POSOS: Predict the expected answer

Romain Vial

romain.vial@mines-paristech.fr

Collège de France, April 2018

Table of Contents

Introduction

Drug misuse could be responsible for more than 144,000 hospitalizations every year in France!

Introduction

People often ask questions about the drugs they use. How to accurately understand the underlying intent? (contraindication, side effects,...)

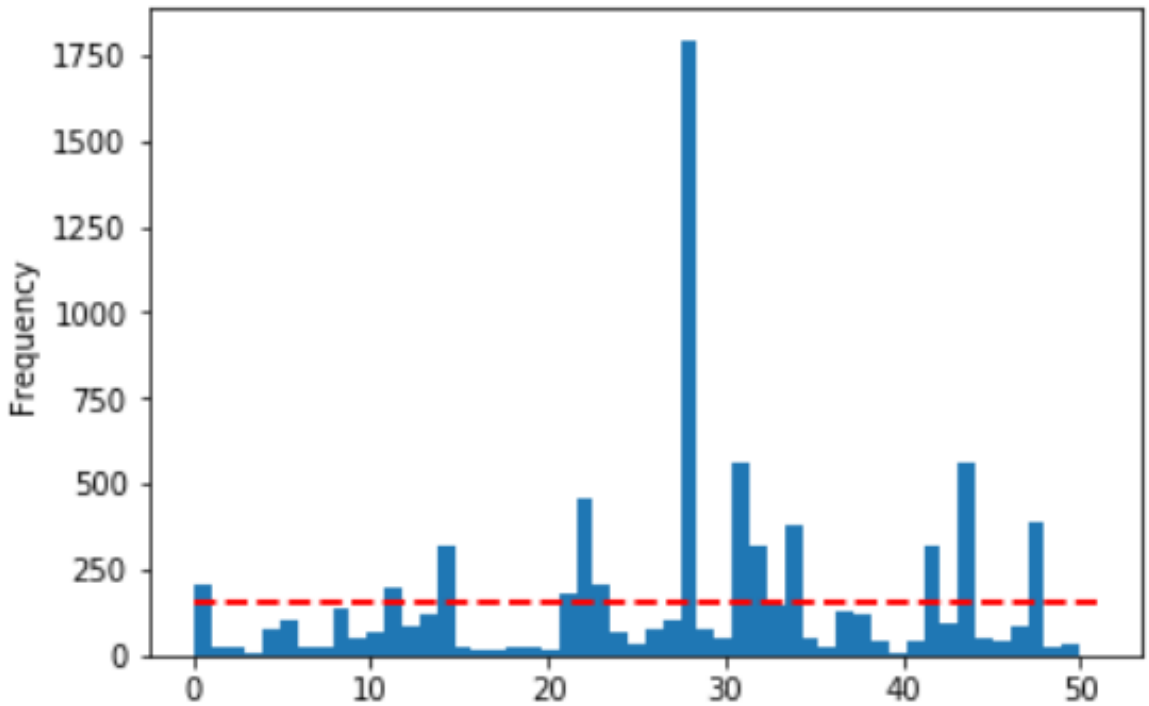
Introduction

Goal: predict the intent associated to a given question among 51 possible categories

Dataset

- 10,063 questions: 8028 for training, 2035 for testing.
- Training set divided in two splits: 80% training, 20% validation

Class Imbalance



Class 28 accounts for 22% while +30 classes account for <1% each

Dealing with Drug Names

"Par quoi remplacer Aerius et Doliprane ?" :

- "Par quoi remplacer Aerius et Doliprane ?" + [2]

Dealing with Drug Names

"Par quoi remplacer Aerius et Doliprane ?" :

- "Par quoi remplacer Aerius et Doliprane ?" + [2]
- "Par quoi remplacer <MED> et <MED> ?"

Dealing with Drug Names

"Par quoi remplacer Aerius et Doliprane ?" :

- "Par quoi remplacer Aerius et Doliprane ?" + [2]
- "Par quoi remplacer <MED> et <MED> ?"
- "Par quoi remplacer <MED0> et <MED1> ?"

Bag of Words

the dog is on the table

TF-IDF

Issue of BoW: some words will be over-represented thus disturb the statistics of rarer but discriminative words

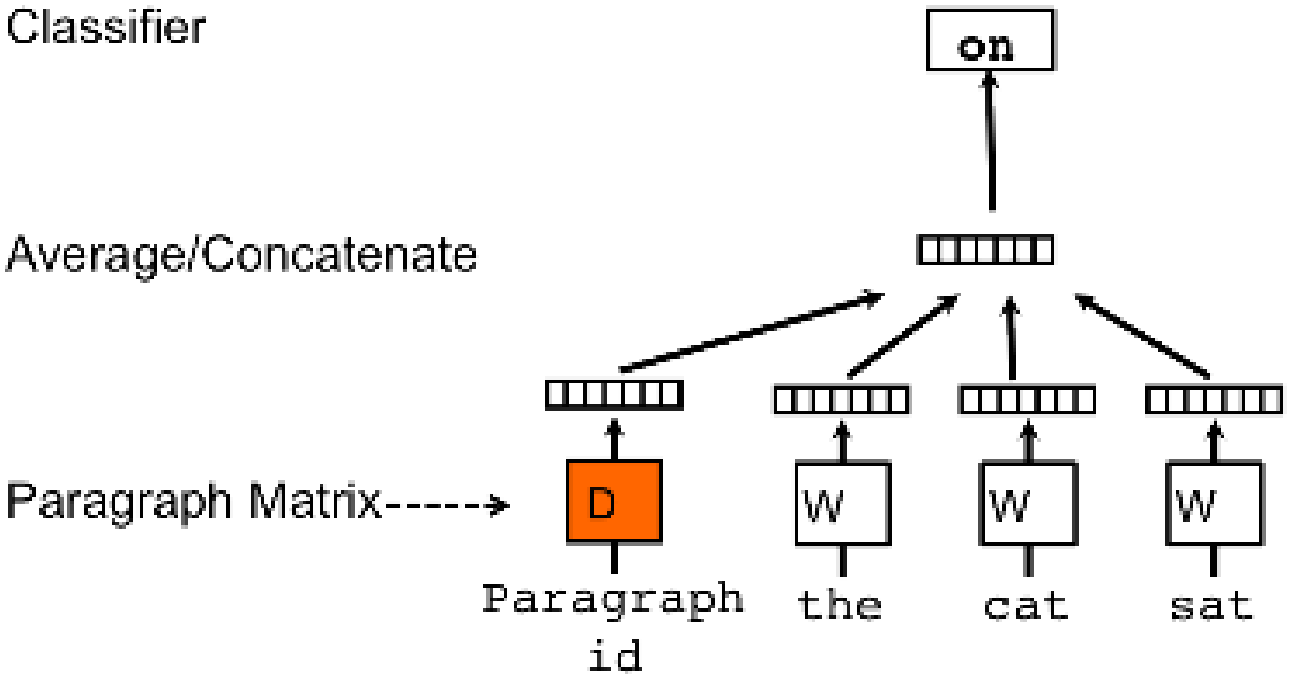
TF-IDF

Issue of BoW: some words will be over-represented thus disturb the statistics of rarer but discriminative words

Reweighing scheme:

$$\begin{aligned} \text{tf-idf}(t, d) &= \text{tf}(t, d) \times \text{idf}(t) \\ &= f_{t,d} \times \log\left(\frac{N}{n_t}\right) \end{aligned}$$

Doc2Vec [Le and Mikolov, 2014]



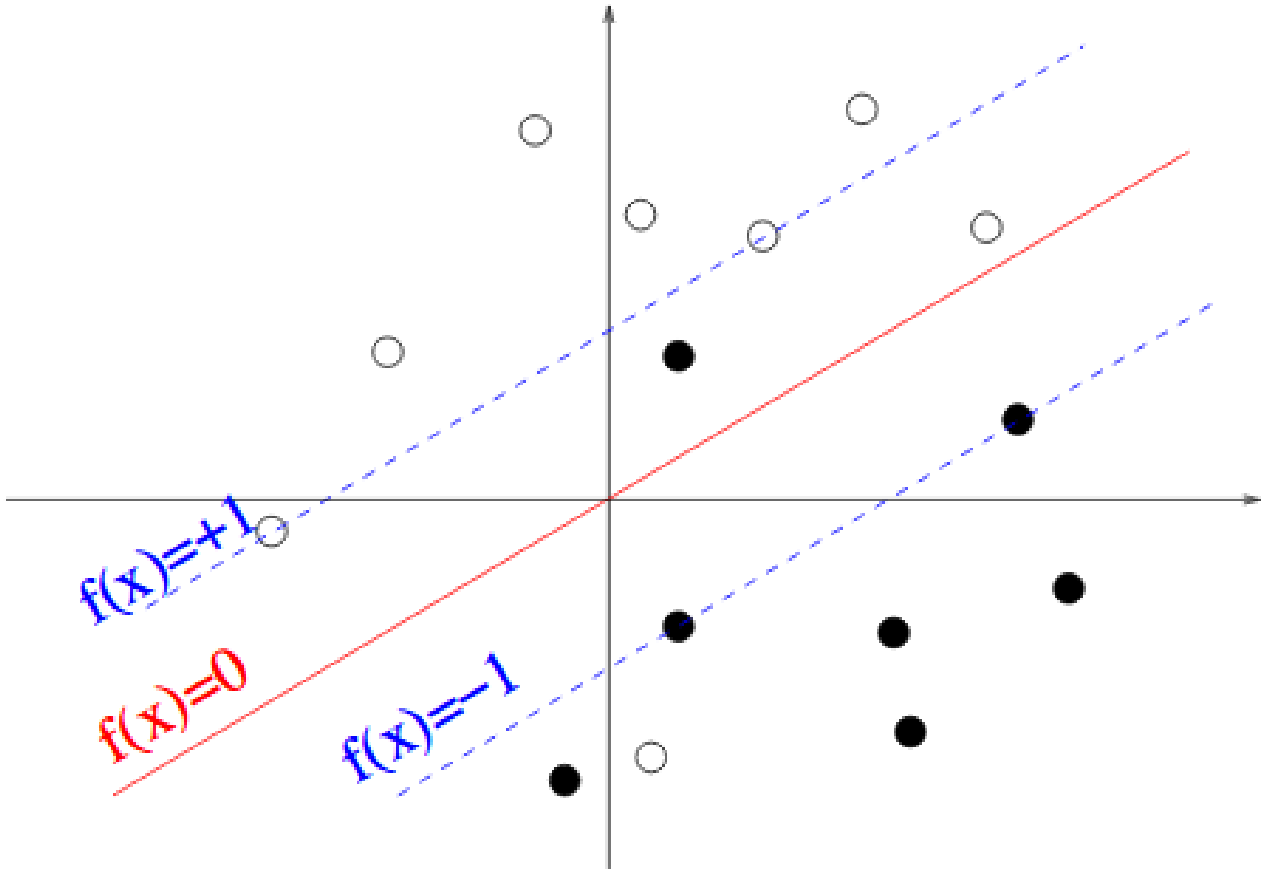
Support Vector Machine [Cortes and Vapnik, 1995]

Idea: (i) map input data $\mathbf{x}$ into a higher dimensional space where the data become separable

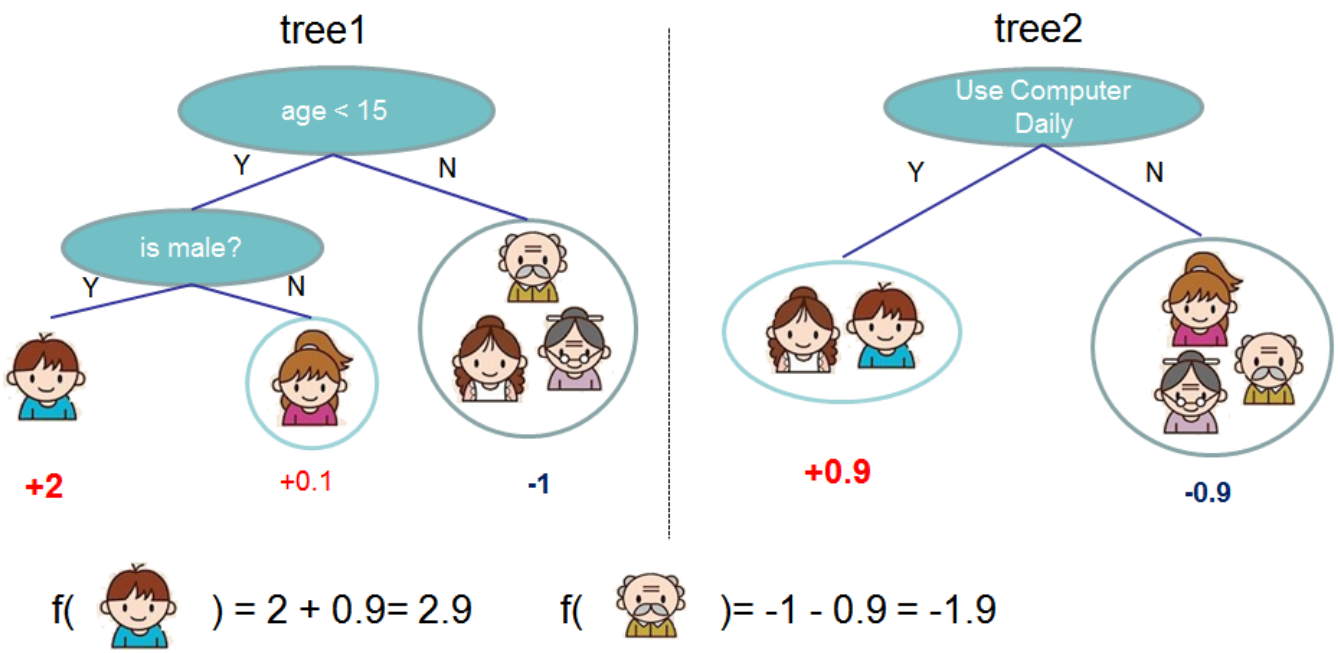
$$\phi : \mathbf{x} \rightarrow \phi(\mathbf{x}) \in \mathcal{H}$$

Support Vector Machine [Cortes and Vapnik, 1995]

(ii) use large margin classifier

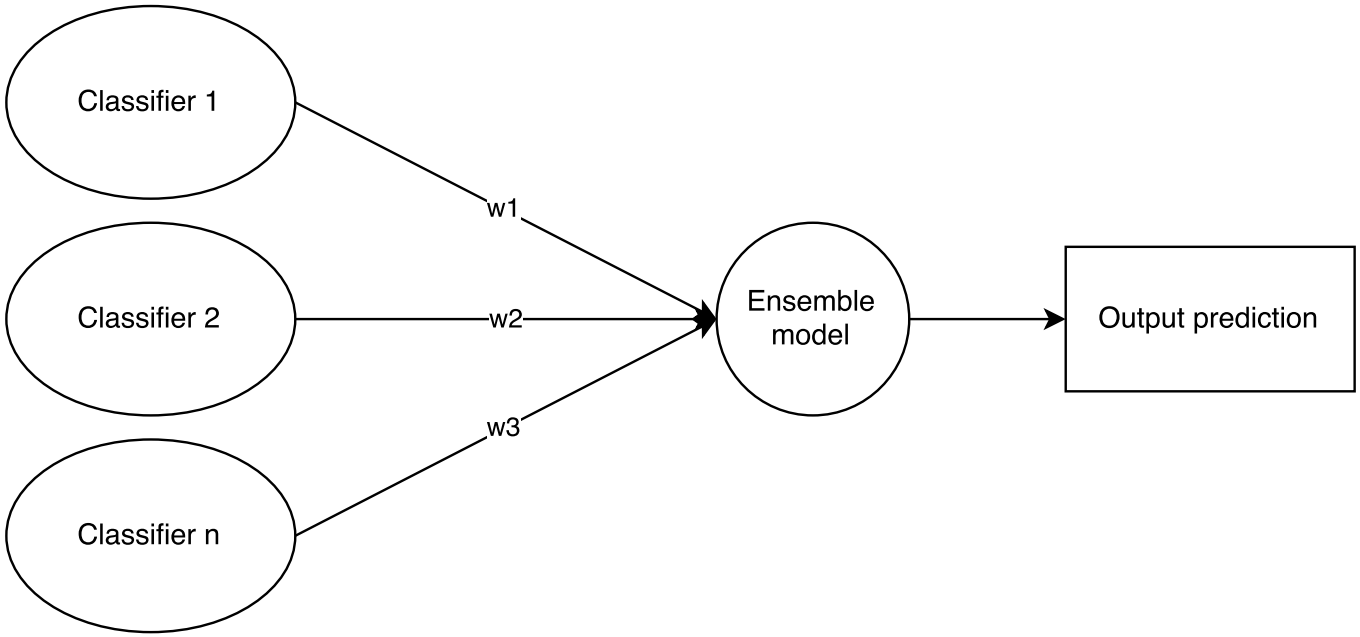


Xgboost [Chen and Guestrin, 2016]



Ensembling

Taking advantage of different models to boost prediction accuracy

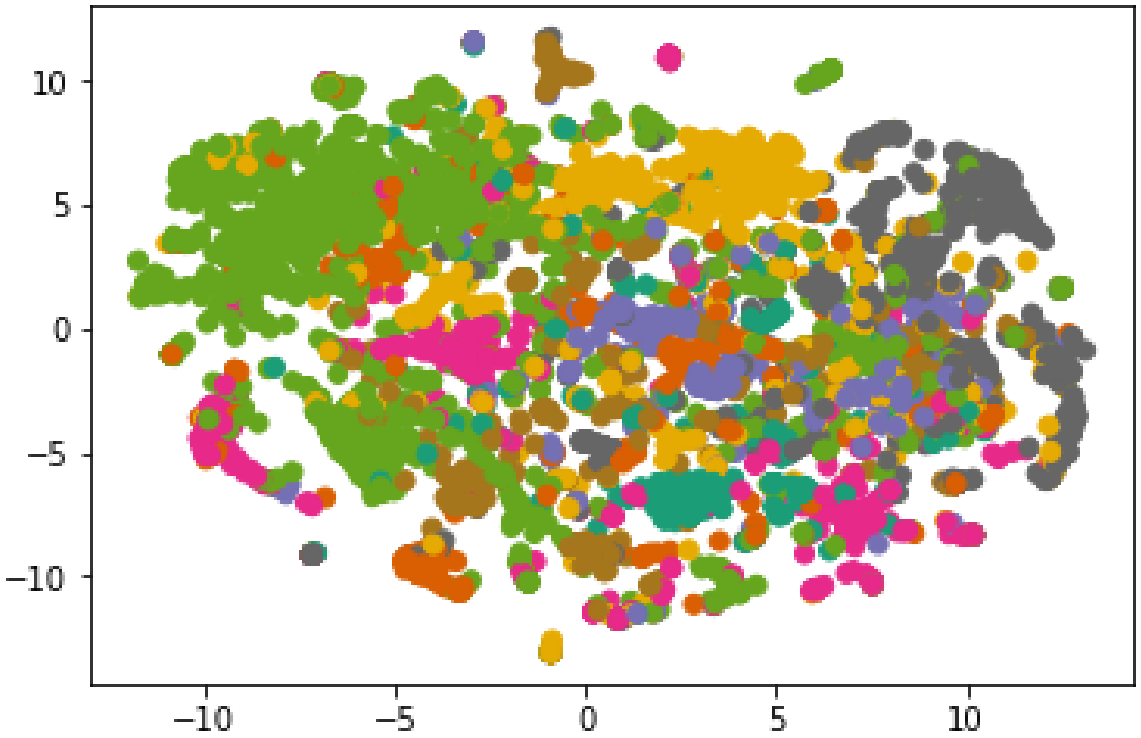


Quantitative Experiments

tf-idf	Doc2Vec	SVM	XGB	train	val
✓		✓		99.41	65.91
✓			✓	94.11	60.50
✓		✓	✓	99.37	**66.71**
	✓	✓		84.28	51.84
	✓		✓	98.65	50.18
	✓	✓	✓	93.98	52.03

Accuracy of 68.60% (8/19) on the test set with tf-idf and SVM/XBF ensemble

Qualitative Experiments



T-SNE representation of the 50-dimensional Doc2Vec embedding of the training split.

Conclusion

- Improve spelling normalization with e.g. noisy channel approaches [Kernighan et al., 1990]

Conclusion

- Improve spelling normalization with e.g. noisy channel approaches [Kernighan et al., 1990]
- Drug embedding by looking at the co-occurrence matrix

Conclusion

- Improve spelling normalization with e.g. noisy channel approaches [Kernighan et al., 1990]
- Drug embedding by looking at the co-occurrence matrix
- Exploring CNN and RNN as a powerful way to learn the features

References

Chen, T. and Guestrin, C. (2016).

Xgboost: A scalable tree boosting system.

In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM.

Cortes, C. and Vapnik, V. (1995).

Support-vector networks.

Machine Learning, 20(3).

Kernighan, M. D., Church, K. W., and Gale, W. A. (1990).

A spelling correction program based on a noisy channel model.

In *Proceedings of the 13th conference on Computational linguistics-Volume 2*, pages 205–210. Association for Computational Linguistics.

Le, Q. and Mikolov, T. (2014).

Distributed representations of sentences and documents.

In *International Conference on Machine Learning*, pages 1188–1196.