raid: Jérémie Dentan, Paul Théron, Louis Gautier

INF582 Challenge: Summary Source Prediction

Oral Presentation

Content

I- Introduction: task presentation

Summary source prediction

New tropical fruit to go on sale in the UK is cross between mango and plum.
The Bouea macrophylla - or mango plum - has been nicknamed the plango.
The fruit has a bright orange edible skin and a sweet taste and soft texture.

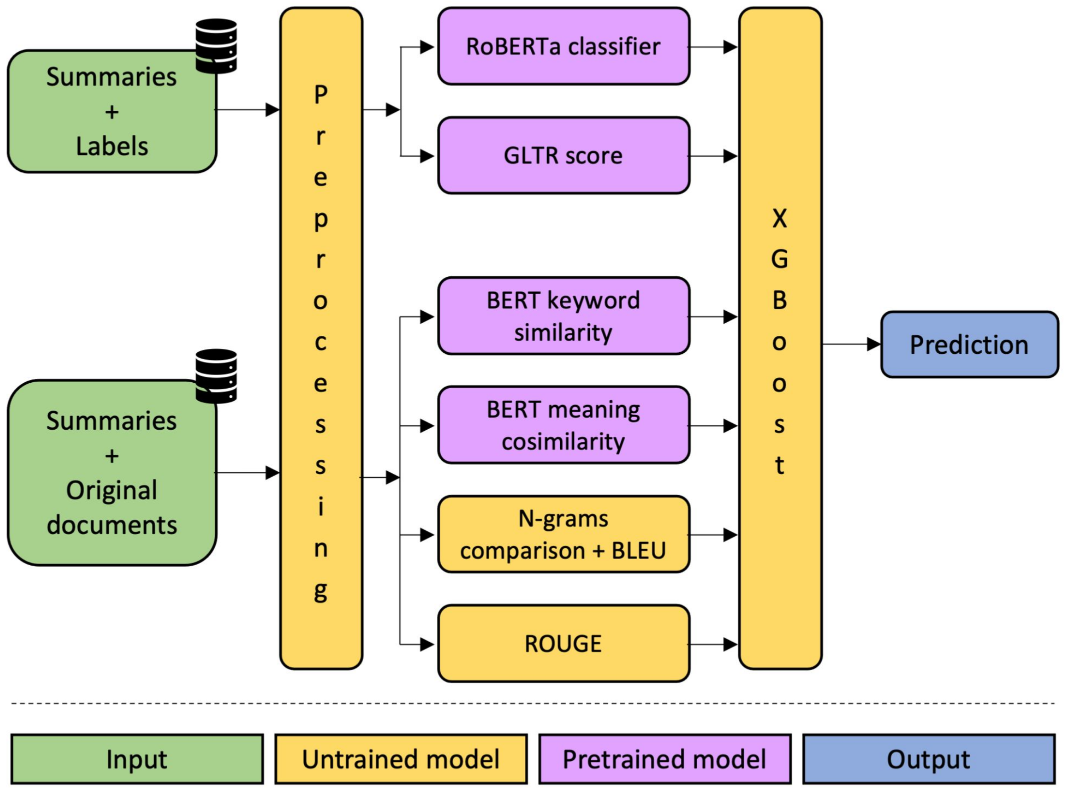
Patrick Cherry will appear on NBC New York on Friday night to apologize.
He has been stripped of his badge, will be placed on desk duty before being
transferred out of the NYPD 's Joint Terrorism Task Force division.

II- Our pipeline

Two types of features :

- Features from summaries only
- Features comparing the summaries to the original documents

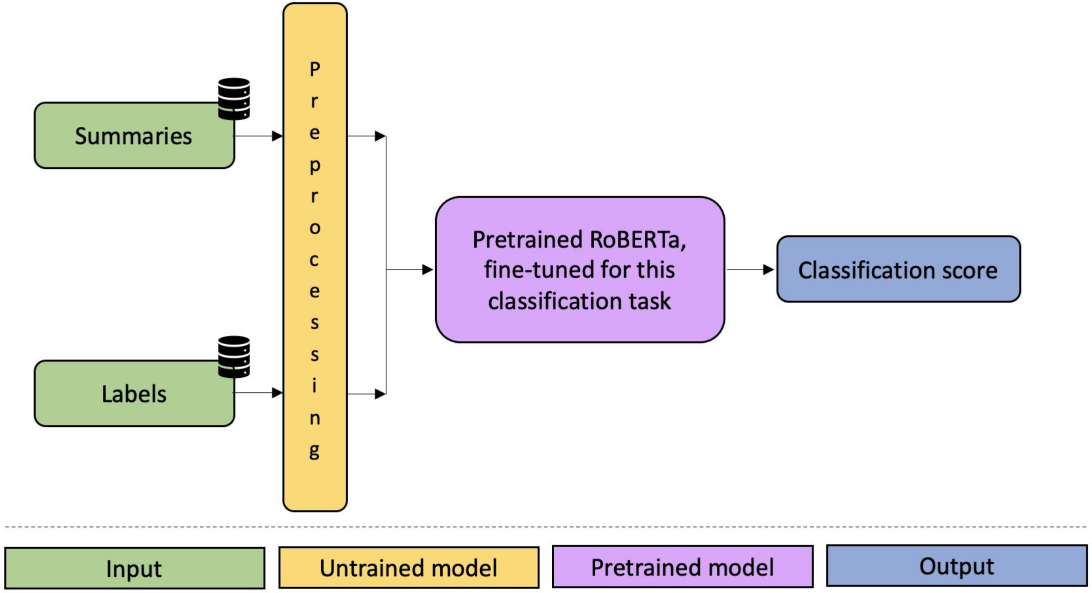
Due to the small size of our datasets, we made extensive use of pretrained models.



III- Summaries features: RoBERTa for classif.

This is the core component of our classifier (score of 0.89 alone).

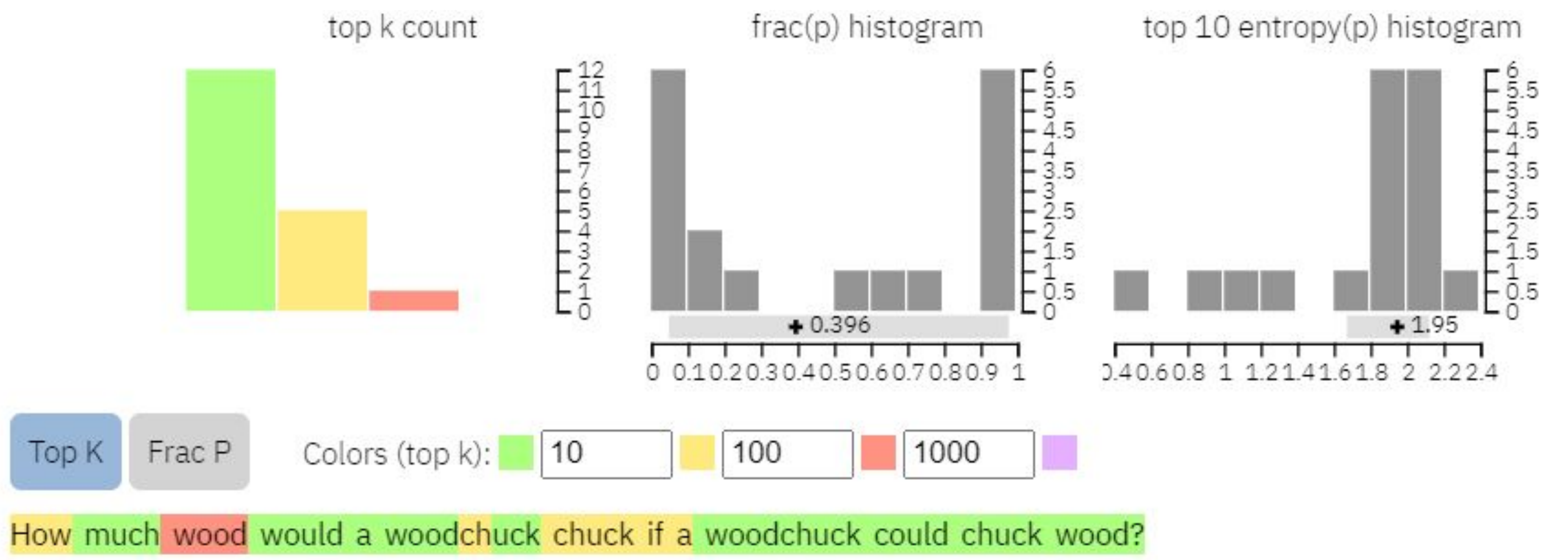
- We could have fine-tuned RoBERTa on the document dataset we had.
- Since this model is really performant, we only did a small preprocessing.



Ref: [2, 3]

III- Summaries features: GLTR

GLTR principle



Ref: [5]

III- Summaries features: GLTR

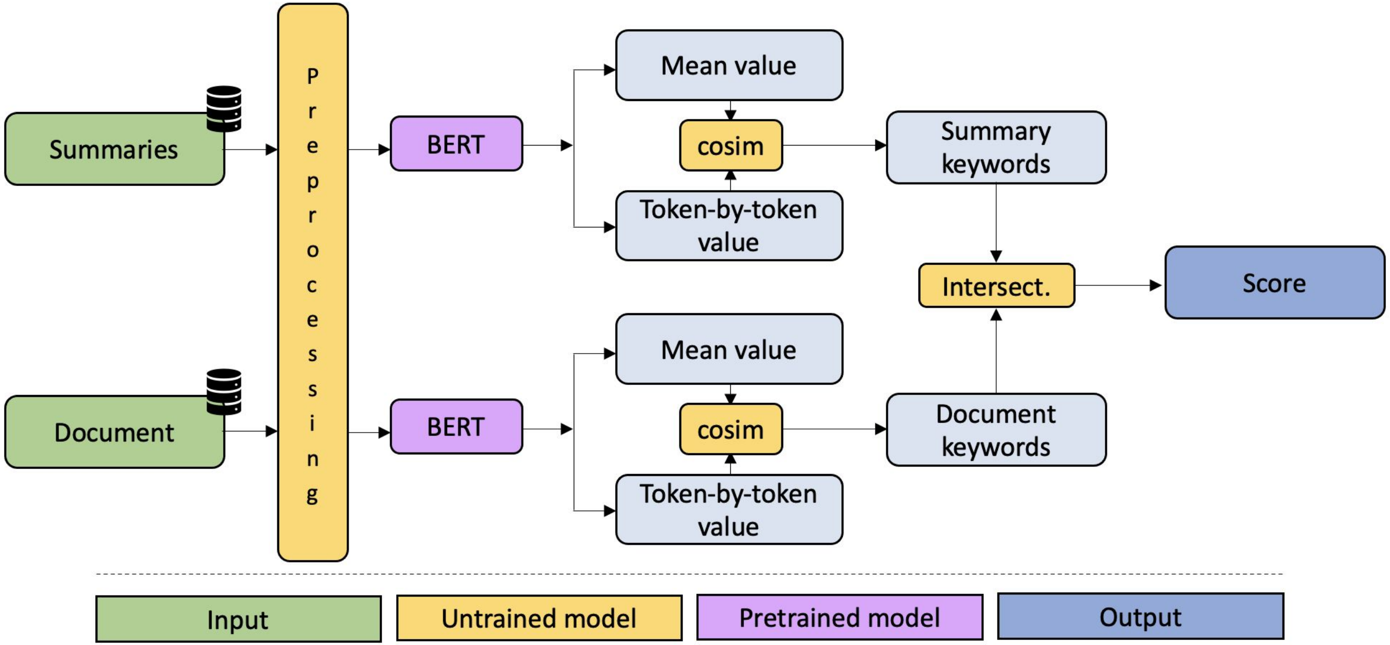
Features we took in GLTR:

- Maximum k in the text
- Average p and k, Standard deviation of the distribution of p and k
- Histogram of k with fixed bounds
- Histogram of p with evenly-distributed bounds and its bounds
- Histogram of entropy with evenly-distributed bounds and its bounds

If we had more texts in the train set:
1D-convolutions would have been adapted

Ref: [5]

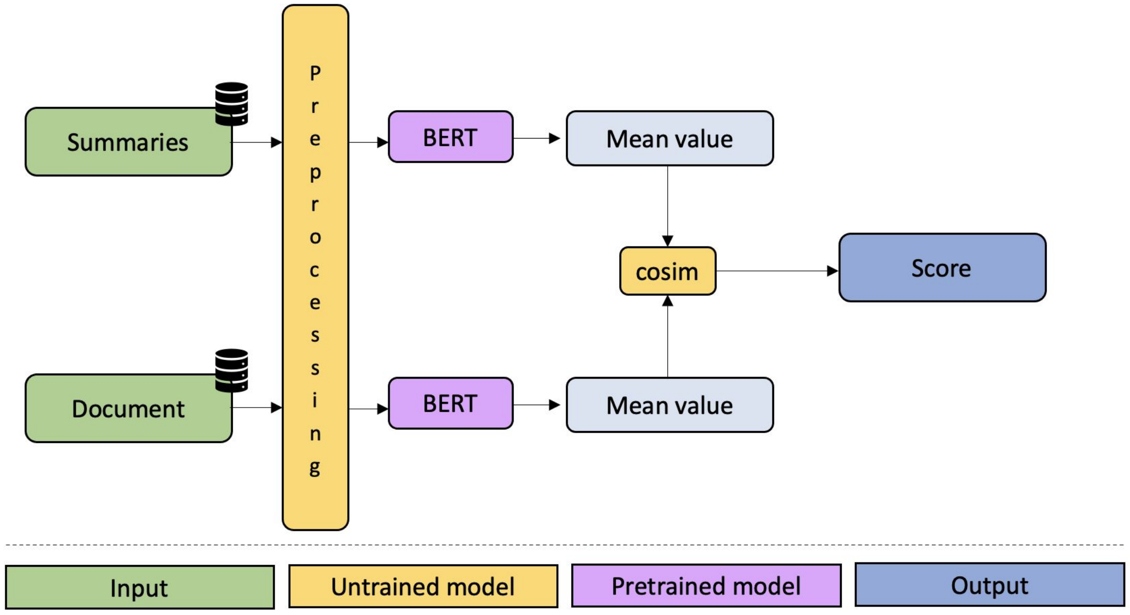
IV- Comparison features: keywords



But this did not really improved the performance…

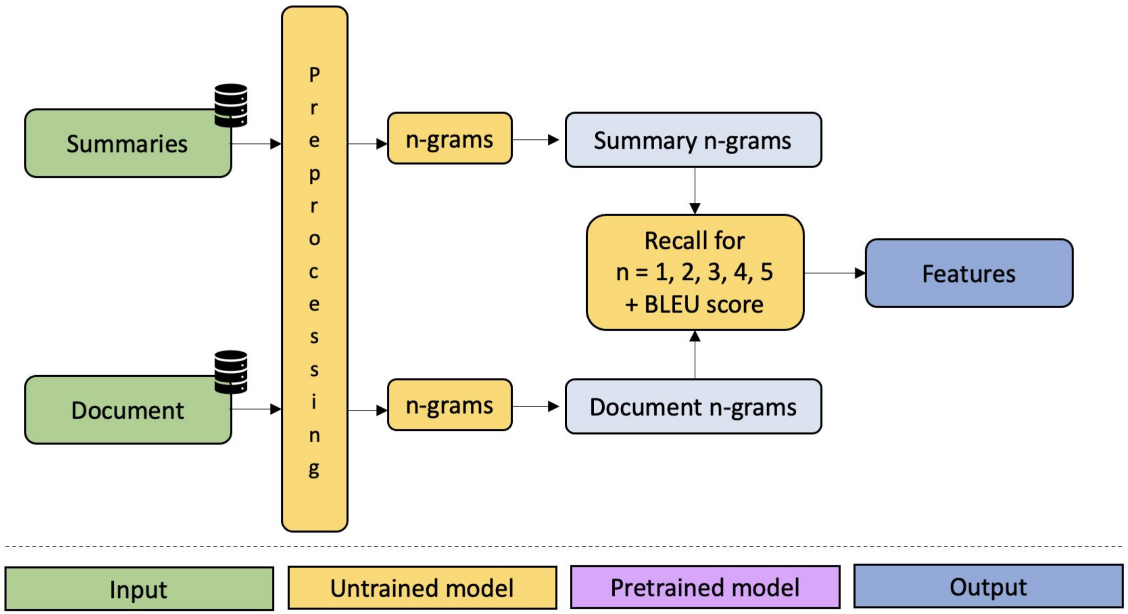
Ref: [7]

IV- Comparison features: meaning with BERT



This improved our performance: 0.90687 to 0.91000

IV- Comparison features: n-grams and BLEU



$$BLEU = \exp\left(\frac{1}{4}\sum_{i=1}^{4}\log(p_i)\right)$$

This significantly improved our performance: 0.91000 to 0.93312

Ref: [8]

IV- Comparison features: ROUGE score

ROUGE: a recall, precision and f1-score on n-grams.

- A score originally designed to assess translation quality
- A generalization of BLEU score
- We computed it for n = 1, 2, 3, 4, 5

$$ROUGE_n = \begin{cases} ROUGE_{R,n} = \frac{|N_{sum} \cap N_{doc}|}{|N_{sum}|} \\ ROUGE_{P,n} = \frac{|N_{sum} \cap N_{doc}|}{|N_{doc}|} \\ ROUGE_{F,n} = \frac{2}{ROUGE_{R,n}^{-1} + ROUGE_{P,n}^{-1}} \end{cases}$$

This significantly improved our performance: 0.93312 to 0.94067

Ref: [9, 10]

V- Classification: building the classifier

Comparison between algorithms:

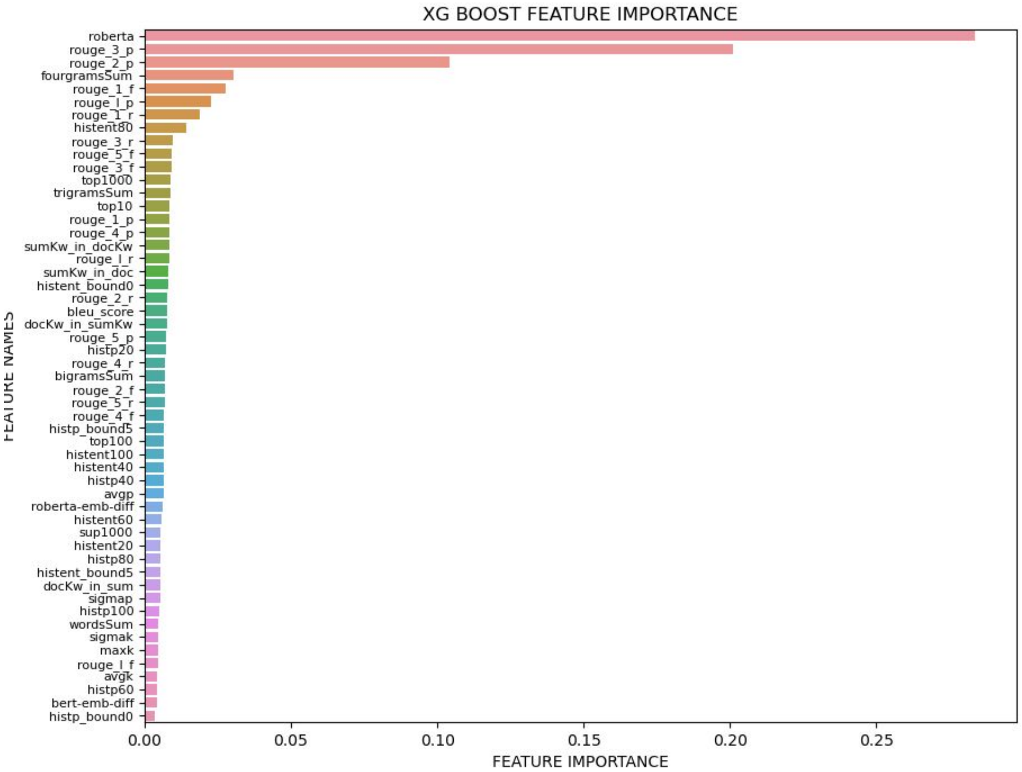
Accuracy on...	Logit model	Random Forest	XG-Boost
Train set	0.951125	1.00000	0.97425
Test set (Kaggle)	0.92500	0.93812	0.93937

XG-Boost hyperparameter tuning:

- *n_estimators,*
- *max_depth,*
- *learning_rate,*
- *colsample_bytree*
- ...

Ref: [11, 12]

V- Classification: performance comparison



	RoBERTa	GLTR	Keywords	N-grams	ROUGE
Score	0.89	0.90687	0.9100	0.93312	0.94067

References

[1] Quoc V. Le, Tomas Mikolov : Distributed Representations of Sentences and Documents, 2014

[2] Devlin, Jacob and Chang, Ming-Wei and Lee, Kenton and Toutanova, Kristina Bert: Pre-training of deep bidirectional transformers for language understanding, 2018

[3] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov: RoBERTa: A Robustly Optimized BERT Pretraining Approach, 2019

[4] Radford, Alec and Wu, Jeffrey and Child, Rewon and Luan, David and Amodei, Dario and Sutskever, Ilya and others Language models are unsupervised multitask learners, OpenAI blog, 2019

[5] Gehrmann, Sebastian and Strobelt, Hendrik and Rush, Alexander M GLTR: Statistical detection and visualization of generated text, 2019

[6] How to Fine-Tune GPT-2 for Text Generation, TowardsDataScience https://towardsdatascience.com/how-to-fine-tune-gpt-2-for-text-generation-ae2ea53bc272.

[7] Maarten Grootendorst, *Keyword Extraction with BERT*, https://towardsdatascience.com/keyword-extraction-with-bert-724efca412ea and it associated github repository https://github.com/MaartenGr/KeyBERT

[8] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. *Bleu: a Method for Automatic Evaluation of Machine Translation*. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

[9] Chin-Yew Lin. 2004. *ROUGE: A Package for Automatic Evaluation of Summaries*. In Text Summarization Branches Out, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

[10] Paul Tardy, ”rouge” git repository https://github.com/pltrdy/rouge

[11] Breiman, Leo Random forests, Machine learning, 2001

[12] Chen, Tianqi and Guestrin, Carlos Xgboost: A scalable tree boosting system, Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016

[13] Liu, Yang and Lapata, Mirella Text summarization with pretrained encoders, 2019