identrics

FILTERING MEDIA CONTENT AND ALERTING ABOUT THE APPEARANCE OF SPECIAL INTERESTED ENTITIES

SUCCESS STORY

Case study of technology implemented in a risk and compliance operating company

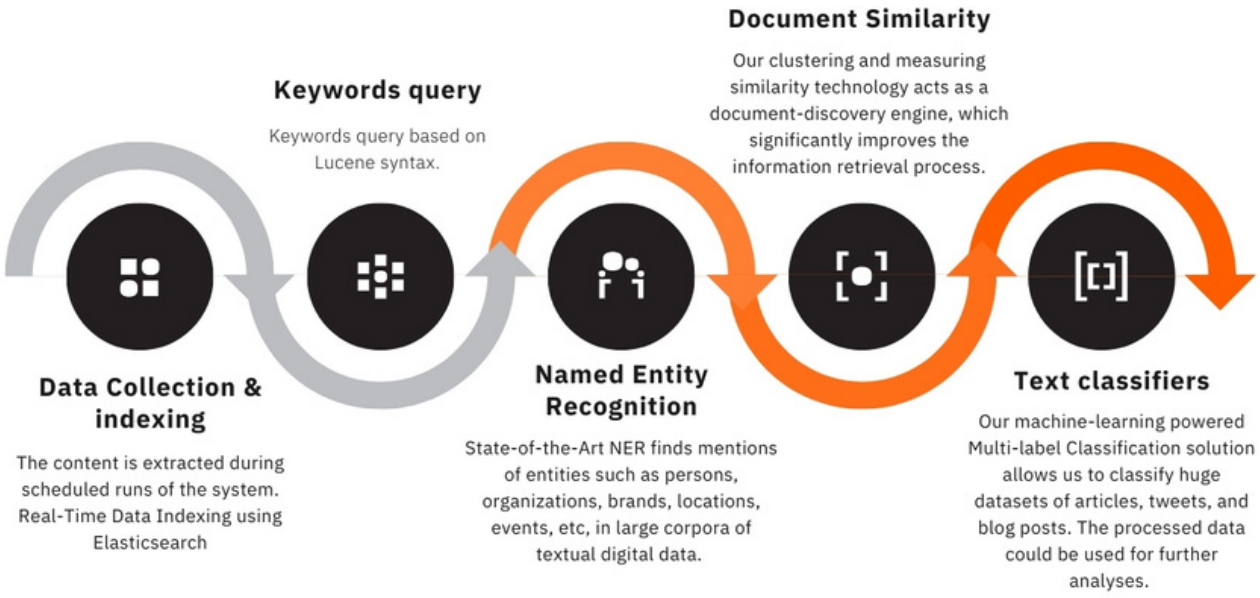
PROJECT OVERVIEW

This is one of our production solutions where our client examines a large array of media publications intending to identify entities (people/organizations) that are mentioned in connection with any events of their interest.

Before the implementation of semantic technologies, the process was divided into two parts, including filtering information through search queries and analyzing it with the help of various experts. One of the main problems the client was facing was the presence of numerous false-positive articles after the search queries.

This issue meant that analysts had to dedicate a significant portion of their time to reading irrelevant content and manually filtering all of the information.

After adopting machine learning models for information management, manual work has been dramatically reduced. As reported by the client, the process of building a database of such entities has been optimized by about 70%.



PIPELINE

Let us outline the services we added to their initial pipeline to gain this improvement:

Named Entity Recognition.
With this automation, we managed to eliminate all articles which do not contain any entities. Regardless of whether a keyword is present in the publication, the model will skip it altogether if no entity or entities are mentioned. This way, we filter around 2% of the data set. Articles mentioning the same entities are then grouped at the output to be more easily analyzed by experts.

Document Similarity.
This technology significantly improves the information retrieval process, as it allows us to create clusters of articles and then leave people to work directly with them; it also serves to eliminate duplicate publications with similarity above a certain threshold, filtering as much as 42% to 73% of it in different languages.

Text Classifiers.
We have developed custom text classifiers for each different language required from our client. This technology allowed us to classify and filter out as much as 40% of irrelevant events.

The process shortly described above is implemented in a pipeline containing custom logic to group and organize the output.

We currently cover the following languages: Spanish; Chinese; English; French; Russian; Dutch; Bahasa; Turkish; Italian; Japanese; German; Danish; Greek; Bulgarian; and Swedish.

CHALLENGE & SOLUTION

The biggest challenge we faced when solving the problem was setting up a clear Human-in-the-Loop (HITL) process and gathering all the example data needed to create classifiers for every language. What usually happens with client data is that it is not directly usable to solve the task. There is almost always something missing and the data does not seemingly want to “tell its story” :) In this particular case, our client had tons of relevant data and not a single example of what irrelevant content was or could be.

Our first mission was gathering irrelevant example data for the classifiers.

We used the definition of what exactly an irrelevant article is and developed some rules to filter such articles out. Then, we introduced Active Learning. Active Learning is one of the cornerstones of Human-in-the-Loop Machine Learning.

A good Active Learning strategy focuses on how you get training data from people, and what is the right data to put in front of people when you do not have the budget or time for human feedback on all of it. We started highlighting the irrelevant articles as defined by the rules and collecting feedback from analysts, asking them to accept or reject the highlighted content. This was the shortest way to start collecting negative data for the classifiers, in order to begin the training of initial models and improve their performance in the same loop.

Our next mission was to get above 0.90 for both precision and recall of the classifiers so we could deploy the models in production.

We approached all of the classification algorithms and neural networks that we usually use to classify text. The results were always close to, but not quite, the desired scores. More data was not the solution because we had already collected a lot – both positive and negative examples.

CHALLENGE & SOLUTION

In this particular case, we got the best results by assembling several lightweight but powerful classifiers and using their combined predictions for the final decision. In classification, this solution is called a hard voting ensemble. It involves summing and weighing the votes for crisp class labels and predicting the class with the most confidence mass. And it does the job!

Given a large enough number of diverse well-tuned models (~10) of accuracy above a certain threshold (~70%), they successfully calibrated each other and both precision and recall were boosted to our desired levels.

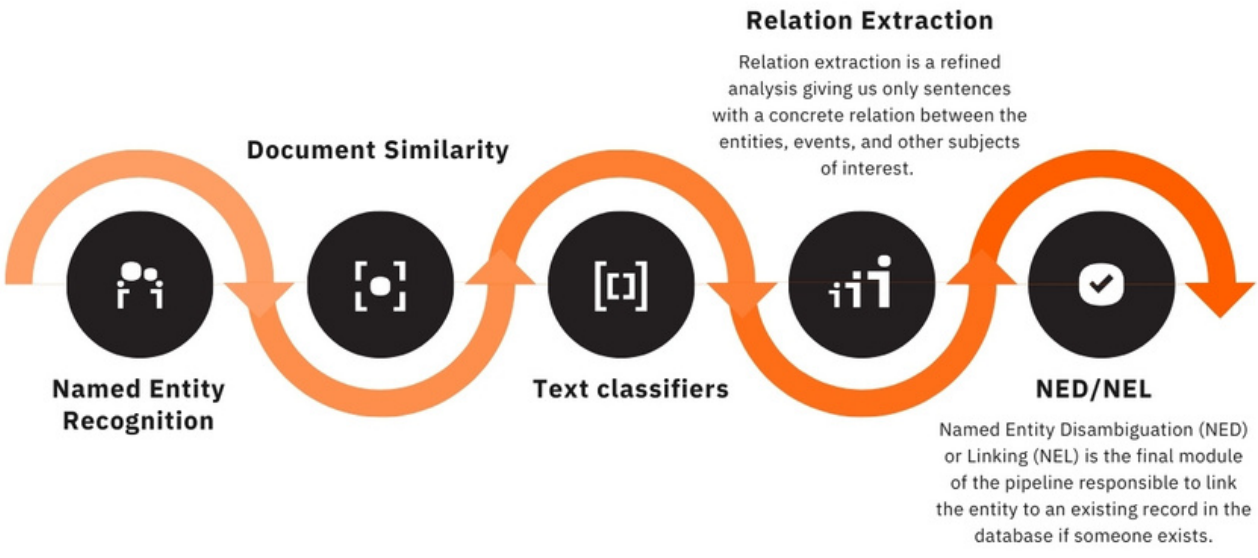
The ensembles are then deployed and monitored in real-time to ensure that they behave and generalize as expected.

Now our client expands this optimization very shortly on new languages (in less than a month per language). We plan to improve it with additional models.

THE IMPROVEMENT

We introduced Relation Extraction and Named Entity Disambiguation in the pipelines to become much more precise. The improved solution refines analysis and alerts only about articles mentioning an object in connection to events that interest the client. In addition, all data points that form a real record in the database can be extracted too. The named entity disambiguation module additionally displays if the entity is already in the database or is a completely new record to be processed.

The improved pipeline is in its final development phase and we cannot wait to see how it will improve our client's performance, even more, when deployed to production.



AUTORS

Here are the people who worked on this paper:

Iva Marinova, Data Scientist

Iva's experience covers deep learning and artificial neural networks, statistical data analysis, and NER. She speaks four languages and holds the unlikely combination of a Master's degree in Computer and Information Systems Security and a Bachelor's in Theatre.

Kristian Krastev, Machine Learning Engineer

Kristian is an R&D engineer and data scientist, working on Machine/Deep Learning applications for Natural Language Processing in the data intelligence industry. He is skilled in Python, Java, and C/C++. A cinema buff, Kristian volunteers at a local film production company and speaks Russian and French.

Meet the rest of the team on our website: identrics.net/team

About us

Identrics intertwines human intelligence with the latest AI & automation technologies to deliver tailor-made solutions that optimize your business processes, boost cost-efficiency and create added value.

What sets us apart is the level of tailoring, consulting and support we provide. Our domain experts train the technology with your own dataset, which allows us to reach precision, achievable usually through manual work.

We currently offer ready-made entity extraction and sentiment analysis models in English, German, French, Spanish, Bulgarian, Italian, traditional Chinese, Portuguese, Russian, Swedish, and Dutch.

Our solutions are immediately deployable via API, easily customizable, and can be augmented via our Data science as a service (DSaaS) capabilities.

Need additional info about our ready-made models or consultation about what our team can do for your business?

Just contact us!