Document Categorization made easy: Start your own machine learning model (without coding)

Stefan Geißler, Kairntech

August 2020

Situation

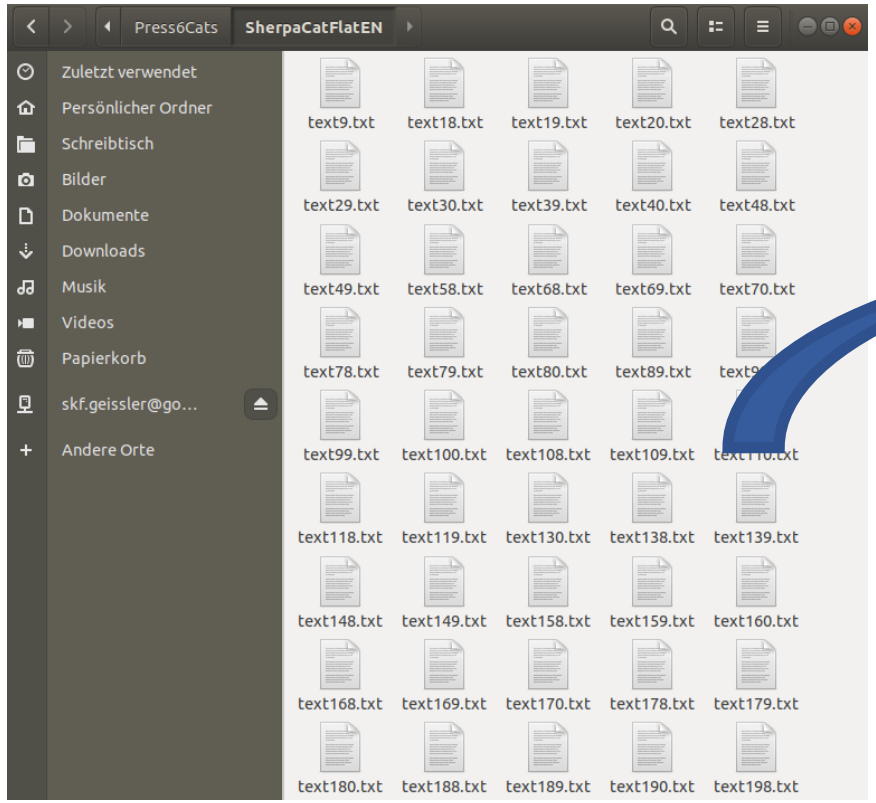
kairntech
AI-powered Text Applications for Enterprise

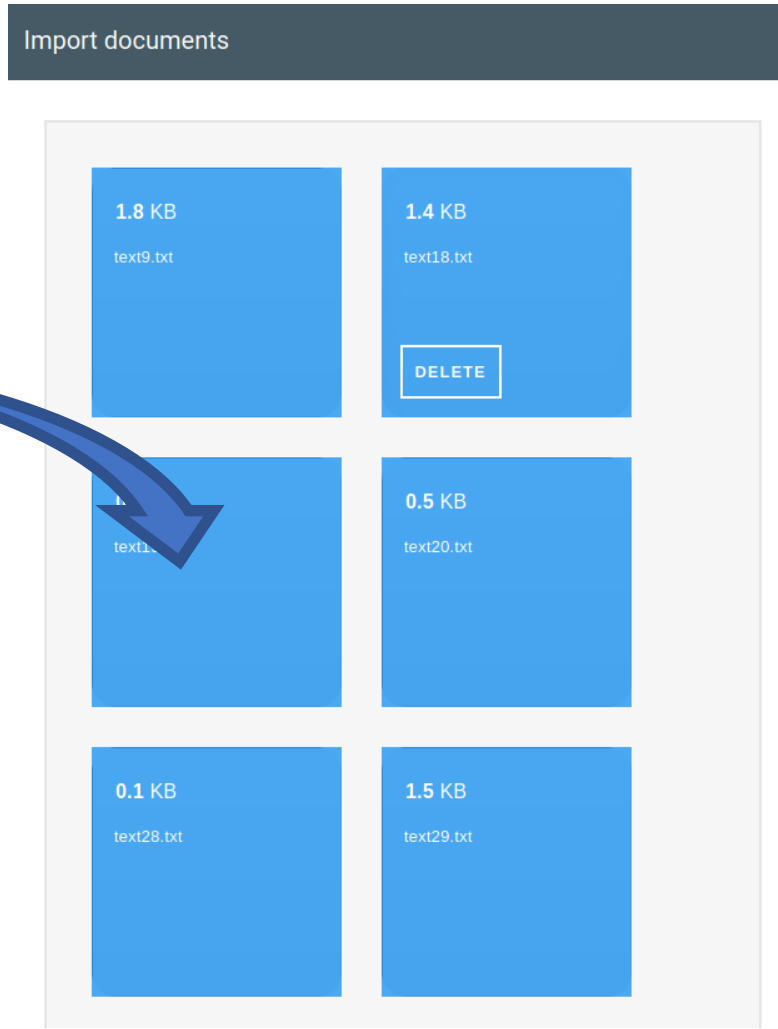
- I need to sort a large number of documents into appropriate folders or categories
- I have no training corpus
- I don’t want to (aehm: I don’t know how to) program
- Still, I want to benefit from these new clever machine learning / AI / Deep Learning technologies that everyone is so excited about

Solution: I do this in the Kairntech Sherpa!

Step 1: Import the documents

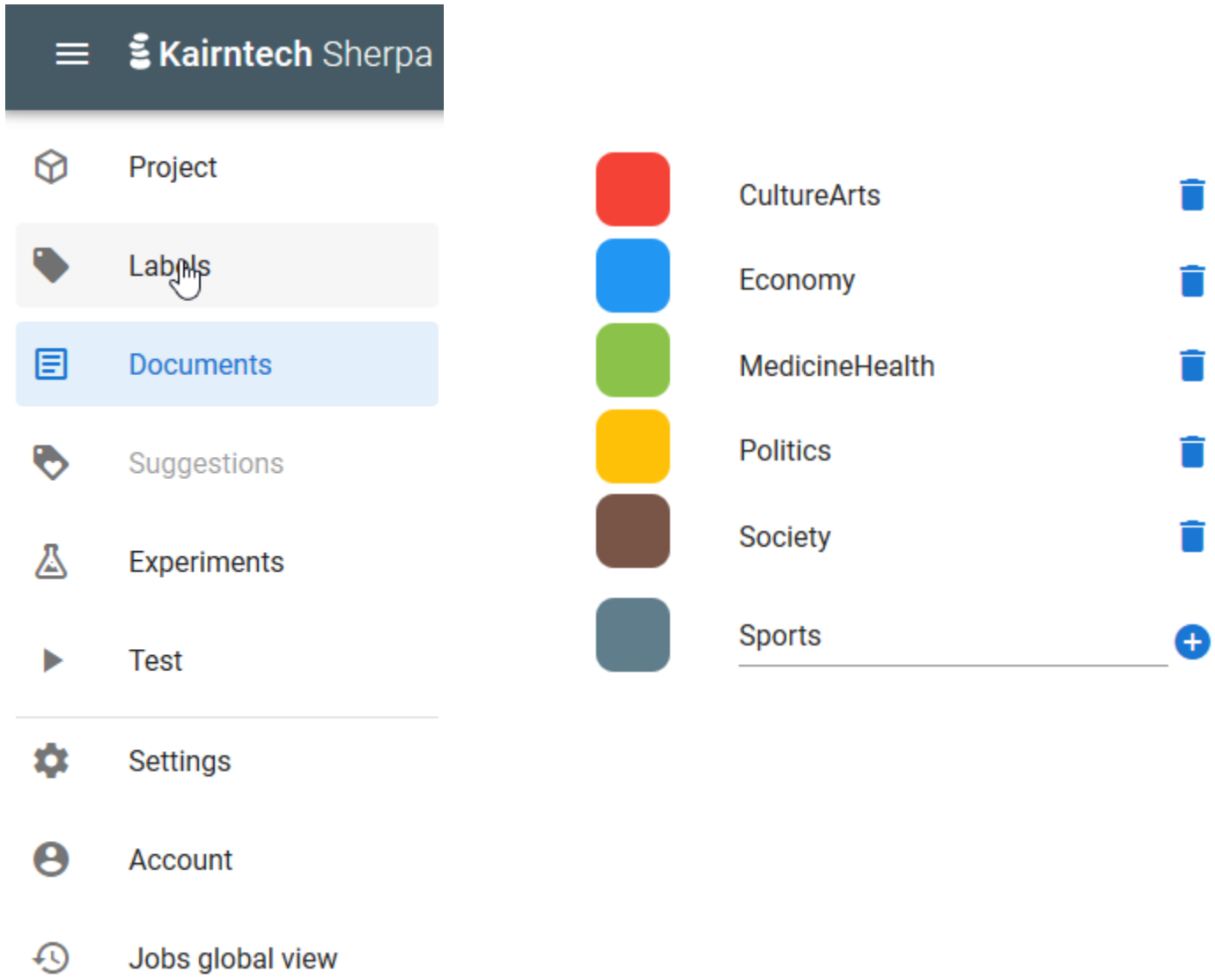
kairntech
AI-powered Text Applications for Enterprise





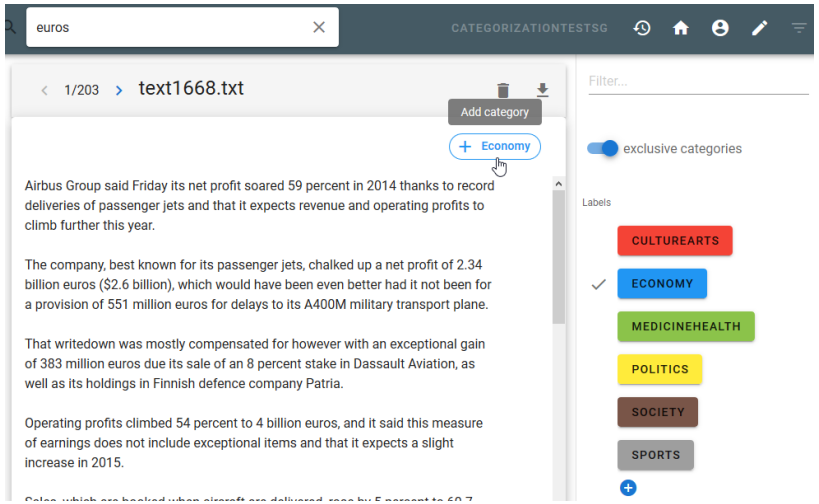
Step 2: Define your categories

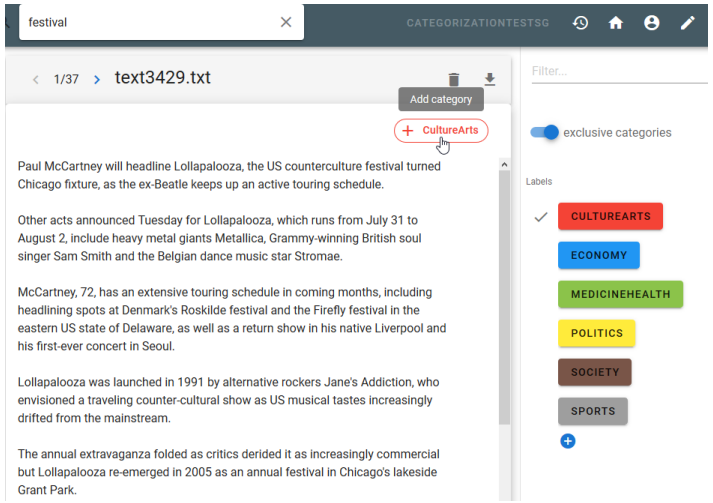
kairntech
AI-powered Text Applications for Enterprise

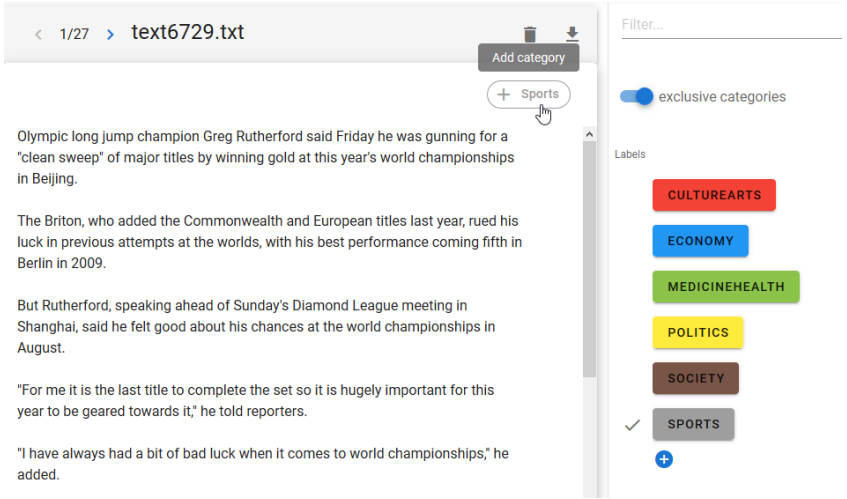


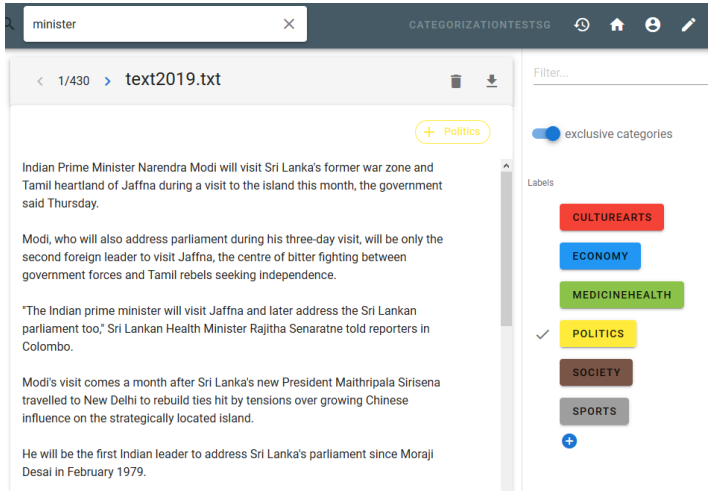
Step 3: Start putting documents into categories

kairntech
AI-powered Text Applications for Enterprise









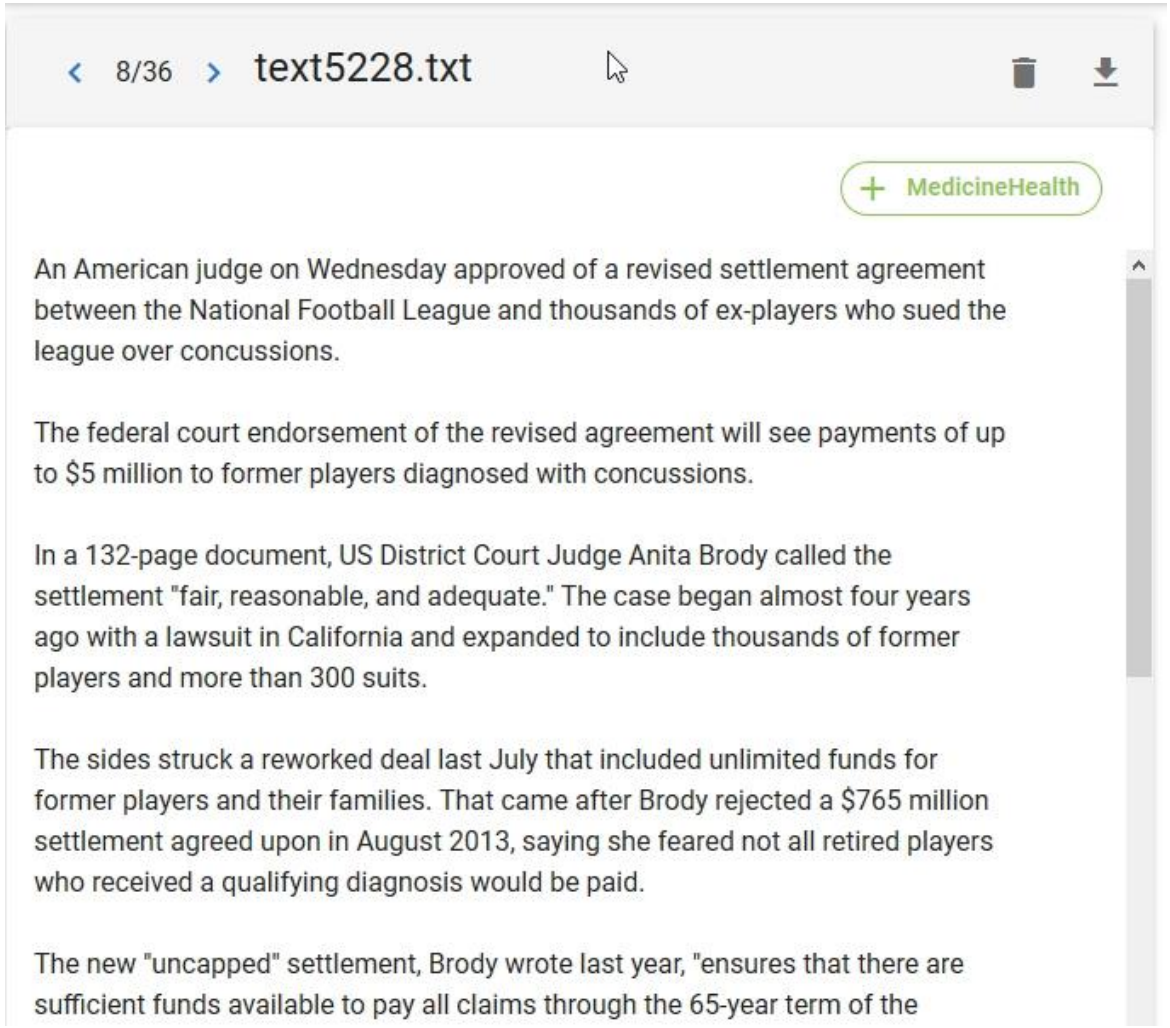
Here, my domain expertise is required. Assumption

- Many (not all) documents that speak about “championships” are about Sports
- “Minister” will indicate Politics.
- “Festival” CultureArts
- “Euros” may suggest Economy

I step through the search results for each of these and quickly collect 2-3 dozen sample documents for each.

Step 3: Unclear cases? Exclude!

kairntech
AI-powered Text Applications for Enterprise

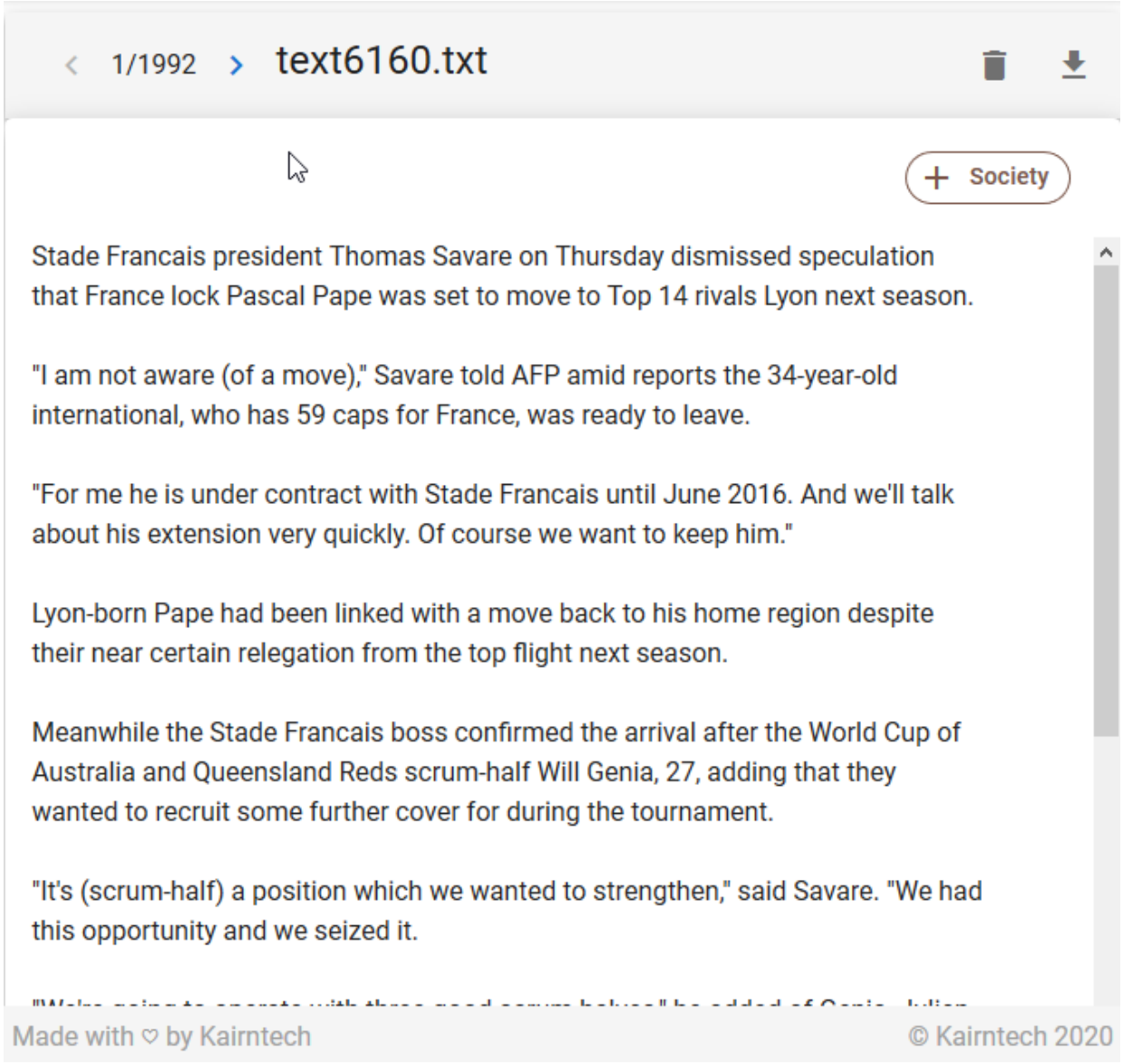


Text mentions the topics Sports and MedicineHealth issues and legal/economic topics …

When unsure, just exclude the document for a start or assign it to more than one category.

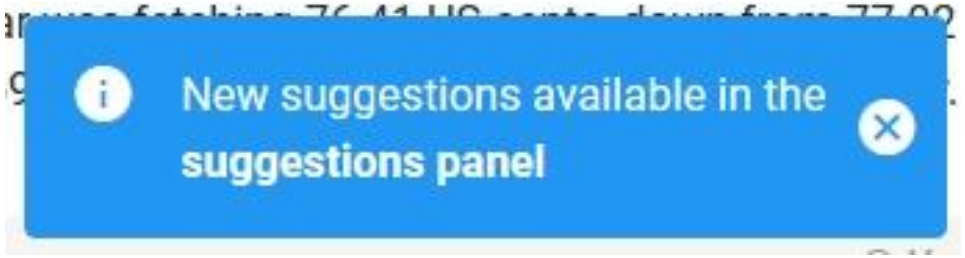
Step 4: System starts to learn ...

kairntech
AI-powered Text Applications for Enterprise



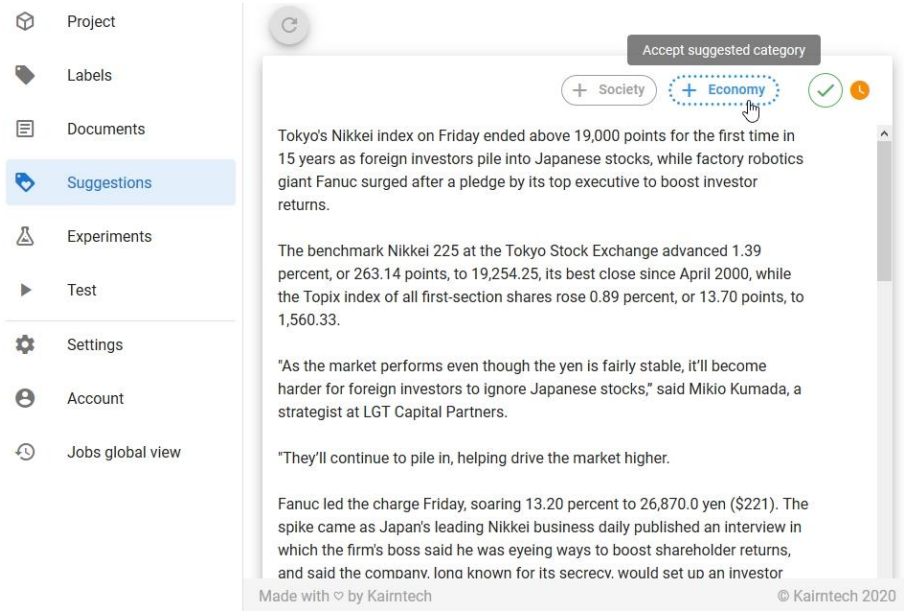
After a few minutes of assigning documents to categories, the system starts to learn.

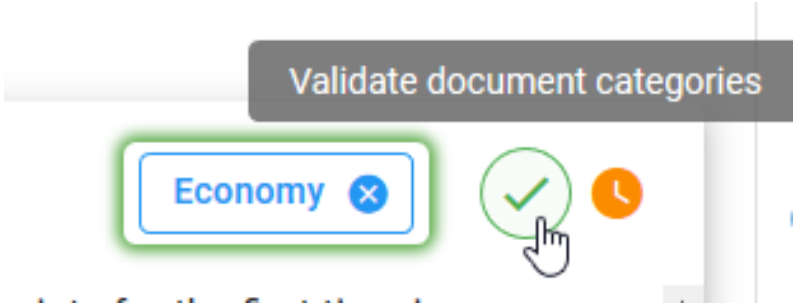
… and generates automatically new suggestions.



Step 4: accepting suggestions (or not) ...

kairntech
AI-powered Text Applications for Enterprise





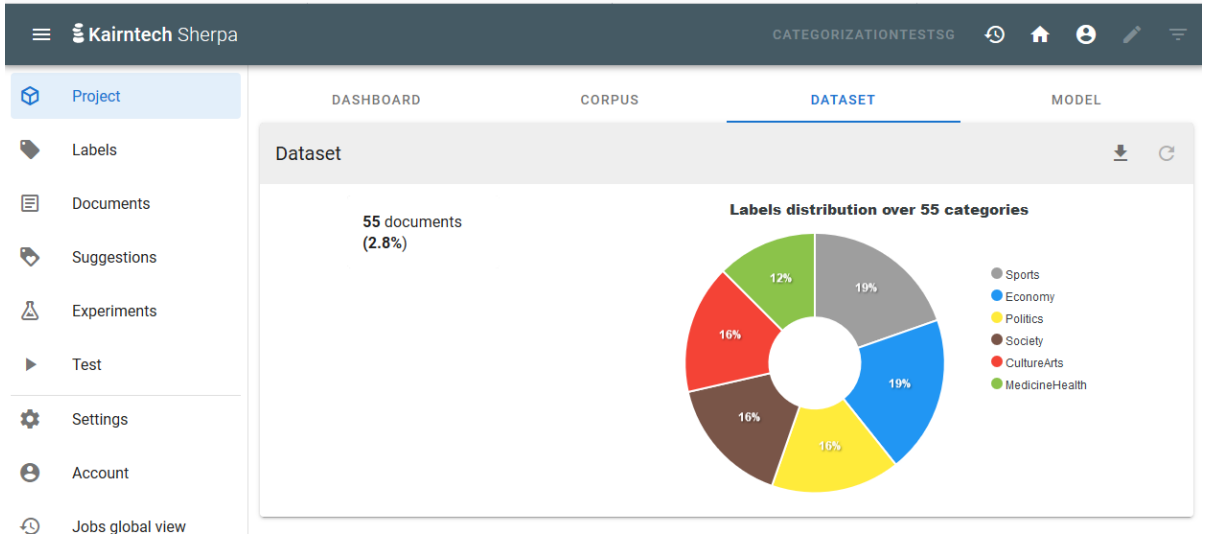
System presents documents it "thinks" are about a given category.

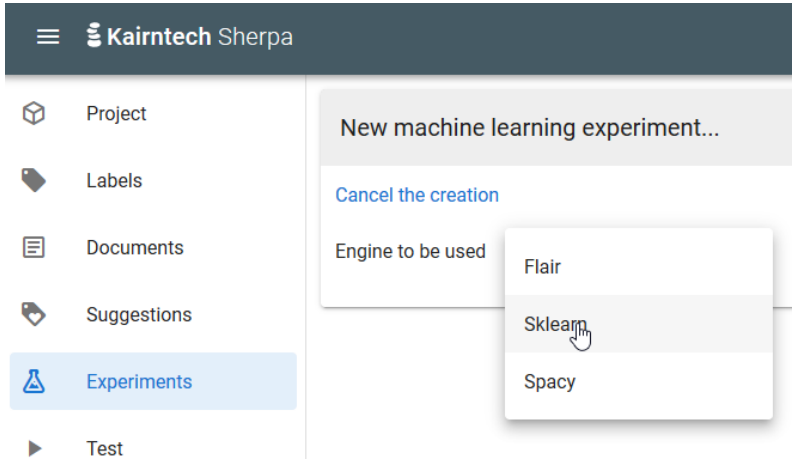
I accept and immediately the next suggestion pops up.

This is faster than having to select documents and their training categories by myself.

Step 5: where am I?

kairntech
AI-powered Text Applications for Enterprise



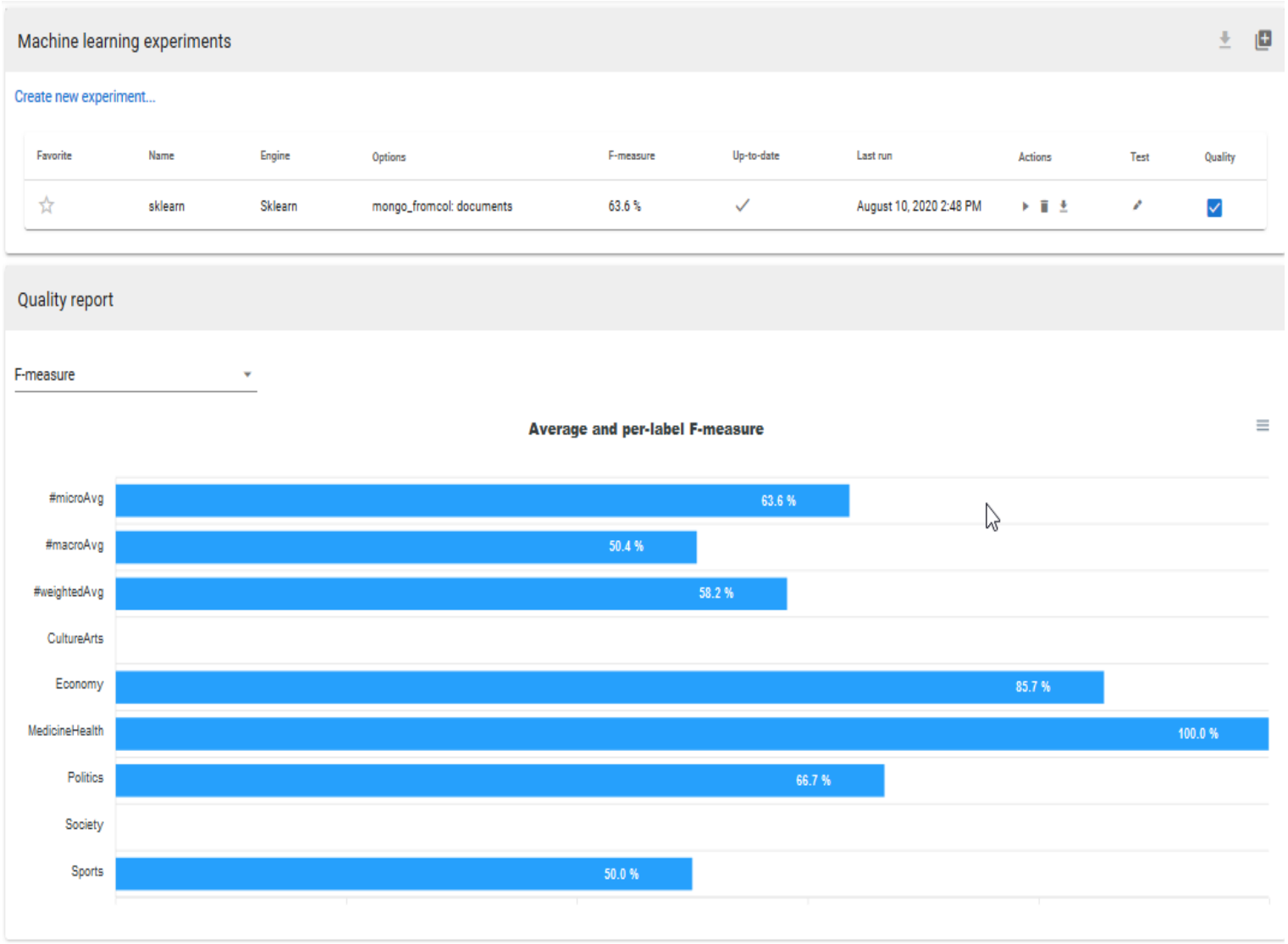


Check how many examples I have created?

Ok, time for a first evaluation!

Step 6: some orientation

kairntech
AI-powered Text Applications for Enterprise



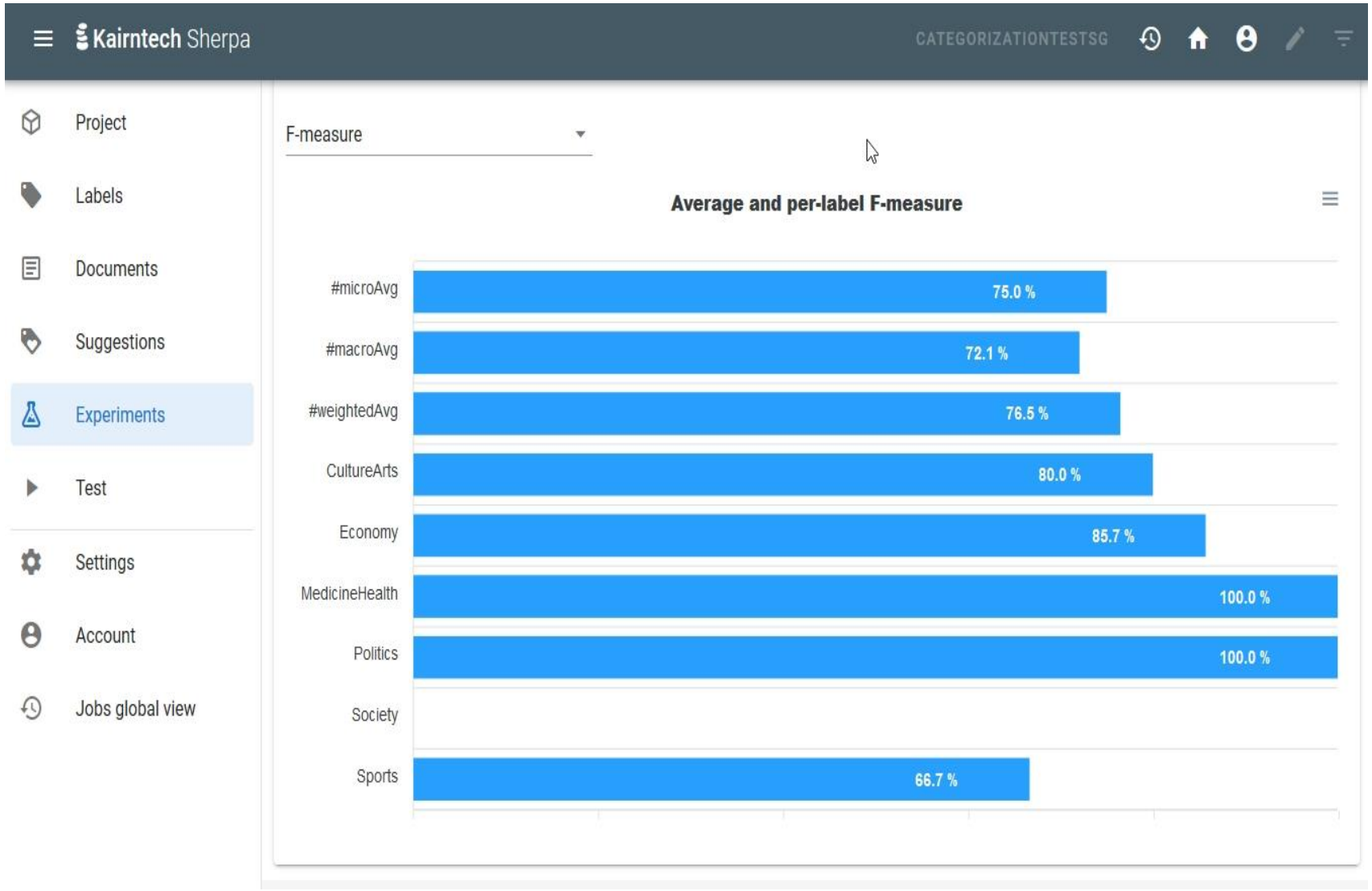
Ok, some categories are already quite ok.

“CultureArts” and “Society” seem to still require some work.

Let’s go back to assigning some more in these categories …

Step 7: … after adding some CultureArt texts

kairntech
AI-powered Text Applications for Enterprise

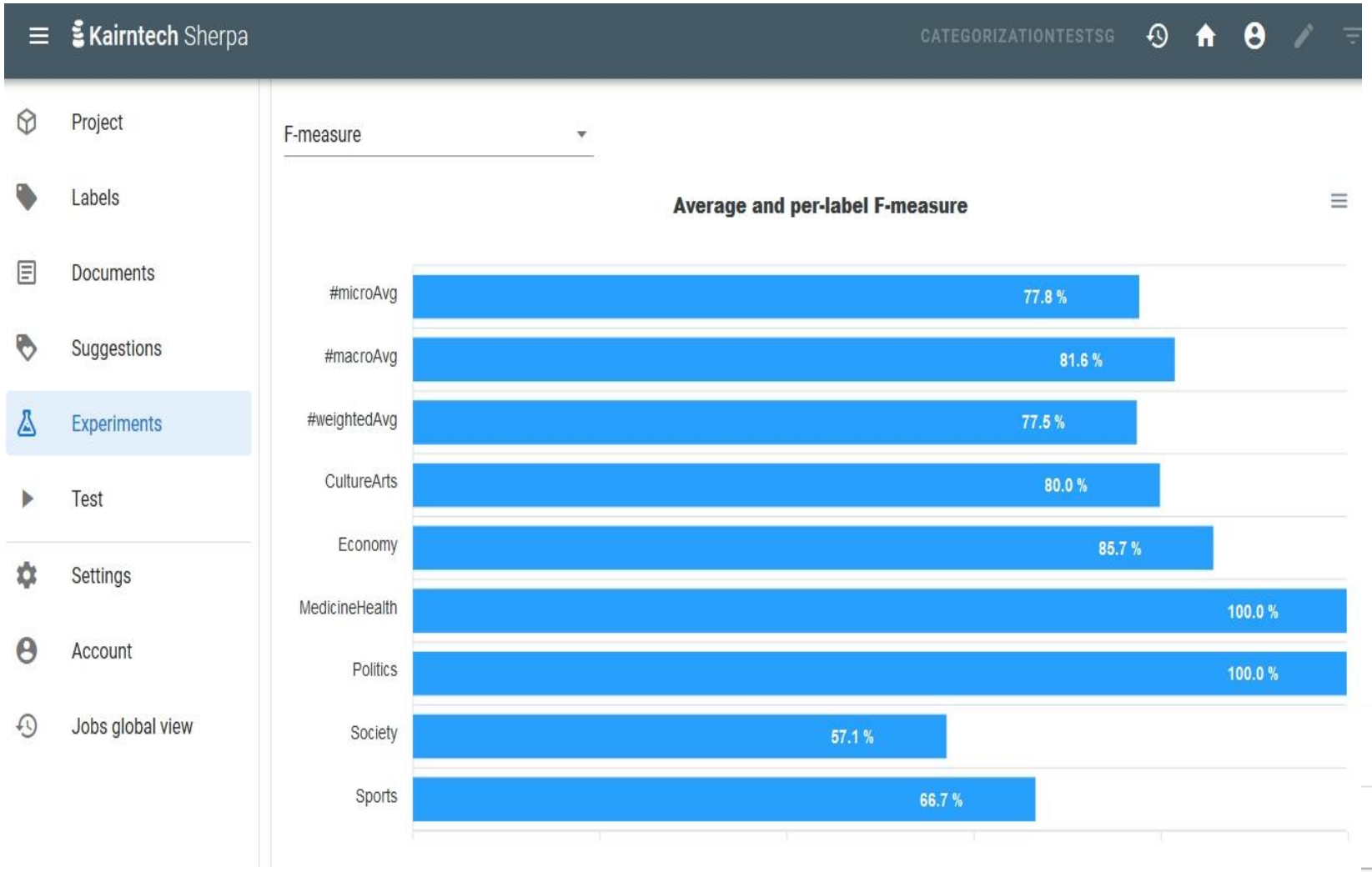


Ah we are getting there!

But “Society” still seems to need some attention …

Step 7: … after adding some CultureArt texts

kairntech
AI-powered Text Applications for Enterprise



See!

Until here we have spent roughly 30-45 min.

We have started from scratch with uncategorized documents using only our domain knowledge

Arriving at ~78% quality.

F-measure	Up-to-date	Last run
77.8 %	✓	August 10, 2020 4:00 PM

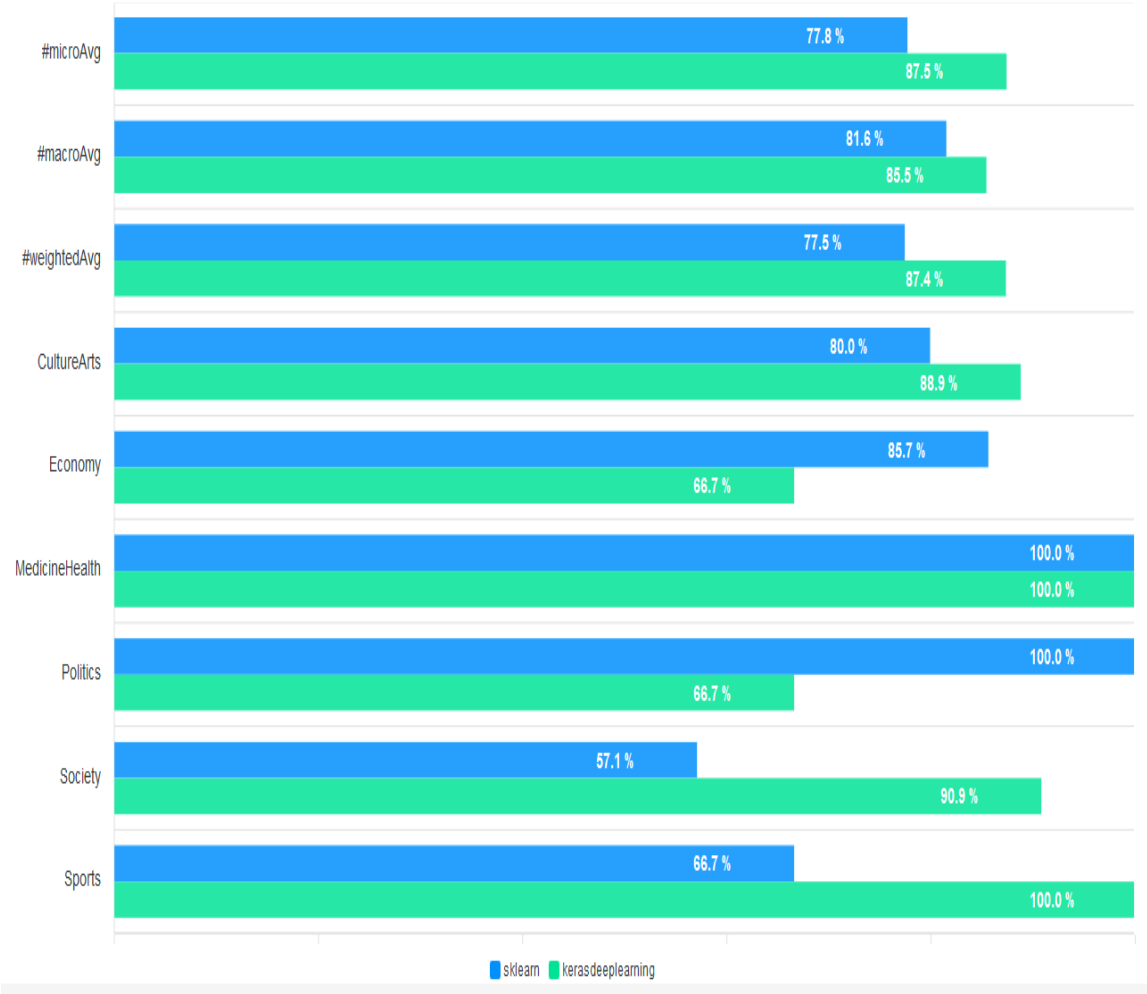
Step 8: Tuning

kairntech
AI-powered Text Applications for Enterprise

- So far we have « only » used the default categorization (quick, applicable to most scenarios, but not the top approach that Sherpa has to offer)
- Let's now apply a Deep Learning based approach – all available from the Web GUI

kerasdeeplearning	Sklearn	train_on: train, test_on: test, mongo_fromcol: documents, classifier: KerasMLPClassifier, featurizer: tfidf	87.5 %

- The Deep Learing based approach boosts the quality from the default 78% up to 87% on out little manually created training corpus.
- Experimentation/training time sofar: Still less than 1 hour.



Conclusion

kairntech
AI-powered Text Applications for Enterprise

- Easy intuitive handling of the application
- Instant, online learning, immediate feedback
- Useful results after very short time
- Need better results?
  - Invest some more time, adding manually labelled data
  - Experiment with other learning methods
- I now have a categorizer and I didn’t have to bother anyone to implement it for me.

The Kairntech Sherpa:
Powerful Machine Learning for non-programmers.

www.kairntech.com