IMT Mines Alès
École Mines-Télécom

EuroMov
DIGITAL HEALTH IN MOTION

Mining Bilingual App Reviews with Pre-Trained Models and ChatGPT

Jialiang Wei, Anne-Lise Courbis, Thomas Lambolais, Binbin Xu, Pierre Louis Bernard, Gérard Dray

GDR GPL, Rennes, June 6th, 2023

CONTENTS

1. MOTIVATION

2. APPROACH AND EVALUATION

- Classification
- Clustering
- Summarization
- Ranking

3. LIMITATIONS

4. CONCLUSION

IMT Mines Alès
École Mines-Télécom

CONTENTS

1. MOTIVATION

2. APPROACH AND EVALUATION

- Classification
- Clustering
- Summarization
- Ranking

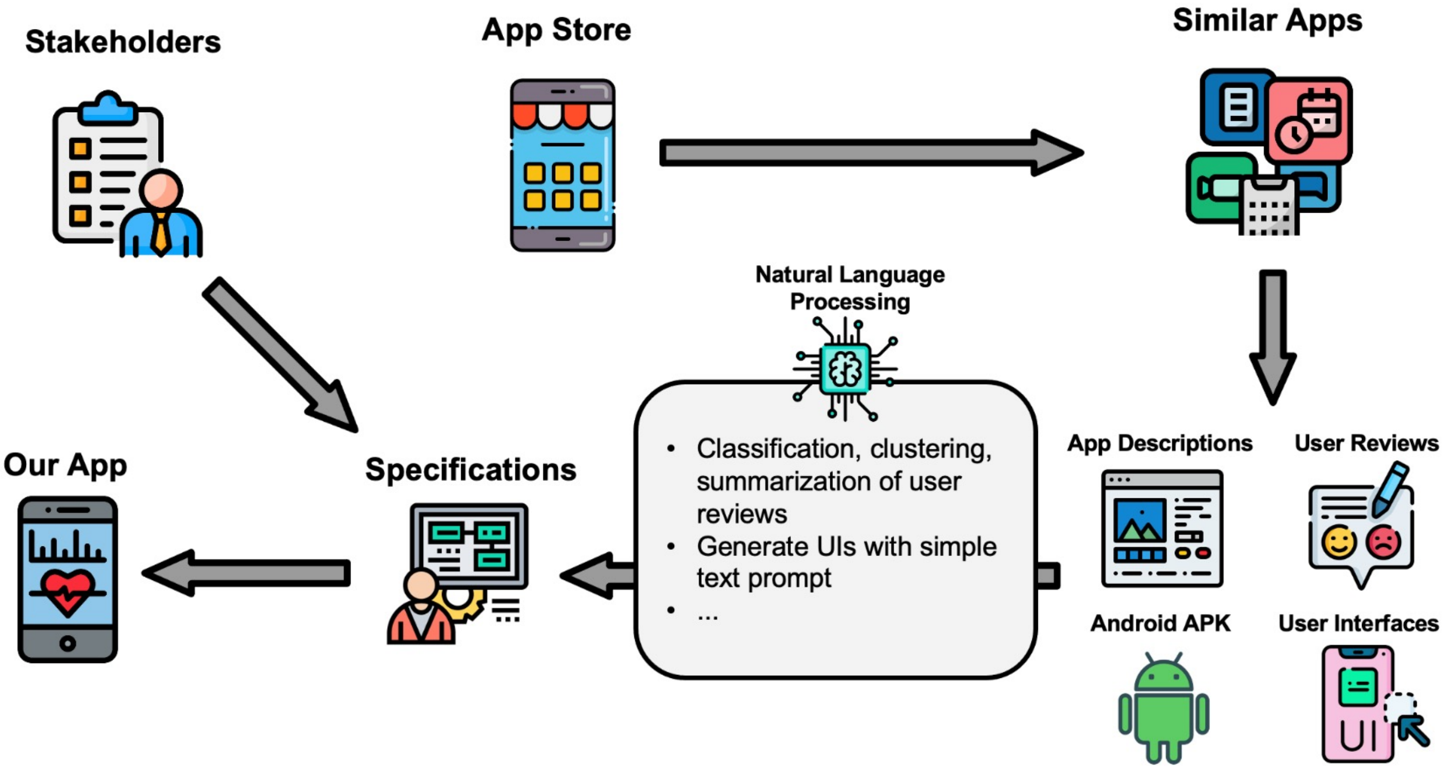
3. LIMITATIONS

4. CONCLUSION

IMT Mines Alès
École Mines-Télécom

MOTIVATION

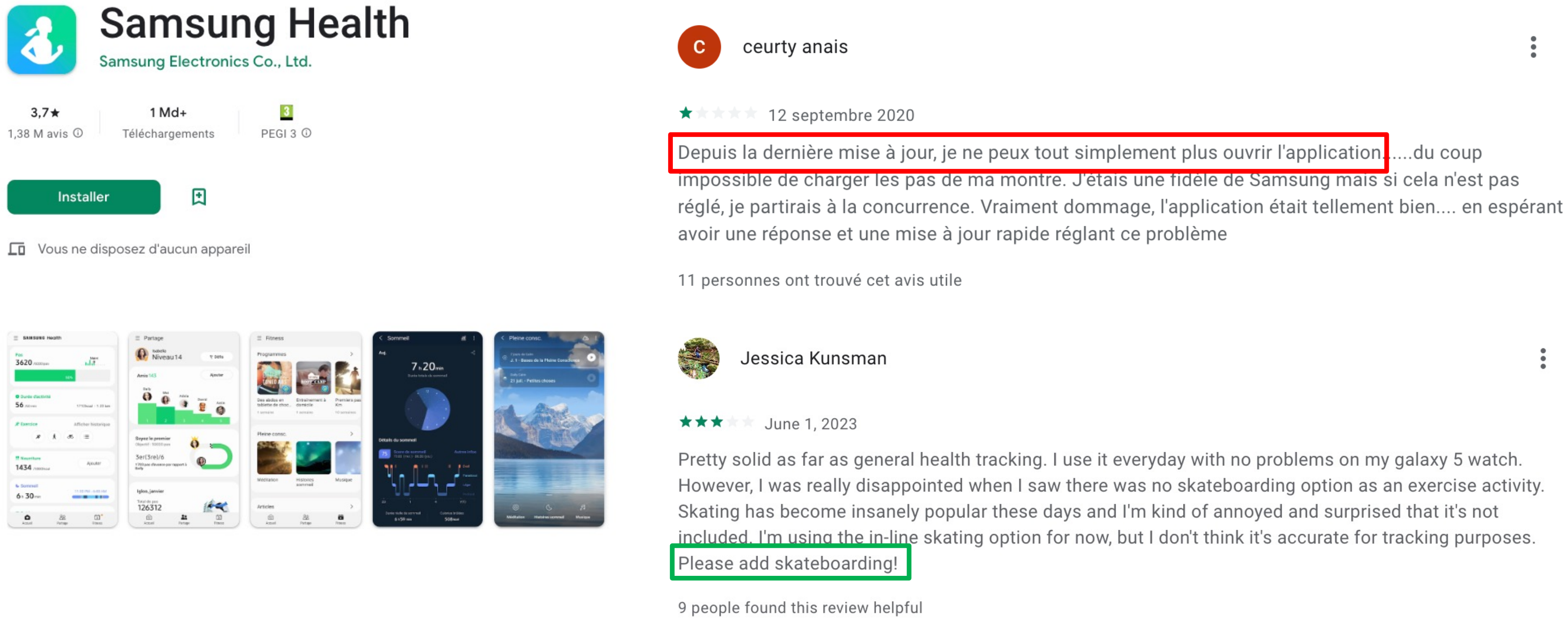
Data-driven requirements engineering



- Classification, clustering, summarization of user reviews
- Generate UIs with simple text prompt
- ...

MOTIVATION

Example of app reviews



Samsung Health

Samsung Electronics Co., Ltd.

3,7★ 1,38 M avis ⓘ | 1 Md+ Téléchargements | PEGI 3 ⓘ

Installer

Vous ne disposez d'aucun appareil

ceurty anais

★☆☆☆☆ 12 septembre 2020

Depuis la dernière mise à jour, je ne peux tout simplement plus ouvrir l'application.....du coup impossible de charger les pas de ma montre. J'étais une fidèle de Samsung mais si cela n'est pas réglé, je partirais à la concurrence. Vraiment dommage, l'application était tellement bien.... en espérant avoir une réponse et une mise à jour rapide réglant ce problème

11 personnes ont trouvé cet avis utile

Jessica Kunsman

★★★☆☆ June 1, 2023

Pretty solid as far as general health tracking. I use it everyday with no problems on my galaxy 5 watch. However, I was really disappointed when I saw there was no skateboarding option as an exercise activity. Skating has become insanely popular these days and I'm kind of annoyed and surprised that it's not included. I'm using the in-line skating option for now, but I don't think it's accurate for tracking purposes. Please add skateboarding!

9 people found this review helpful

https://play.google.com/store/apps/details?id=com.sec.android.app.shealth

MOTIVATION

Pre-Trained Models and ChatGPT

- Pre-Trained Models (PTMs)
  - PTMs are neural networks that have been previously trained over a large corpus
  - PTM can be employed to generate contextual word embeddings for text, or alternatively fine-tuned for specific downstream tasks, like classification
  - Example: GPT [1], BERT [2], CamemBERT [3], XLM-R [4]
- ChatGPT
  - It is fine-tuned from GPT-3.5 [5] using Reinforcement Learning from Human Feedback (RLHF) [6]
  - ChatGPT has achieved a state-of-the-art performance in crosslingual summarization [7]

[1] A. Radford and K. Narasimhan, “Improving Language Understanding by Generative Pre-Training,” 2018.

[2] J. Devlin et al., “BERT: Pretraining of deep bidirectional transformers for language understanding,” in NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, vol. 1, oct 2019, pp. 4171–4186.

[3] L. Martin et al., “CamemBERT: a Tasty French Language Model,” in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 7203–7219.

[4] S. Ruder, A. Søgaard, and I. Vulic, “Unsupervised cross-lingual representation learning,” in ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Tutorial Abstracts, nov 2019, pp. 31–38.

[5]T. Brown et al., “Language Models are Few-Shot Learners,” in Advances in Neural Information Processing Systems, 2020, vol. 33, pp. 1877–1901

[6]P. F. Christiano et al., “Deep reinforcement learning from human preferences,” Advances in Neural Information Processing Systems, vol. 2017-December, pp. 4300–4308, 2017.

[7] J. Wang et al., “CrossLingual Summarization via ChatGPT,” 2023. [Online]. Available: http://arxiv.org/abs/2302.14229

CONTENTS

1. MOTIVATION

2. APPROACH AND EVALUATION

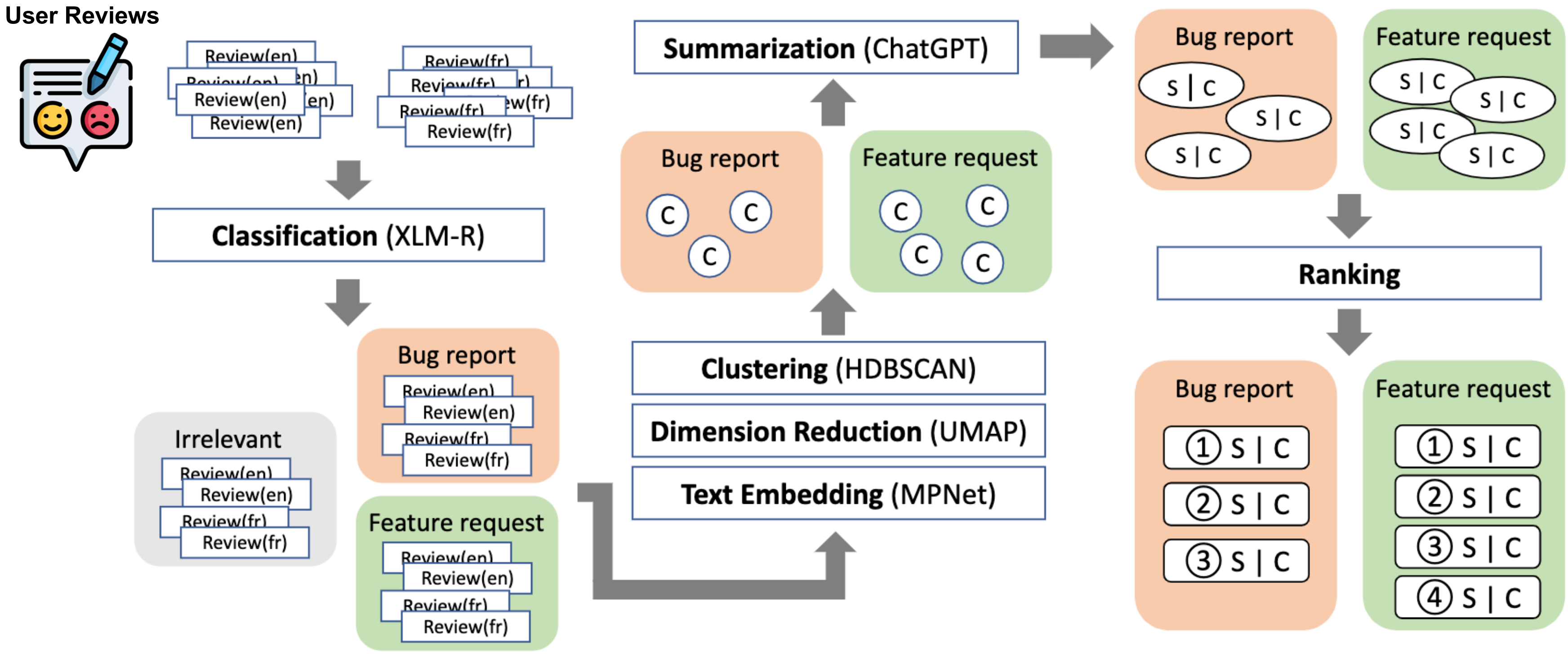
- Classification
- Clustering
- Summarization
- Ranking

3. LIMITATIONS

4. CONCLUSION

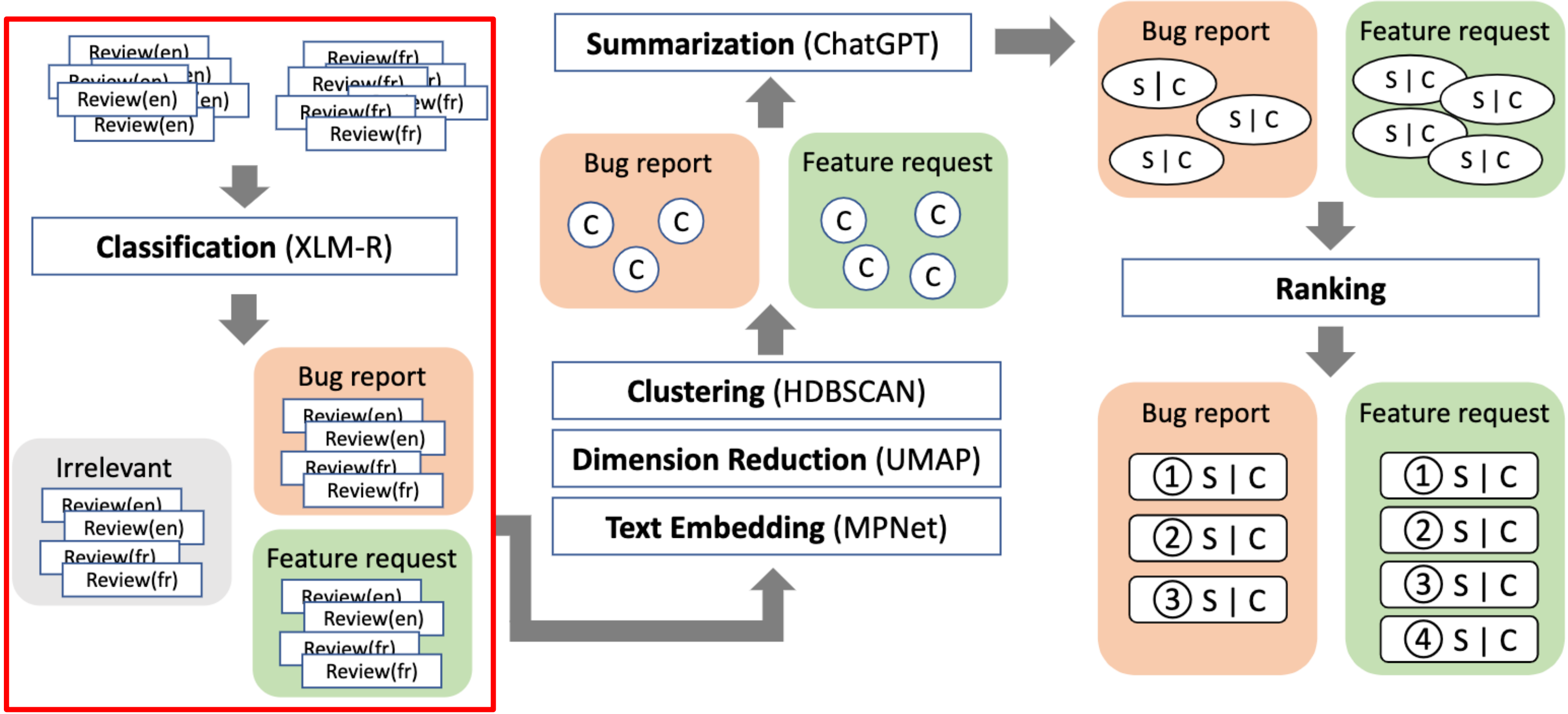
OVERVIEW OF OUR APPROACH

Mini-BAR



MINI-BAR

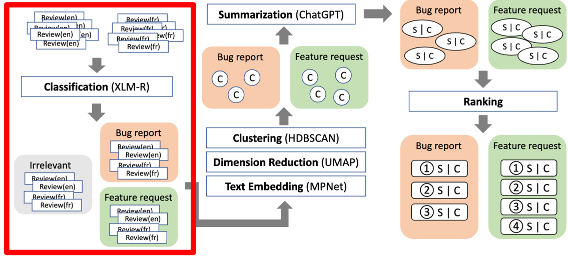
Classification



CLASSIFICATION

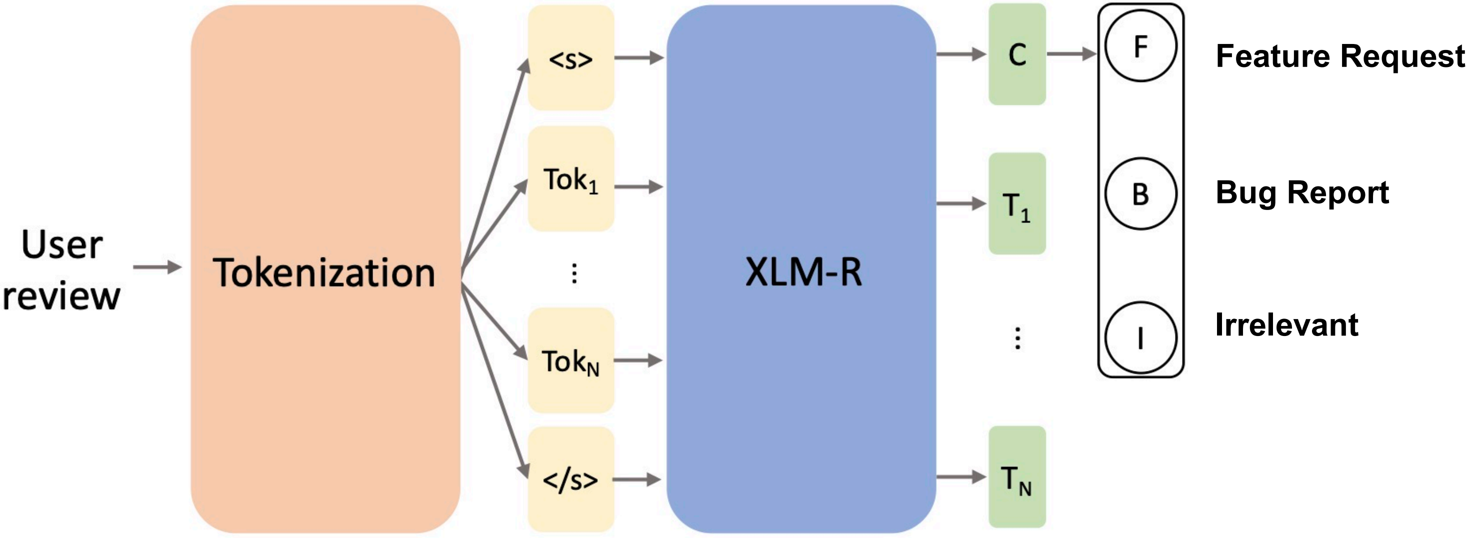
Overview

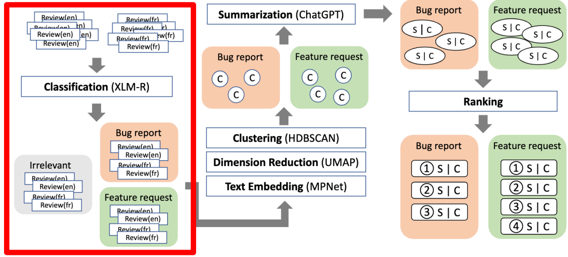
- Classify automatically the user review into three categories:
  - Feature request
    - “Please bring a feature to add some custom watch faces”
  - Bug report
    - “Can’t sync sleep data since last update”
  - Irrelevant
    - “Best app ever! ”



CLASSIFICATION

Method

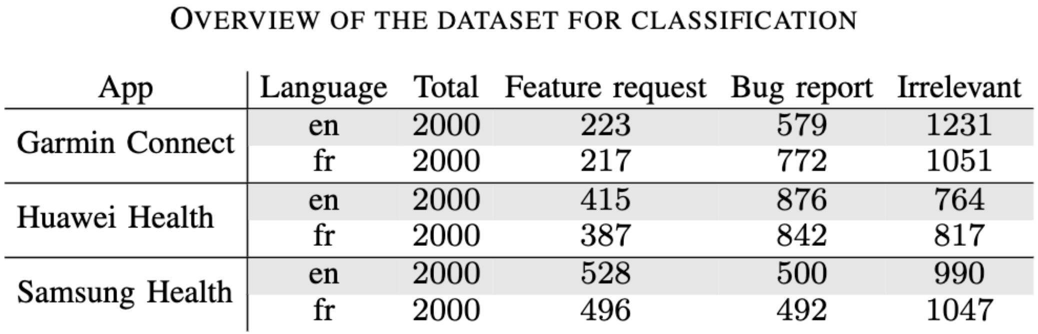




S. Ruder, A. Søgaard, and I. Vulic, “Unsupervised cross-lingual representation learning,” in ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Tutorial Abstracts, Nov. 2019, pp. 31–38, doi: 10.18653/v1/p19-4007.

CLASSIFICATION

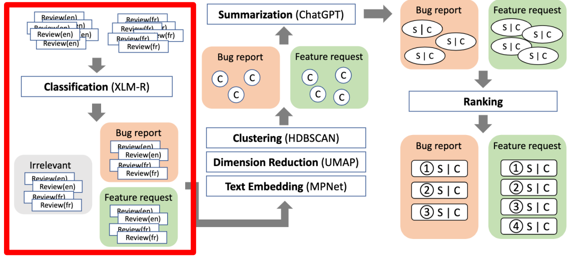
Dataset



OVERVIEW OF THE DATASET FOR CLASSIFICATION

App	Language	Total	Feature request	Bug report	Irrelevant
Garmin Connect	en	2000	223	579	1231
	fr	2000	217	772	1051
Huawei Health	en	2000	415	876	764
	fr	2000	387	842	817
Samsung Health	en	2000	528	500	990
	fr	2000	496	492	1047

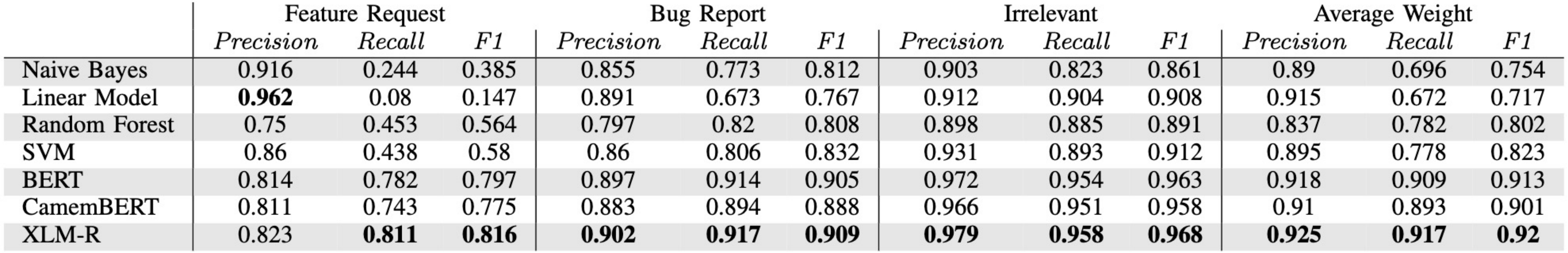
- Data split
  - 20% Test set
  - 80% Train set



CLASSIFICATION

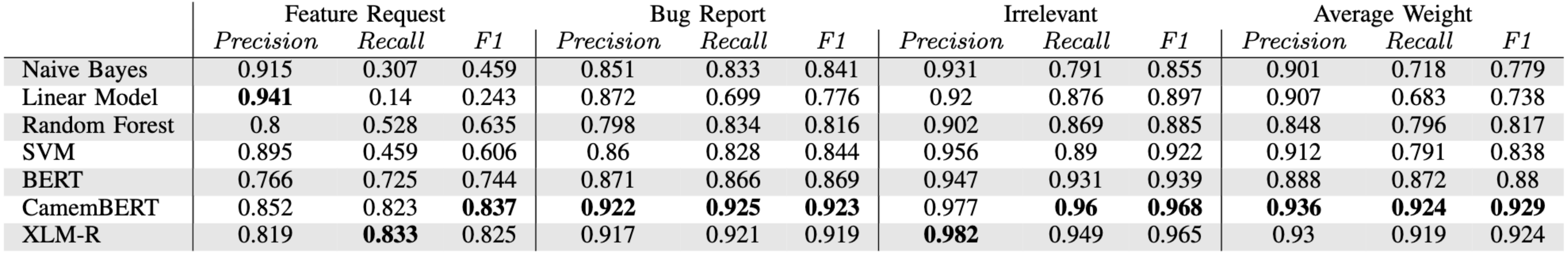
Results

Classification accuracy on English user reviews of three apps

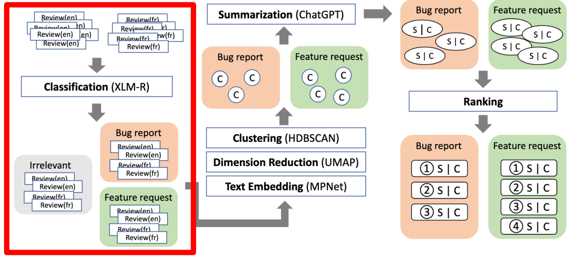


	Feature Request			Bug Report			Irrelevant			Average Weight		
	Precision	*Recall*	*F1*	*Precision*	*Recall*	*F1*	*Precision*	*Recall*	*F1*	*Precision*	*Recall*	*F1*
Naive Bayes	0.916	0.244	0.385	0.855	0.773	0.812	0.903	0.823	0.861	0.89	0.696	0.754
Linear Model	**0.962**	0.08	0.147	0.891	0.673	0.767	0.912	0.904	0.908	0.915	0.672	0.717
Random Forest	0.75	0.453	0.564	0.797	0.82	0.808	0.898	0.885	0.891	0.837	0.782	0.802
SVM	0.86	0.438	0.58	0.86	0.806	0.832	0.931	0.893	0.912	0.895	0.778	0.823
BERT	0.814	0.782	0.797	0.897	0.914	0.905	0.972	0.954	0.963	0.918	0.909	0.913
CamemBERT	0.811	0.743	0.775	0.883	0.894	0.888	0.966	0.951	0.958	0.91	0.893	0.901
XLM-R	0.823	**0.811**	**0.816**	**0.902**	**0.917**	**0.909**	**0.979**	**0.958**	**0.968**	**0.925**	**0.917**	**0.92**

Classification accuracy on French user reviews of three apps

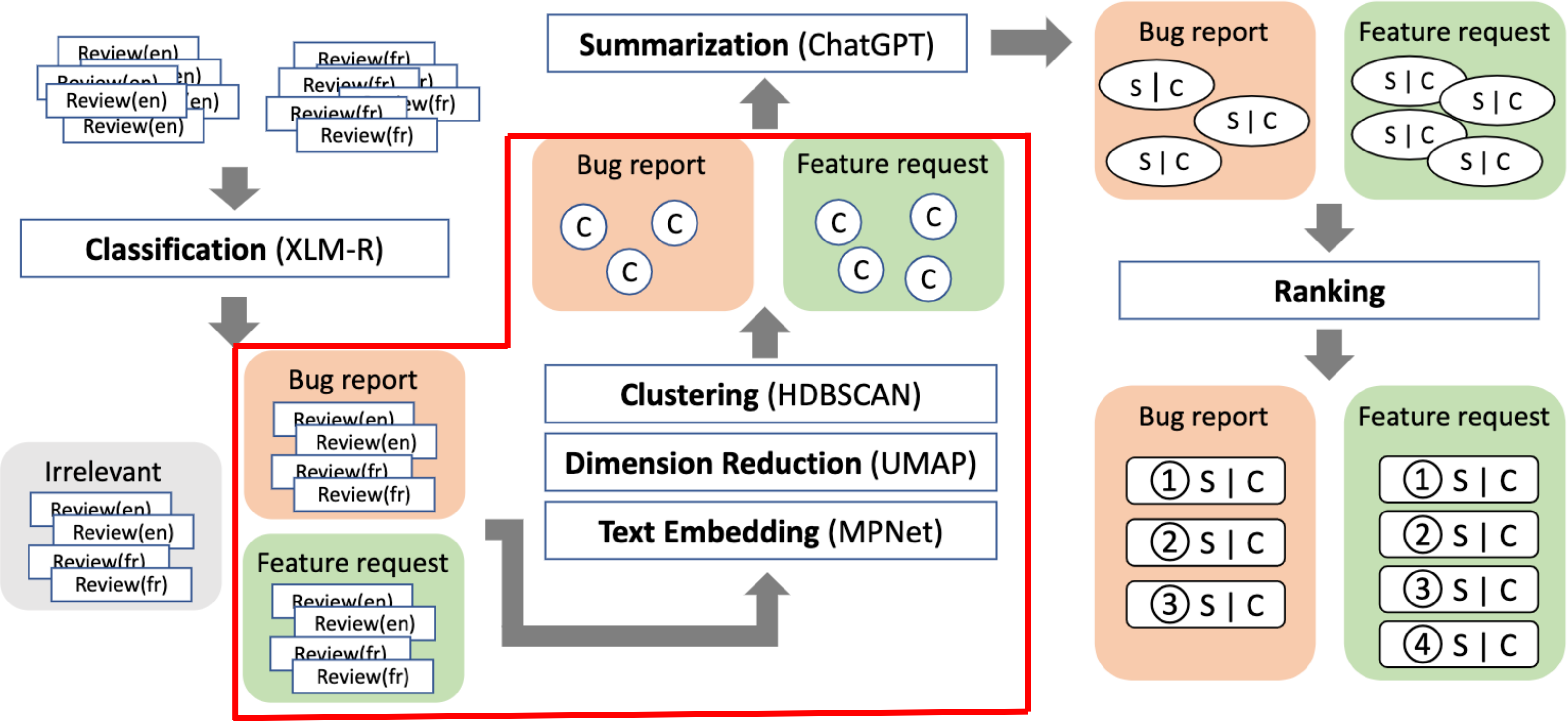


	Feature Request			Bug Report			Irrelevant			Average Weight		
	Precision	*Recall*	*F1*	*Precision*	*Recall*	*F1*	*Precision*	*Recall*	*F1*	*Precision*	*Recall*	*F1*
Naive Bayes	0.915	0.307	0.459	0.851	0.833	0.841	0.931	0.791	0.855	0.901	0.718	0.779
Linear Model	**0.941**	0.14	0.243	0.872	0.699	0.776	0.92	0.876	0.897	0.907	0.683	0.738
Random Forest	0.8	0.528	0.635	0.798	0.834	0.816	0.902	0.869	0.885	0.848	0.796	0.817
SVM	0.895	0.459	0.606	0.86	0.828	0.844	0.956	0.89	0.922	0.912	0.791	0.838
BERT	0.766	0.725	0.744	0.871	0.866	0.869	0.947	0.931	0.939	0.888	0.872	0.88
CamemBERT	0.852	0.823	**0.837**	**0.922**	**0.925**	**0.923**	0.977	**0.96**	**0.968**	**0.936**	**0.924**	**0.929**
XLM-R	0.819	**0.833**	0.825	0.917	0.921	0.919	**0.982**	0.949	0.965	0.93	0.919	0.924



MINI-BAR

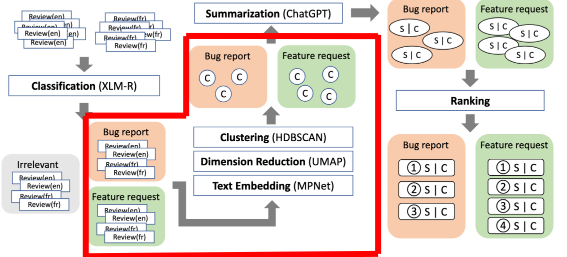
Clustering



CLUSTERING

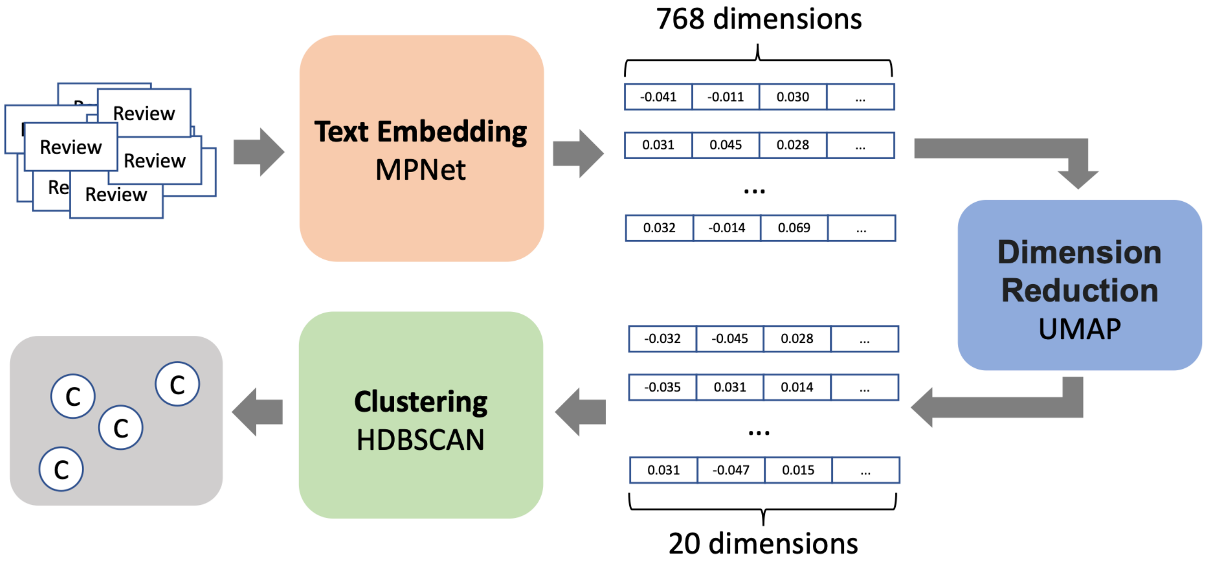
Overview

- Clustering the user reviews based on their semantic similarity:
  - Example of a cluster about “offline”
    - “Dommage que la connexion 4g soit indispensable pour fonctionner. ”
    - “Please for god sake make it to work offline also. ”
    - “Is not work offline.”
    - “It used to work offline. Now I have to log in just to see my old data. ”
    - “Useless without internet.”
  - Example of a cluster about ”French language”
    - “Est il possible de mettre l'application en français ? Car elle est en anglais”
    - “Est il possible de l'avoir en français ? MERCI”
    - “Notices d'information : en français c'est possible ?”

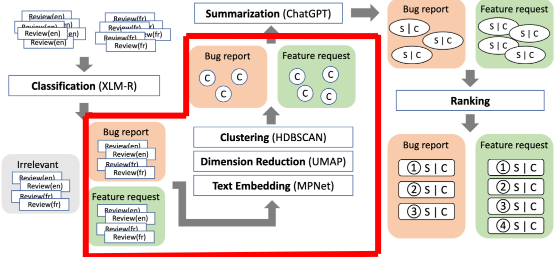


CLUSTERING

Method



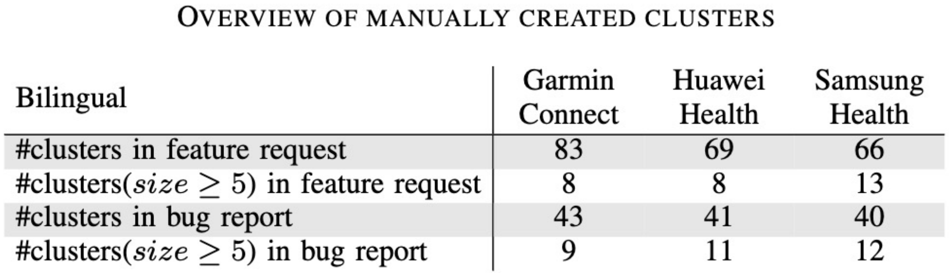
K. Song, X. Tan, T. Qin, J. Lu, and T. Y. Liu, “MPNet: Masked and permuted pre-training for language understanding,” Advances in Neural Information Processing Systems, vol. 2020-Decem, no. NeurIPS, pp. 114, 2020.
L. McInnes, J. Healy, and J. Melville, “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction,” *arXiv preprint arXiv:1802.03426*, feb 2018.
L. McInnes and J. Healy, “Accelerated Hierarchical Density Based Clustering,” in 2017 IEEE International Conference on Data Mining Workshops (ICDMW), 2017, pp. 33–42.



CLUSTERING

Dataset

- Dataset



OVERVIEW OF MANUALLY CREATED CLUSTERS

Bilingual	Garmin Connect	Huawei Health	Samsung Health
#clusters in feature request	83	69	66
#clusters($size \geq 5$) in feature request	8	8	13
#clusters in bug report	43	41	40
#clusters($size \geq 5$) in bug report	9	11	12

CLUSTERING

Result

- Evaluation Metric: V-measure
  - Quantify the similarity between the clustering results and the ground truth

- Result

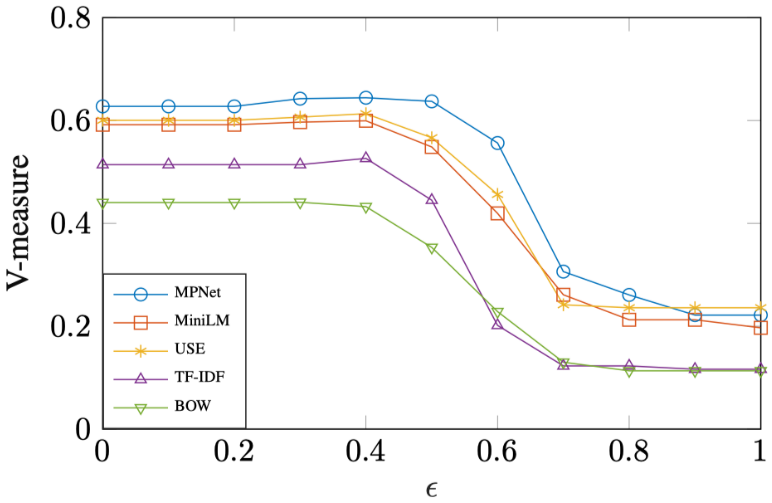
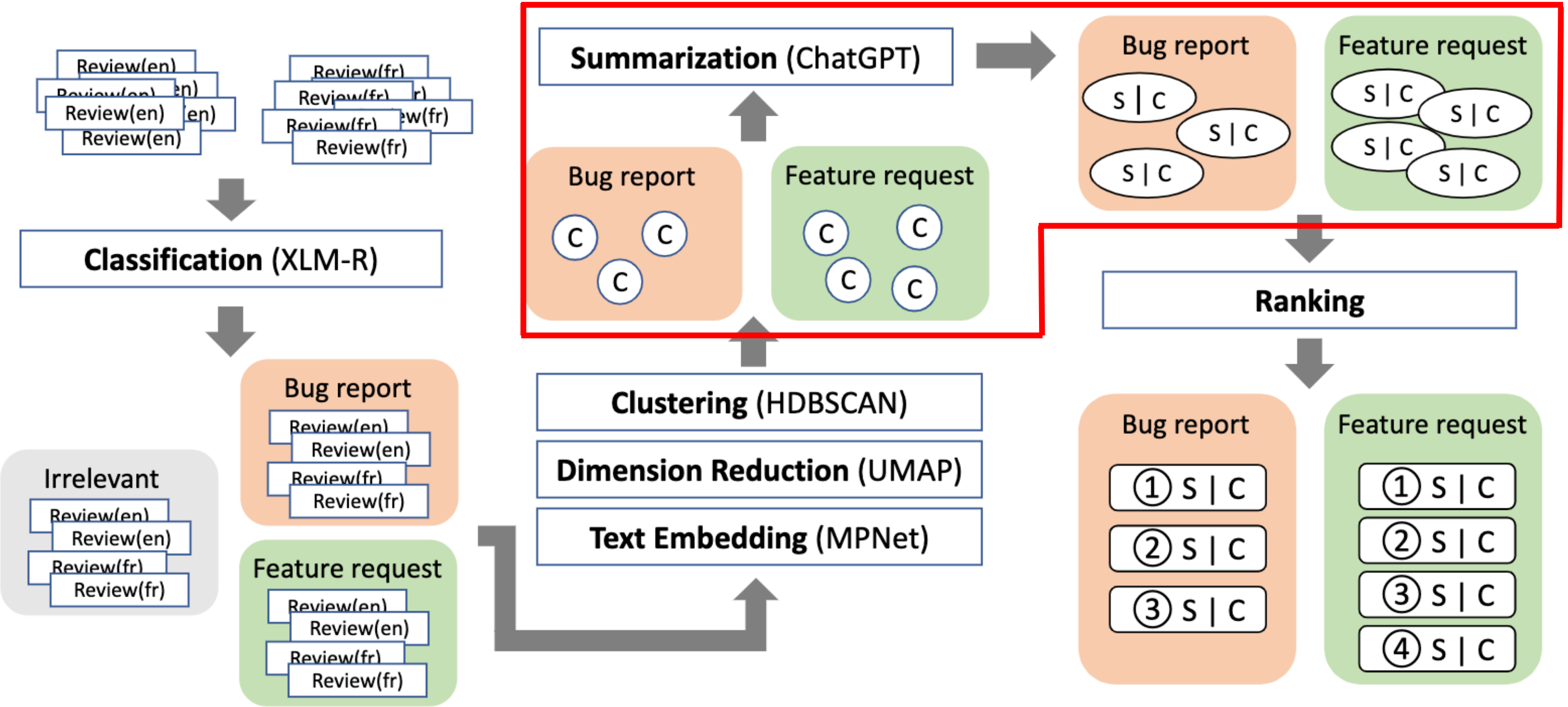


Fig. 7. V-measure score on bilingual user reviews

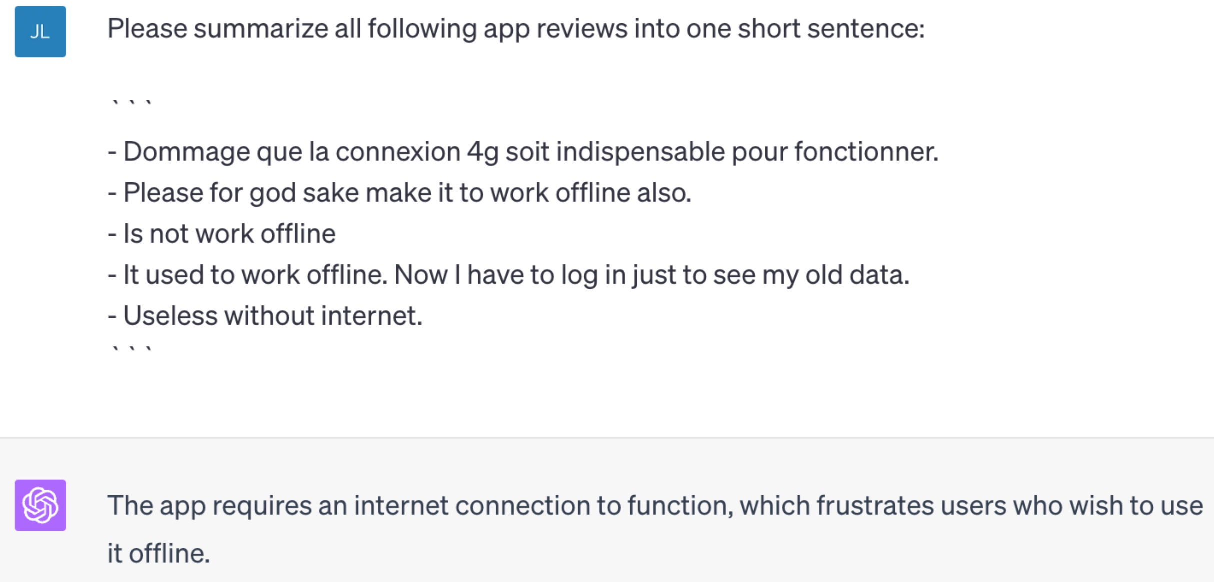
MINI-BAR

Summarization



SUMMARIZATION

Overview



JL

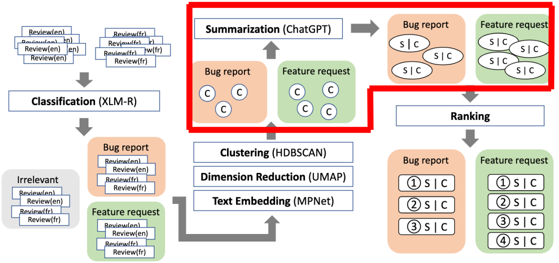
Please summarize all following app reviews into one short sentence:

` ` `

- Dommage que la connexion 4g soit indispensable pour fonctionner.
- Please for god sake make it to work offline also.
- Is not work offline
- It used to work offline. Now I have to log in just to see my old data.
- Useless without internet.

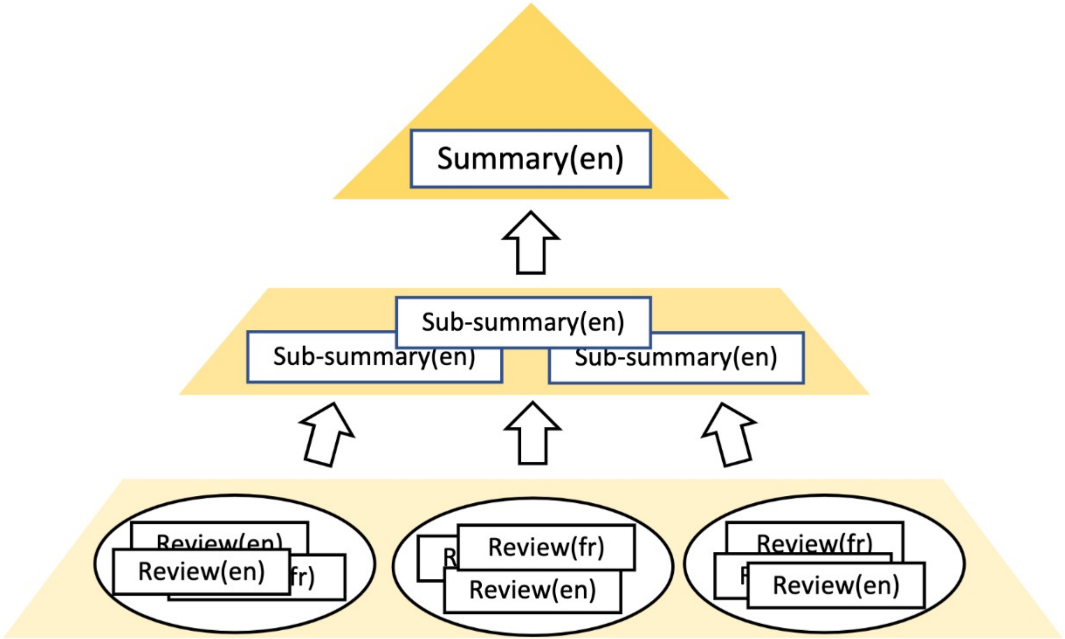
` ` `

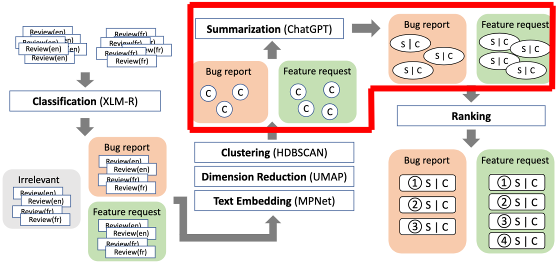
The app requires an internet connection to function, which frustrates users who wish to use it offline.



SUMMARIZATION

Method



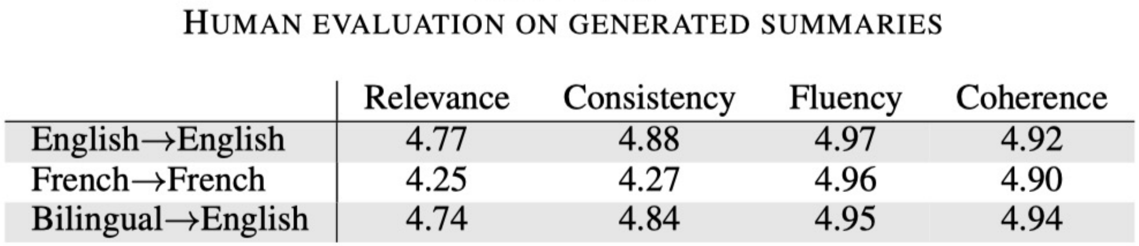


SUMMARIZATION

Evaluation

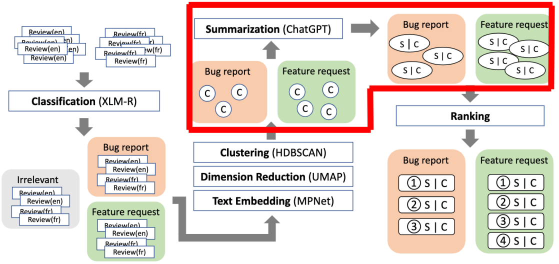
- Evaluation Metric
  - **Relevance -** selection of important content from the source
  - **Consistency** - the factual alignment between the summary and the summarized source
  - **Fluency** - the quality of individual sentences
  - **Coherence** - the collective quality of all sentences

- Results



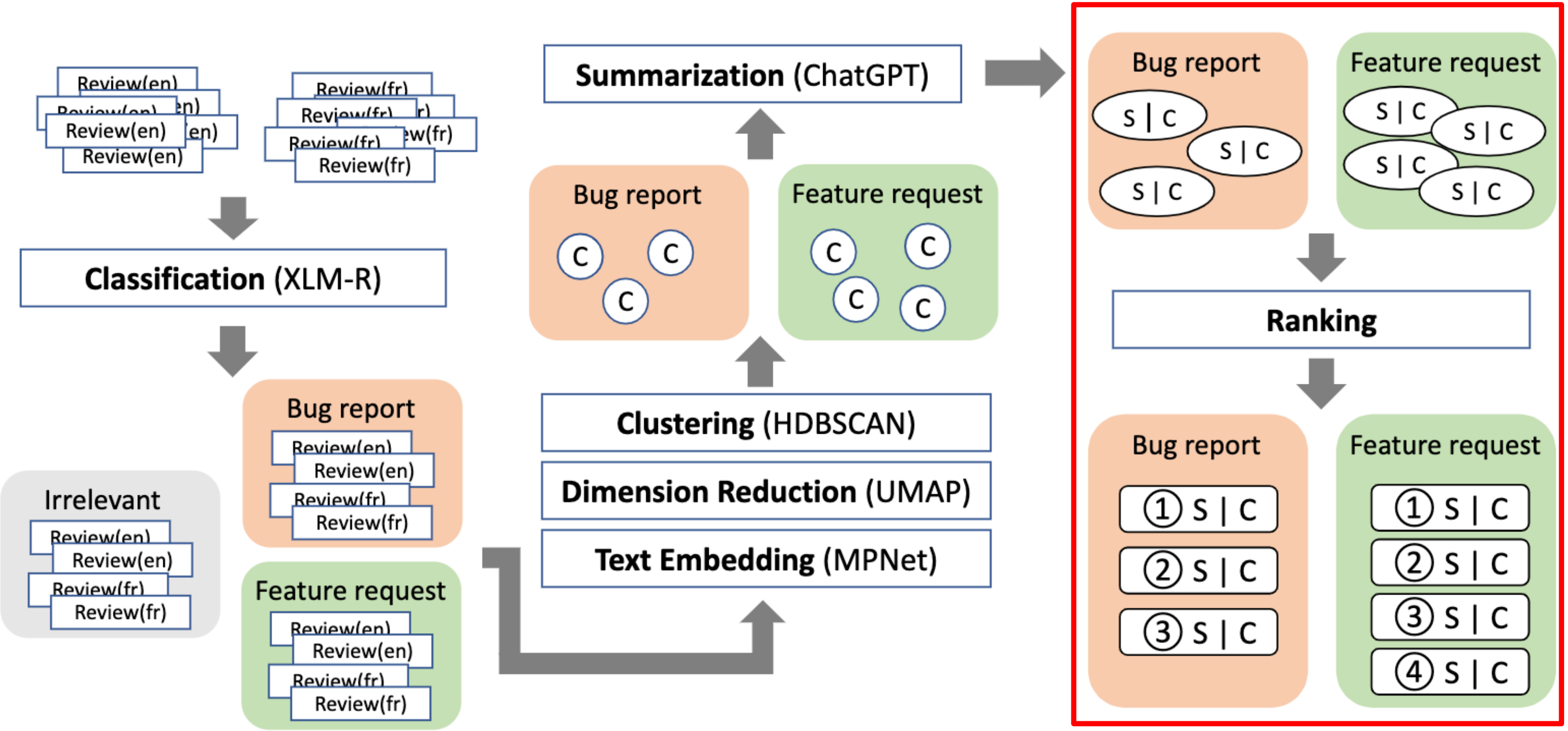
HUMAN EVALUATION ON GENERATED SUMMARIES

	Relevance	Consistency	Fluency	Coherence
English→English	4.77	4.88	4.97	4.92
French→French	4.25	4.27	4.96	4.90
Bilingual→English	4.74	4.84	4.95	4.94



MINI-BAR

Ranking

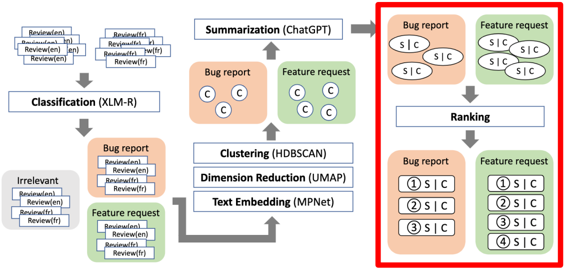


RANKING

Overview

- Rank the user review clusters by:
  - $|reviews|$ Quantity of user reviews
  - $|thumbsup|$ Thumbs up number
  - $\overline{rating}$ Average rating

$$ClusterScore = \frac{w_{rev} \cdot |reviews| + w_{th} \cdot |thumbsup|}{w_{ra} \cdot \overline{rating}}$$



CONTENTS

1. MOTIVATION

2. APPROACH AND EVALUATION

    - Classification
    - Clustering
    - Summarization
    - Ranking

3. LIMITATIONS

4. CONCLUSION

LIMITATIONS

- App reviews from one category
  - Garmin Connect, Huawei Health, Samsung Health
- Issues of using ChatGPT
  - Data privacy
  - Cost
  - Availability

CONTENTS

1. MOTIVATION

2. APPROACH AND EVALUATION

   - Classification
   - Clustering
   - Summarization
   - Ranking

3. LIMITATIONS

4. CONCLUSION

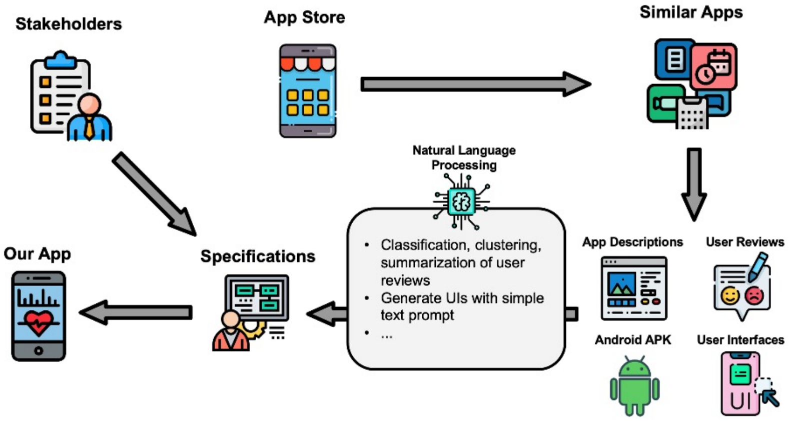
IMT Mines Alès
École Mines-Télécom

CONCLUSION

- Mini-BAR (Mining Bilingual App Reviews):
  - Classify reviews with an F1 score of 0.92
  - Create meaningful clusters of reviews with a V-measure greater than 0.64
  - Produce highly satisfactory summaries of reviews

- Next step:
  - Support more languages
  - Train with user reviews from apps in various categories
  - Perform classification and clustering at sentence level rather than review level
  - Employ alternative large language models for summarization tasks.

ANY QUESTIONS ?



Stakeholders

App Store

Similar Apps

Natural Language Processing

- Classification, clustering, summarization of user reviews
- Generate UIs with simple text prompt
- ...

App Descriptions

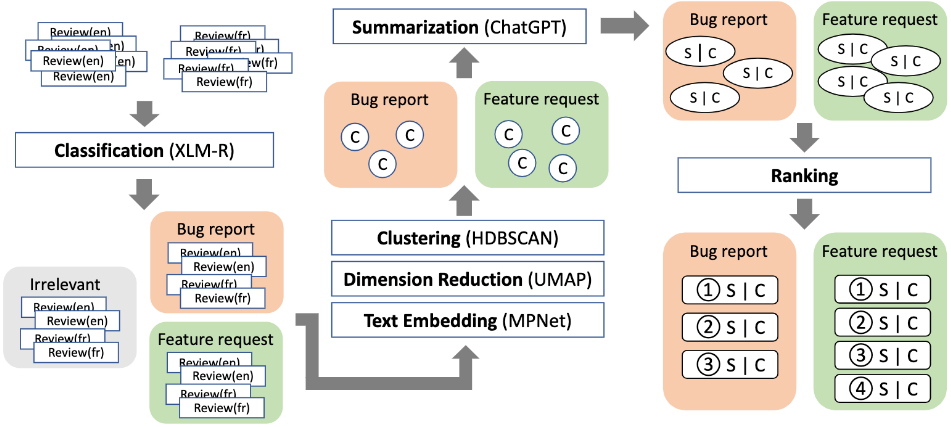
User Reviews

Android APK

User Interfaces

Our App

Specifications



Review(en) Review(en) Review(en) Review(en)

Review(fr) Review(fr) Review(fr) Review(fr)

Classification (XLM-R)

Irrelevant: Review(en) Review(en) Review(fr) Review(fr)

Bug report: Review(en) Review(en) Review(fr) Review(fr)

Feature request: Review(en) Review(en) Review(fr) Review(fr)

Text Embedding (MPNet)

Dimension Reduction (UMAP)

Clustering (HDBSCAN)

Bug report: C C C

Feature request: C C C C

Summarization (ChatGPT)

Bug report: S | C S | C S | C

Feature request: S | C S | C S | C S | C

Ranking

Bug report: ① S | C ② S | C ③ S | C

Feature request: ① S | C ② S | C ③ S | C ④ S | C

- J. Wei, A.-L. Courbis, T. Lambolais, B. Xu, P. L. Bernard, and G. Dray, “Towards a Data-Driven Requirements Engineering Approach: Automatic Analysis of User Reviews,” in APIA 2022 - 7e Conférence Nationale sur les Applications Pratiques de l’Intelligence Artificielle, Juin 2022, Saint-Étienne, France.
- J. Wei, A.-L. Courbis, T. Lambolais, P. L. Bernard, and G. Dray, “Towards Boosting Requirements Engineering of a Health Monitoring App by Analysing Similar Apps: A Vision Paper,” in 2022 IEEE 30th International Requirements Engineering Conference Workshops (REW), 2022, pp. 75–80, doi: 10.1109/REW56159.2022.00020.
- J. Wei, A.-L. Courbis, T. Lambolais , B. Xu, P. L. Bernard, and G. Dray, “Boosting GUI Prototyping with Diffusion Models”, accepted by IEEE 31th International Requirements Engineering Conference (RE), 2023
- J. Wei, A.-L. Courbis, T. Lambolais , B. Xu, P. L. Bernard, and G. Dray, “Mining Bilingual App Reviews with Pre-Trained Models and ChatGPT”, under review

Thank you

Jialiang WEI
jiailang.wei@mines-ales.fr

IMT Mines Alès
École Mines-Télécom