CRP Final Report

ILLUIN Technology

Sarah-Charlotte Gru

Yinzhe Huang

Brunehilde Senellart de Vrière

Zhengxiao Ying

Lauren Yoshizuka

ACM Reference Format:

Sarah-Charlotte Gru, Yinzhe Huang, Brunehilde Senellart de Vrière, Zhengxiao Ying, and Lauren Yoshizuka. 2022. CRP Final Report: ILLUIN Technology. In *Proceedings of Corporate Research Project (CRP '22).* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 ABSTRACT

This report provides the details, implementation, and rationale relating to the corporate research project (CRP) for the team assigned to the company, Illuin. The goal of this project is to deliver an end-to-end pipeline that facilitates the creation of a database from web-scraped data using Natural Language Processing (NLP) techniques. The database must be able to construct a usable and updatable visualization tool. Illuin can scale and repurpose this pipeline to fit any future business need they may have in the future. By clarifying the rationale in the design of the system design, Illuin has a clear vision of their data stream and can implement the pipeline seamlessly.

2 INTRODUCTION

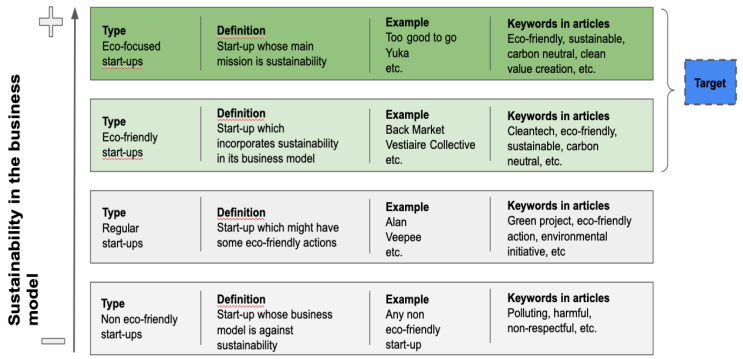
Illuin Technology is one of the leaders in France in the field of NLP, especially since the publication of the FQuAD research article on the first French dataset in Question Answering. Illuin continues to innovate in order to remain at the forefront of this field, both from an academic point of view and in the application of these innovations, in order to put research at the service of the greatest number of people.

Expected work and deliverables:

- Data scraping module that produces a usable dataset
- Exploration of different Machine Learning (ML) issues on said dataset
- Benchmark of different ML approaches and their limitations
- Interface for presenting results, updated in real time

3 PROBLEM DEFINITION

Detection of eco-friendly startups from news articles represented the initial goal of the project as requested by Illuin. In order to achieve this, we define "eco-friendly":



However during the course of our work we faced some problems regarding the classification of eco-friendly startups. First, though we established a grid to define the level of eco-friendliness, in reality it was difficult to assign a start-up in a category. Because the articles gives only a brief overview of the start-up work but not a complete

CRP '22, 2021/2022, Centralesupelec

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

explanation of its business model, we often faced the problem of not knowing whether the startup was actually eco-friendly focused or not.
Secondly, we also had some difficulties to find subsequent sources of eco-friendly startups. For example, we web-scraped BusinessWire website, with the keyword "green startup" however out of 250 articles that were web-scraped only 1 was somewhat eco-friendly. For all of those reasons, we decided in accordance with Illuin to switch our focus from greentech to another sector. As the interest of our work for Illuin was the pipeline rather than the project, they agreed on this switch.

We chose to focus our work on **biotech startups**. Biotech investment have risen tremendously in the past years, meaning that there is an increased need from V.C. firms to know more about the potential investments they could make in this lucrative sector. In 2021, 41B dollars were invested in biotechs startups and their global valuation has multiplied by 4.1 since 2017. Another reason which lead us to choose this sector, is that it is easily classifiable both for us and for the model. We do not have to establish a scale but rather a binary decision: is the startup a biotech or not.

After further discussions with the company, we discovered the main objective focuses more on the **creation of a functional and explainable workflow** that is capable of extracting data on the internet, classifying it depending on the problem, and transforming and outputting the relevant data.

The project objectives are formalized in the following list:

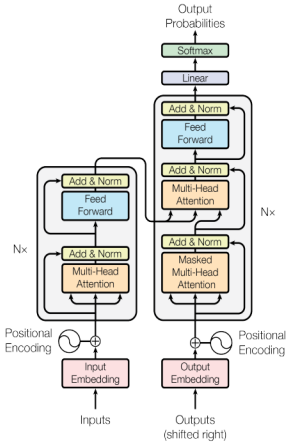
- Develop a multi-source scrapping and structuring pipeline for the extracted data
- Clean and prepare the data (outlier removal, NLP preprocessing)
- Analyze the extracted data, for example (depending on the result of the scrapping) :
  - Classify the companies by company size, sectors of activities...
  - Use an entity extraction model or a question answering model to recover company names, location, activities, revenue...
- Develop of a graphical dashboard interface to visualize the results and interact with the data

4 RELATED WORK

Extensive research is available in the realms of web-scrapping, NLP, BERT modelling, NER modelling, and QA modelling. Since some of us already had experience with web-scrapping, we forgo-ed further research into best practices in order to save time and begin work on data preparation and modeling tasks.

The bulk of our research was spent on learning more about models:
Binary classifier: When looking fot the most appropriate model for our project, it quickly appeared to us through related work, that the BERT model would be the best option.

First of all, through its architecture the BERT model has the capacity of understanding the context of any text.This was a crucial advantage for us, as some news article can be complex, and present a different meaning without any understanding. The bidirectional aspect of the model, allows to process both the text from end to finish and from finish to end. Through the paper "Attention is all you need" [2] we got the chance to have a better understanding of BERT structure shown in the following schema. We can see the importance of the encoder which embeds each words according to its importance compared to other words in the sentence. Then the decoder takes the encoder inputs and put it back to a text format.



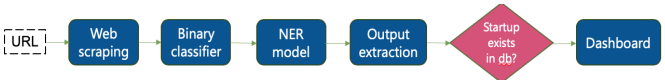
Through other readings of research papers we observed that the BERT model was quite performing when it came to text classification comparing to already state of the art algorithm. In their paper about comparing BERT against traditional machine learning models for text classification, researchers showed that BERT outperformed all tested models [3]. This research helped us to have an overall better understanding of BERT model, and to use it for both our NER model as well as binary classifier.

Finally, we thought it would also be interesting to look at papers on startup database. The research called 'Benchmarking Venture Capital Databases' helped us to understand the market, gave us key numbers in terms of payment, impact of potential competition for our tool[1]. If Illuin wants to go further with our project, we

believe that this paper can be also of great help to them to define better the existing offer and existing companies evolving in this sector.

5 METHODOLOGY

An essential step in the project was the construction of a workflow from data collection to final output creation:



5.1 Data collection - Webscraping

The first step of our system design was to get raw articles from different websites. Illuin provided us a list of start up news websites; for each site we first manually checked if it was scrapable, and then if we could find any relevant articles. We decided to focus on techcrunch, tech.eu, venturebeat, businesswire, crunchbase, and eu-startups.

The web-scraping python code is quite straightforward. First, the URL of relevant websites is manually obtained. Then the URL is sent to the data extraction crawlers that navigate to each article page and extract the raw data. The output results in the production of a folder that contains all the articles as text files. We used the library Beautiful Soup to write the web-scraping code.

Since some of us had already worked on web-scraping before, building the code for this step only required a few days. Finding the websites and checking manually the relevance of the articles took a bit more time.

We had to repeat this step several times, since we changed the type of articles we were looking for from eco-friendly to biotech startups. As a result, occasionally we experienced issues pertaining to the quantity of data needed to run the different models.

5.2 Data annotation

Our instinct was to use a supervised machine learning algorithm, hence the next step was to annotate the raw data. Illuin provided us with their company's annotation software, *Etiquette*, to annotate. We chose to highlight the **startup name**, the **founders**, the **investors**, the **creation year** and **location** of each news article's web-scraped text.

Since we had to redo the webscraping step, we also had to redo the annotations of all the articles. The first time, we started annotating around 200 articles, which is how we realized we the data deficiency problem of eco-friendly startups. This lack of data prompted us to pivot to biotech startups.

Moreover, we needed to agree on how we would annotate: should we highlight every time we see the start up name? Sometimes in the articles only the CEO name was mentioned, should we highlight it as founder? If multiple start-ups were in the same articles, should we highlight them all? In many cases, there are more than one investor; do we tag every investor mentioned?

We agreed on some rules: for example, we decided to put every C-suite executive in "founders".

The second time, we annotated 350 articles related to biotech startups. We felt this was a better foundation of data quality for our models.

The annotation part was very time consuming, as we had to do this manually for every article. However, it is a vital step in the model, so we spent a lot of time ensuring our work here was consistent and thorough. The performance of the model actually depends on the quality of the annotation. Thus, these decisions we agreed upon regarding the annotation methodology affected the output to quite an extent.

5.3 Classification

The next step was to classify the raw data into two classes: related to the field we wanted to focus on or not.

At first we thought about only filtering on predefined keywords but the real problem with filtering is that it does not take the meaning of a sentence or the context of words into account. Under the guidance of our coach Thomas, we decided to use a supervised model to classify our text files into biotech or not.

Bidirectional Encoder Representation Transformers (BERT) is a pre-trained model that helps machines learn to decipher the representation of text with respect to context at very high accuracy rates. The BERT architecture is composed of several transformer encoders stacked together.

Thomas suggested a BERT model, because cutting edge results on NLP are obtained with transformer models.

To increase the performance of our model we split the data into train and test sets (25%) and did some **hyper-parameter tuning** on the number of epochs, activation function, batch size, and layer dropout.

We also improved the performance by changing the balance between biotech and non-biotech articles. Initially, the model performed better when we had a balanced dataset, but then we managed to obtain an overall accuracy of 91% on an unbalanced dataset. We will detail more the performance of the model in the Evaluation section below.

The classification model took a lot of time as we had to find the best classifier, tune hyperparameters and run multiple times on different datasets as we added more articles.

By default, our model used only 5 epochs, a batch size of 32, and an output space dimensionality of 1. With 700 articles, we had an overall accuracy of 73%.

We changed the different parameters to see how we could improve the performances.

- **number of epochs**: By increasing the number of epochs, we manage to improve the performances. However, it also takes a lot more time. With 30 epochs, the model runs in 1 hour and half and performs very well.
- **activation function**: the best performance is obtained with the Sigmoid function. For the same parameters (10 epochs, batch size = 24 and dropout 0.1), we tried other activation functions, like ReLU (maximum performance of 50%),softmax (maximum performance of 50%) and sigmoid which obtained an accuracy of 81%.
- **batch size**: Reducing the batch size to 24 helped increase the performance of the model. We tried reducing more, to 12, all parameters equal, and it increased the accuracy but it also increased the number of False Positives. Depending on the use case and business recommendations, leaving it to 24, to lower the number of False positives, might be a better solution. Increasing the batch size to 60 again increased the number of False Positive.
- **layers dropout** We reduced the dropout from 0.1 to 0.001 and not only it decreased the accuracy but also lower the performance with a poor confusion matrix (50% recall). On the other hand, increasing it to 0.9 reduced the accuracy to 67% but also lowered the recall for class 0 and precision of class 1 to 48% and 62% respectively.

We adjusted these parameters, and the best overall accuracy we obtained on the test set is is of **91%**.

5.4 Data Extraction - NER model

After classifying the articles containing essential information we needed, the next step was to extract the information from the plain text. As mentioned in the literature review, the tasks of Named Entity Recognition are well-researched. Our work in NER can be split into several steps:

Data Model Definition: the data model indicates the kinds of entities to be extracted. In our case, in order to build a database containing the key information of startups, we defined five entity types:

- *The names of start-ups*: the long or short format of the company names, helping the users of the database to firstly recognize the start-ups.
- *The investors of the start-ups*: the investors who at any time had stakes in the start-ups, helping the users of the database to know the fundraising means of the start-ups.
- *The founders of the start-ups*: the founders or top-management of the start-ups, helping the users of the database to know the managerial groups of the start-ups.
- *The location of the start-ups*: the base of the companies, both country-wise and city-wise, helping the users of the database to evaluate the geographical locations of the start-ups.
- *The founding year of the start-ups*: helping the users of the database to know the maturity of the start-ups.

IOB Encoding: the annotated files include the entities and their positions in the articles. However, the model can not directly utilize such data. In the training set of the BERT-based NER model, x should be the tokenized text, while the labels should be the class of each token. However, in many cases, the entities consist of more than one token. Therefore, the labels should also indicate the position of the chunks. Here are the explanation of I, O, and B:

- The I- prefix before a tag indicates that the tag is inside a chunk.
- An O- prefix before a tag indicates that a token belongs to no chunk.
- The B- prefix before a tag indicates that the tag is the beginning of a chunk that immediately follows another chunk without O tags between them.

Model Construction: In our NER model, since the scale of annotated data is in the hundreds, we conducted transfer learning in order to utilize the pre-trained power of BERT model. The model we fine-tuned is BERT-Base, containing 12 layers, with 768 cells

in the hidden layers each, and 12 multi-attention heads, i.e., 110M parameters in total. Since we are fine-tuning the pre-trained model in the startup scenario, we set a relatively small learning rate of 1e-5. Thanks to the computation power of NVIDIA Tesla P100, we can fine-tune the model within 1 hour.

Tricks to Improve the Performance: Initially, although the model claimed a high overall accuracy, the model performs relatively poor in terms of encoding some of the labels. With the support from both the coaching side and the company side, we performed several adjustments to improve the performance in all labels:

- *Data quality improvement*: In order to improve the model performance in token classification, we must teach the model which contexts would display the particular entities. Therefore, after the first experiment of NER task, we re-annotated 350 biotech articles in which the data quality is carefully checked. Well-equipped with high-quality data, the foundation of the NER task was strengthened.
- *Information density improvement*: As we noticed that even though the articles contained essential information, some of the sentences did not show any key/relevant information for the model. After removing the sentences with all *O* labels, the density of information improved.
- *Sentence Re-organization*: Although all the remaining sentences contained relevant information, we realized that inputting the model one sentence at a time might actually weaken the model's performance. This is because the max length of input for BERT-Base model is 128 tokens, while some of the short sentences only contain tens of tokens. Therefore, we combined the sentences with the 128-max-length restriction in order to squeeze the prediction power of BERT-Base model.

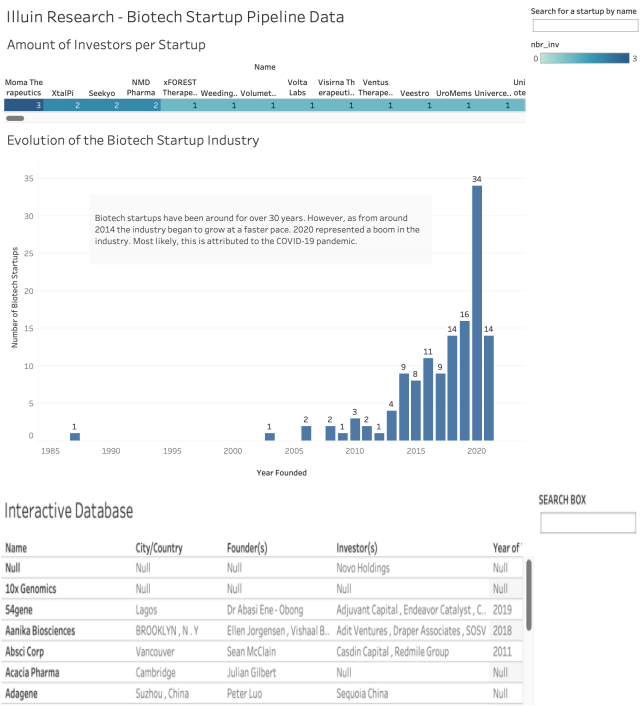
5.5 Output creation - Tableau dashboard

The last step was to create a user-friendly output. We decided to do it on Tableau as we were all comfortable with this software, and because it is a popular data visualization tool for many companies.

The visualizations in Tableau were built based on the output "database" of the combined BERT and NER models. This database is in fact an excel file that is connected to the Tableau file. This Tableau output is useful for both the customer user and Illuin. For the customer, they can type any possible field, such as the name of the startup, year founded, investor, founder, or country, and all the related information appears on the screen of the Interactive Database. As for the dashboard, this is more of quick overview for the user to get a fast understanding of where the biotech market is at. This dashboard has information about number of investors per biotech startup as well as the growth in biotech startups year-over-year.

For Illuin, the Tableau file helps to monitor the KPIs and the performance of the system design. Additionally, as this project is aimed to be scalable and adaptable for Illuin's future needs, the creation of a simple yet informative dashboard made the most sense.

The final Dashboad and interactive database in Tableau look like this:



5.6 Joining models

Throughout the duration of the project, the work was divided between group members, thus the very last step was to join all the models and try running it. This took some time as we had to change some output formats as well as the dataloader. We then had to evaluate the end-to-end performance of the entire pipeline.

6 EVALUATION

6.1 Data collection

For the data collection part, the quality evaluation was done manually. Indeed, we assessed the data quality, that is composed of completeness, consistency, conformity, accuracy, integrity, and timeliness while annotating.

6.2 Data classification

For classification, we needed to choose a proper evaluation metric. At first, we focused on the accuracy, but then we focused more on the F1 score per class and confusion matrix.

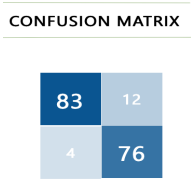
By default, our model was used only 5 epochs, a batch size of 32, and an output space dimensionality of 1. With a balanced dataset, we had an overall accuracy of 73%.

We adjusted these parameters, and the best overall accuracy we obtained on the test set is of **91%**.

	precision	recall	f1-score	support
0	0.874	0.954	0.912	87
1	0.950	0.864	0.905	88
accuracy			0.909	175
macro avg	0.912	0.909	0.908	175
weighted avg	0.912	0.909	0.908	175

We managed to obtain these results with 30 epochs, a batch size of 24, dropout of 0.1, binary cross entropy as loss function, and Adam optimizer.

The associated confusion matrix for these results is as follows:



As is shown, the model performs quite well on both classes. With this model, we managed to get 350 biotech classified articles to run in the following NER model.

One key question around this part of the workflow was on the cut of predicted class. Indeed, in the model this rate is at 50%, meaning, for example, that if the model predicts a class of 0.6, it actually belong to class 1. We tried changing this rate to see its effect on the confusion matrix and the performance of our workflow. We reduced it to 45%. The results were not the same depending on the size of the database but overall, it increased the number of false positives.

Depending on the cut, type I and type II error vary. The higher the cut %, the higher false negative we get and the opposite when it gets lower. We advise in terms of business to **limit as much as possible false positive errors**, and thus to keep a high cut rate.

We also observed similarities between the type I errors articles:

- **Type I errors** Most of the type I errors have some "health" related words. Examples: lifespan (recycling start-up), healthy (food tech)
- **Type II errors** There seem to be no explicit reasons for type II errors. Some of them have biotech words some don't, some are short and some are long and complex. They usually have a prediction probability between 0.4 and 0.5.

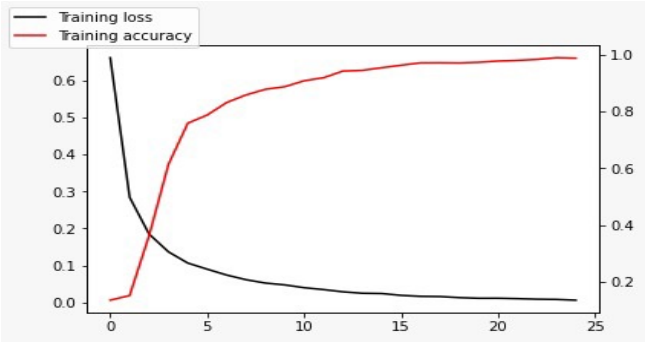
6.3 NER model

Since in the articles, the label *O* takes a share of more than 80%, we cannot rely solely on the accuracy to evaluate the performance of the model. Instead, we evaluate both the precision and the recall, so that the model can both classify the exact labels in the right way and avoid misclassifying the labels. With precision and recall being calculated, we can use the F1 value to evaluate the model performance. The F1 score of each label can be seen in following picture:

	precision	recall	f1-score	support
B-f	0.77	0.92	0.84	77
B-i	0.76	0.74	0.75	105
B-p	0.61	0.74	0.67	50
B-s	0.68	0.77	0.72	92
B-y	0.93	0.88	0.90	32
I-f	0.80	0.92	0.86	85
I-i	0.85	0.76	0.80	137
I-p	0.27	0.26	0.27	23
I-s	0.69	0.86	0.77	36
I-y	0.00	0.00	0.00	1
O	0.99	0.99	0.99	8574
accuracy			0.97	9212
macro avg	0.67	0.71	0.69	9212
weighted avg	0.97	0.97	0.97	9212

It is clear that the model can correctly classify most of the label except one: I-y. I-y means the second or latter tokens of an entity "year". However, most of the entities "year" are single token entities (as a year is simply a 4-digit integer). After checking the source of I-y, we found that it is because some of us took both the month and the year into consideration when annotating the data. With only several examples, the model can hardly learn how to classify such a label. Therefore, it failed to recognize the year label in the validation set. It can be fixed by removing all the mis-annotations. In general, we can conclude that our model has good performance in extracting key information in the start-up context.

Here we can observe the accuracy evolution according to the epochs
:



6.4 End-to-end performance

Since the goal of our project is to create an end-to-end tool of from data collection to data visualization. It is important to also evaluate the transaction rate of the whole pipeline. In order to evaluate the end-to-end performance we built a pool of 40 articles for the evaluation set. After filtering the articles through the binary model, only 17 remained (100% biotechs with no type 1 error, 93% of accuracy). For the 17 articles identified by the classifier and then processed by the NER model, 76.5% of all labels are correctly identified. More specifically, out of these 17 articles, startup names in every article, founder names in 6 articles, investor names in 15 articles, city/country information in 12 articles and year of foundation in 15 articles are correctly labelled. Also, information in these 17 articles have all been successfully added to the existing database, which is built based on the previously annotated articles.

7 BUSINESS RECOMMENDATIONS

The aim of the project is to have a clean, up-to-date database of new startups. In order to increase the business value of our tool, we recommend to Illuin to develop the following axis of research.

- Labels: Develop the NER model with more labels which are essential in the investment world such as the valuation, the funding round, etc. Through those, the database will have a better impact and will be more useful in the real world.

- Classifier: Develop a sector classifier. Even though our classifier only applied for biotech, in order to maximize the tool's reach, a sector classifier seems essential. It will be one of the main requests of potential client.

- Sources: Develop the source pool and prevention of bias. The more qualitative the sources are, the better the database is. The more feature engineering done on the text of the sources, the better the database is. Thus, to keep finding and working around the sources is essential.

Through those axis of development we could create a complete tool which can be of value for multiple potential clients as it can respond to multiple types of requests at a type unlike our current tool which is only applicable to biotech startups.

Further we would advise to insure the quality of our database, to establish some Key Performance Indicators. The following KPIs we think can help with this goal.

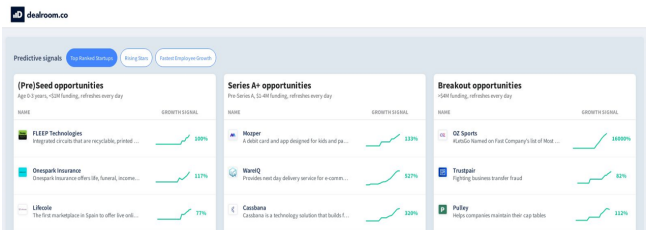
- **Up-to-dateness**: we will define with the client the frequency at which the entire workflow must be run to update the database. In the dashboard, the client is able to see the date at which the model was run.

- **Consistency**: Check regularly if we have redundant value.

Our project has a concrete business value behind it and can **generate some revenue streams for ILLUIN**. Through our model we can propose a tool which responds to an existing need of potential clients which are both the V.C. firms and database companies.

We analysed the current market of startup database and we observed a simple market with two main actors : the clients and the providers. The clients are mainly composed of V.C. firms, but also journalists and researchers. They usually want to have access to those database to learn more about the current state of the startups sector in a specific field, to conduct some analysis such as due diligence or to analyse the investments of competition. On the other hand, providers are made up of companies whose business model is to create complex and well documented startup database.

Additionally, the following image shows the type of products proposed by Dealroom, one of the actor in this sector. As we can see, the information are well detailed, we can find along basic information (name, sector, etc.) some analysis (last news on the company, etc.). To have access to an entry-level service of this type, V.C. firms pay for example 50.000 dollars per year for CB-Insights Analytics database.

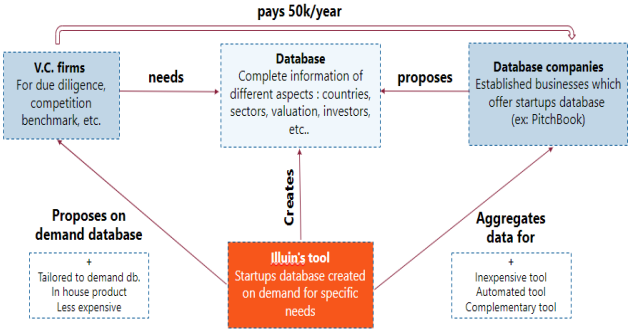


Illuin's primary business model is to propose AI tool such as chatbot for companies in the need. Based that and on the observations we described earlier, we see a clear market opportunity for Illuin and the tool we developed. Indeed, in this market of startup database, there only exist a general offer which is quite expensive. Illuin could disrupt this market by proposing a tailored and on-demand database or tool depending on the client. We think it could be interesting to propose Illuin's services for both the V.C. firms but also for startup database companies (such as Dealroom, CBInsights, etc.).

First, we would make sense to propose to propose a tailored on-demand database to V.C. firms through the tool we developed. It would be a simpler version compared to the current offer of the market, as it would not have any analysis but only the primary most important information: startup name, country, investors, etc. We could easily propose this product at a much lower price than competition (50k/year for entry-level product) as the cost associated with its construction are quite low. Illuin would only need to get the pipeline running, to have the final result. Additionally, the tool can easily be on demand. Through the labels and the classifier we can target the client's wishes. For example, it would be possible to build a database of only foodtech startups founded after 2020 in the UK with valuation lower than 25millions and who have already undergo with funding round A. Offering such a tailored product would be time saving for customers as they do not need to go through the entire database of more generalist database. Overall, through this offer we could disrupt the market by creating two categories of products : high-end generalist expensive database, tailored simple affordable database. It would create new demand that is not already on the market such as niche small V.C. firms which can't afford database currently proposed.

Secondly, we think it would be a good idea to propose Illuin's services to companies which create those startup database. Based on the high fee they charge, we can imagine that they have high costs, mainly related to research and aggregation work. Through the pipeline we built, we could cut the human time spent to do this aggregation work. We could propose this tool to recover the basic information of startups (name, investors, etc.) online, which would allow the company to focus on the technical analysis product. This would cut their costs, but also allowed them to develop more a high-end product. This would reinforce this idea of two very differentiated products on the market.

This schema illustrates our business recommendation in terms of market penetration with our tool.



8 CONCLUSION

During this 6-month Corporate Research Project, we worked for ILLUIN Technology and created an end-to-end pipeline to create a database from news articles on the web. The database gathers information from 350 biotech startups such as the name, the founder's name, the investors, the year it was created and the location.
Our pipeline is composed of webscraping of new articles then a BERT binary classifier (biotech focused or not), then a NER model to extract relevant information and finally a dashboard where the user can have access to the database and to monitoring of KPIs.
We managed to obtain really great performance for each models (BERT and NER) and an overall end-to-end performance of x? We also established data quality indicators to determine at which frequency the pipeline must be run and on its usability.
This pipeline has concrete business value and can generate some revenue streams for ILLUIN.
ILLUIN can propose on demand databases for VC firms in need for complete information on new startups but also it can create database for companies whose business model is based on database creation. VC firms would go to ILLUIN to get a shaped-on-demand database, which would be less expensive than going to a Database Companies, while Database Companies would be interested in ILLUIN tool for automation or to complete their pre-existing databases.

REFERENCES

[1] Reinger Braun Andre Retterah. [n.d.]. Benchmarking Venture Capital Databases. ([n. d.]). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3706108
[2] Niki Parmar-Jakob Uszkoreit Llion Jones Aidan N. Gomes Lukasz Kaiser Illia Polosukhin Ashish Vaswani, Noam Shazeer. [n.d.]. Attention is all you need. ([n. d.]). https://arxiv.org/pdf/1706.03762.pdf
[3] Eduardo C. Garrido-Merchan Santiago Gonzalez-Carvajal. [n.d.]. Comparing BERT against traditional machine learning text classification. ([n. d.]). https://arxiv.org/pdf/2005.13012.pdf