White paper

Comparison of Uchidata's algorithm with Facebook's text mining algorithms

By Arnaud de Myttenaere, PhD, Uchidata CEO & Founder.

About the author

Awarded by datascience.net (acquired by EY) as the best french Data Scientist 2014 & 2015, Arnaud de Myttenaere holds a PhD in Mathematics from Paris 1 University and he is a Machine Learning expert (certified by french Ministry of Higher Education and Research). He is the CEO and founder of Uchidata.

Prior to that, Arnaud worked 3 years as a Data Scientist at Viadeo. He graduated in 2013 from ENSAE ParisTech and Ecole Normale Supérieure.

http://uchidata.com/

Last summer Facebook open-sourced FastText, its algorithm to analyze and classify textual data, developed by Facebook's Artificial Intelligence Research (FAIR) lab in Paris.

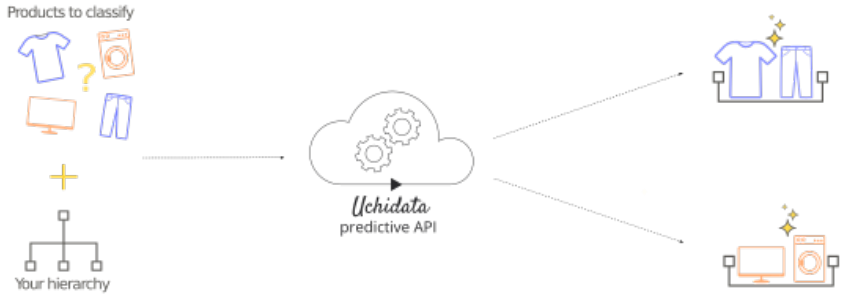
This publication created a buzz in the Machine Learning community, since the algorithm shows great performance on reference classification tasks and is very fast compared to other state-of-the-art algorithms.

> [Facebook] can train FastText on more than one billion words in less than ten minutes [...], and classify half a million sentences among 312K classes in less than a minute.

Joulin *et al.* (2016)

Text analysis is a very large problem and has applications in multiple fields : for example, text comprehension, sentiment analysis, or text tagging. Here at Uchidata, we also focus on machine learning applications for e-commerce. For example, we classify e-commerce products in a fully automated way, using the information of a product provided by the seller (brand, title, description, price, etc).

To do so, we have built a robust methodology in order to set up very accurate classification models, accessible in real time through an easy-to-use API. During the past months, Uchidata's API has analyzed and classified millions of products among thousands of categories and in 5 distinct languages (English, French, German, Italian and Spanish) for several European marketplaces.



UCHIDATA VS FASTTEXT

To illustrate FastText's performance in text classification, Facebook's engineers used several public datasets, and split each set into two parts : one to learn a model, the other to test it. Each dataset and benchmark can be downloaded by cloning and running FastText's code [1].

As described in the original article (Joulin *et al.* (2016)) and in various blogposts (see Kevin J. Ryan (2016) or John Mannes (2016) for example), FastText is highly efficient and

1. https://github.com/facebookresearch/fastText

outperfoms many state-of-the-art algorithms.

To compare Uchidata's algorithm with FastText, we repeated the same methodology and tested our predictions on various test datasets. FastText and Uchidata's accuracies are reported in table 1.

	Datasets							
	Ag	Sogou	dbpedia	Yelp P.	Yelp F.	Yahoo	Amazon. F.	Amazon. P.
Uchidata	92.53%	96.97%	98.84%	96.00%	63.53%	73.37%	60.41%	95.08%
FastText	92.5%	96.8%	98.6%	95.7%	63.9%	72.3%	60.2%	94.6%
VDCNN	91.3%	96.8%	98.7%	95.7%	64.7%	73.4%	63.0%	95.7%

TABLE 1 – Performance comparison with FastText and VDCNN on various public datasets, source : Joulin *et al.* (2016)

Results show that **Uchidata's algorithm outperforms FastText in most cases**. Compared with the performance presented in the original paper, these results also show that Uchidata's algorithm is sometimes more accurate than VDCNN, another algorithm recently developed by Facebook Lab using Very Deep Convolutional Neural Networks (see Conneau *et al.* (2016))

Regarding computational time, Uchidata's algorithm is fast enough to classify products on demand and in real time. **Although Uchidata's algorithm is slightly slower than fastText, it is remarkably faster than VDCNN** (see tables 2 and 3).

	Datasets		
	Yahoo	Amazon F.	Amazon P.
Number of classes	10	5	2
Number of observations	1.400.000	3.000.000	3.600.000
Time (Uchidata)	8m. 18s.	1h. 08m.	31m.
Time (FastText)	27s.	32s.	52s..
Time (VDCNN)	1d. 17h.	5d. 20h.	5d. 20h.

TABLE 2 – Time comparison (train) on various public datasets, source : Joulin *et al.* (2016)

ABOUT UCHIDATA'S ALGORITHM

The initial version of Uchidata's algorithm was awarded by Cdiscount.com in 2015, after a product classification challenge on datascience.net. The challenge was to analyze around 15 million products, and to classify them into more than 5000 categories. Using this algorithm we reached the 3rd place out of a total of 838 challengers.

Today, after 1.5 years of research and developments we proudly present Uchidata 2.0. Our

	Datasets		
	Yahoo	Amazon F.	Amazon P.
Number of classes	10	5	2
Number of observations	60.000	650.000	400.000
Time (Uchidata)	5s.	1m. 28s.	29s.
Time (FastText)	5s.	9s.	10s..
Time (VDCNN)	2h.	7h.	7h.

TABLE 3 – Time comparison (test) on various public datasets, source : John Mannes (2016)

new library is now optimized to learn a model fitting a specific text classification problem, and to compute predictions with high speed and accuracy.

REFERENCES

A. Conneau, H. Schwenk, L. Barrault, and Y. Lecun. Very deep convolutional networks for natural language processing. *arXiv preprint arXiv :1606.01781*, 2016.

John Mannes. Facebook's Artificial Intelligence Research lab releases open source fastText on GitHub, Techcrunh, August 2016.

A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv :1607.01759*, 2016.

Kevin J. Ryan. Facebook's New Open Source Software Can Learn 1 Billion Words in 10 Minutes, Inc, August 2016.