Automatic Entity Matching

Corporate Research Project - Société Générale

Joshua Fan – Sanjana Gupta – Aline Helburg – Anmol Katiyar – Emma Riguidel

Context

In 2018, Société Générale was fined by the U.S. Department of Justice $1.3 billion due to "illegal and non-transparent transactions" to countries currently under sanctions or embargoed by the United States, including Iran, Sudan, Cuba and Libya between 2003 and 2013. The bank processed more than 2,500 transactions valued at $13 billion. These transactions should have triggered alerts that subjected them to further review. Following this regulatory fine, Société Générale has taken actions such as: "recruiting additional compliance officers", "reinforcing (...) [their] alert management teams" and "emphasizing to all employees the importance of compliance". In 2015, BNP Paribas was fined $8.9 billion for the same reasons.

Such enormous fines have led Société Générale to enhance compliance and the bank has chosen a conservative approach: all suspicious transactions must be checked. To do so, the names, country of origin, country of destination, date of birth of sender and receiver have to be checked. However, Société Générale handles huge amounts of data as well as an important diversity in names (language differences, acronyms and synonyms). The current pipeline does not match Société Générale's conservative position and the amount of false alerts are flooding the checking system. The need for an automatic and accurate entity matching system is crucial to reduce operational costs (by reducing the manual reviews), avoid the regulatory fines and preserve the bank's reputation.

Here, we focus on only name identification and linkage, not on countries, date of birth, etc.

1 Objectives

1.1. Dealing with data diversity

Société Générale is an international bank that is present in 66 countries and has 25 million customers. Such amount of data across different languages and customs raises issues.

Some names have an equivalent in other languages. Indeed, a person named William in the U.S.A. can also have Will or Bill as surname, is named Guillaume in France and Guglielmo in Italy. The centralized Société Générale's checking system should be able to identify those names as similar. A transaction made in France from Guillaume Gates to an Italian bank account belonging to Guglielmo Gates should be identified as a transaction between the same person.

More than the differences in languages, there are different customs depending on the countries. For example, typical Russian last names end in "ov" for men and "ova" for women. It means that the checking system should be able to draw a distinction between Ivanov (male) and Ivanova (female) as it is not the same person.

The system should also be able to handle misspellings and should tolerate some alterations in word order, for example 'Bill Gates' should be considered the same as 'Gates Bill'.

Finally, the middle names or prefixes should also be considered. William Harry Gates is not the same person as William John Gates.

This multiple, dispersed and heterogeneous information leads us to pose the question: given the names of two entities how can we confirm that they correspond to the same entity? We seeked to test different approaches to develop a model that will be able to respond to this problem.

1.2. Dealing with Société Générale's identification pipeline

A list of sanctioned entities and potential risky clients are given to the financial crime team of Société Générale. There are multiple client lists: generated by the U.S.A, France, European Union, the United Nations, etc. In order to make sure that all given entities do not circumvent the screening process, a few measures were in place to ensure compliance with the regulations. Société Générale has several processes in the alerting pipeline to make sure that they will not miss any risky client to align with

their conservative position. The problem with their pipeline is that the amount of false alerts is too important. They are flooding the system. The workforce needed is extremely important. Société Générale estimates the true negative to be 0.01% of the total of alerts generated.

Therefore, our project aims to propose a step that would disqualify false positives while not missing any true negatives. It consists of implementing a natural language processing model that checks the names of the sender and receiver to make sure they are not part of any sanctioned entities.

The overall objective is the following: our model should be able to handle data diversity in order to improve Société Générale's pipeline.

2 Data Description

Due to Société Générale's activities, we did not have access to their data. They are highly confidential. Therefore, we built our own dataset using a publicly available dataset. The original dataset consisted of a list of names (first name + last names).
We built our dataset using synonyms of the names contained in the original dataset. We used Wikidata to search for alternative spellings of names. Wikidata, hosted by the Wikimedia Foundation, is a multilingual website. It serves as a central repository of open data accessible to Wikimedia projects like Wikipedia, as well as other individuals or entities. The data provided by Wikidata is available for use under the Creative Commons public domain license.

For example, the synonyms of the name Bill are:
Bill: Guillaume, William, Willem, Wilhelm, Guillén, Guilherme, Vilém, Vilmos, Guillermo, Gulielmus, Will, Gwilherm, Ghilherme, Gwilym, Viljem, Wiliam, Guglielmo.
We kept only synonyms with latin characters.

Then, once we had the synonyms, we built the dataset. The dataset consisted of positive and negative matches.
For the positive matches, we combined a synonym of the initial name, we did this for the initial first name and/or the initial last name. For example: Bill Gates - William Gates is a positive match since Bill and William are synonyms.
For the negative matches, we combined a first name and a last name that are not the same as the initial name. Bill Gates and John Gates are negative matches since Bill and John are not synonyms - even though they share the same last names: a difference in one of the names is enough to be considered as a negative match.
Therefore, once these two steps are completed, we obtained the following dataset:

Table 1. Sample of the dataset

Name 1	Name 2	Match
Bill Gates	Guillaume Gates	1
Bill Gates	Gates Vilmos	1
Will Gates	Gwilym Gates	1
Bill Gates	Bill Murray	0
Bill Gates	John Gates	0

The dataset obtained has 1 800 000 combinations of 400 000 names. The dataset was balanced between positive and negative matches.

3 Analytics

We built a deep learning model. The goal of the model was to be able to differentiate two names, if they belong to the same entity or not.
The model that showed the best performance is a bidirectional Gated Recurrent Unit (Bi-GRU) - based deep Siamese network model. A Siamese model is a type of neural network architecture that is designed to compare and measure similarity between two inputs - here, names. The architecture of a Siamese model (figure 1) consists of two identical subnetworks which share the same weights. Each subnetwork takes one name and processes it independently, generating a fixed-size embedding vector. These embeddings capture the essential characteristics of the input name. We then compared the outputs from both networks using a similarity metric, a cosine similarity. The final prediction will be a linear function of this similarity. The closer to 1, the more similar the data.

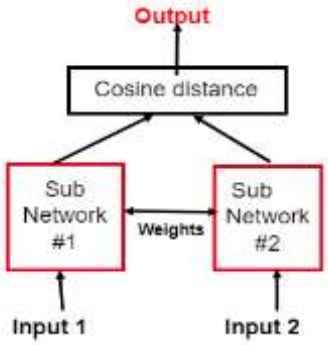

Figure 1. Simplified architecture of our siamese model

In order to evaluate our model, we used different metrics. In our project, to align our model with Société Générale's conservative position, we mainly took into consideration the accuracy and the precision as metrics to evaluate the performance. Precision measures the accuracy of the model in identifying positive matches. A higher precision indicates fewer false positives. Accuracy measures correctly matched entities out of the total number of entities evaluated.

Our aim was to reduce the false positive while not missing any more true negative (i.e. not creating false negatives). Therefore, it was a trade-off between false positives and false negatives: we were aiming for the lowest false positive rate that does not create false negatives.

We achieved the following performance:

Table 2. Performance of the model

	Validation dataset
Accuracy	96.6%
Precision	96.1%

Overall, we are satisfied with the final performance as it was obtained after a tedious process to test different architectures alongside different parameters. If time had allowed, we could have worked on improving further the performance.

4 Proposed solution and value creation

We have developed an automatic entity matching system for Société Générale that showcases high performance and offers the potential to partially replace manual checks. By implementing this system, Société Générale can reduce operational costs while effectively handling differences in names.

The value creation of our model lies in the cost reduction that comes from the reduction of manual checks. Human resources needed will be less important. Furthermore, by strengthening its checking pipeline, it will prevent Société Générale from having regulatory fines in the future.

Around 100,000 alerts are generated per year which lead to manual checks. Each 'dangerous' transaction has to be manually checked by a team of analysts at Société Générale. Each transaction is checked 3 times. 99.9% of the alerts are false positives. Thus, our goal was to build an efficient model that can handle name diversity and reduce alerts. The proposed solution will be able to cut off by 30% the number of false positives while not creating any false negatives (missing alerts).

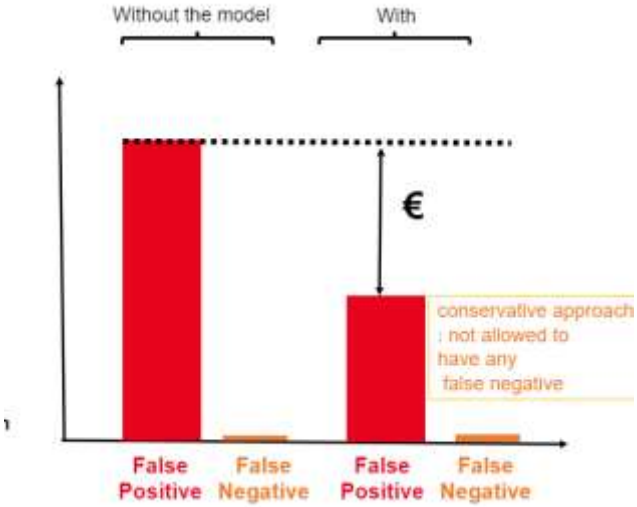

Figure 2. Value creation

Conclusion

By refining and expanding upon our system, Société Générale can stay at the forefront of entity matching technology, improving efficiency, accuracy, and risk mitigation. The ongoing commitment to innovation and improvement will ensure that Société Générale remains equipped with a state-of-the-art solution to address the evolving challenges in the financial industry as well as retain and improve its conservative position.

ACKNOWLEDGMENTS

We would like to thank Grigory Sharkov and Lamiae El Filali for the supervision and their accompaniment for these 6 months. It was a pleasure to learn alongside them. We all learned a lot during this project.

We would like to also thank Thomas Lewiner, who guided us throughout the project during the coaching sessions.

Thanks to Fragkiskos Malliaros, CentraleSupélec, ESSEC Business School and Société Générale for this partnership and giving us the opportunity to work on such a challenging and interesting project.