	0.815	0.760	0.772

Results - Classifier Chains Word2Vec

Best Algorithm - SVC; Accuracy - 0.5987

	Precision	Recall	F1-score
rating	0.824	0.818	0.812
bug report	0.745	0.761	0.758
feature request	0.747	0.438	0.541
user experience	0.629	0.408	0.497
micro average	0.702	0.608	0.621
macro average	0.612	0.516	0.521
weighted average	0.673	0.638	0.628
samples average	0.725	0.679	0.729

Results - Label Powerset TF-IDF

Best Algorithm - SVC; Accuracy - 0.5677

	Precision	Recall	F1-score
rating	0.828	0.825	0.826
bug report	0.799	0.789	0.762
feature request	0.653	0.154	0.249
user experience	0.718	0.124	0.211
micro average	0.701	0.598	0.617
macro average	0.615	0.512	0.516
weighted average	0.678	0.632	0.623
samples average	0.721	0.675	0.725

Results - Label Powerset Word2Vec

Best Algorithm - RandomForestClassifier; Accuracy - 0.56788

	Precision	Recall	F1-score
rating	0.816	0.775	0.795
bug report	0.667	0.850	0.748
feature request	0.476	0.240	0.319
user experience	0.538	0.186	0.276
micro average	0.719	0.604	0.677
macro average	0.624	0.513	0.535
weighted average	0.692	0.640	0.646
samples average	0.739	0.691	0.701

Conclusions

Conclusions and Future Work

The best results were obtained with the adaptive algorithm ridge classifier.

We could not get the results we hoped for, obtaining a generally low f1-score for all approaches tried. We believed it is mostly because of class imbalance. With more time we would have tried SMOTE or class weight to combat this.

We also wish we could have spent more time in error analysis to understand what went wrong and how we could have fixed it.

We think our approach should have favoured understanding the problem and the best solutions to try out instead of trying several approaches, in the future we will keep this lesson in mind.

Perguntas