3. News Analysis

3.1.Problem Statement

As the information regarding to assets becomes available, the pricing indexes are adjusting accordingly, and the quicker you take the action, the more benefit you will be getting. Retrieval of this information is mostly based on the financial news that are presented by known news agencies or the economists' tweets or blog posts. Scraping this information and predicting the impact may warn the user to take an action.

This project was given by the internship department of Yapı Kredi Technology as a team project consisting of five interns. It is aimed to scan certain news sources to be used in investment decisions, credit evaluations and financial audit processes and to create a database by recording the contents. Then, the content that is collected will be processed by the NLP tools and prediction will be made regarding to named-entity and sentiment analysis.

Many investors use the information they have obtained by learning about the events taking place in the country and the world when making their investment decisions. While the difficulty of accessing information is already present here, it is necessary to produce the presence of alarm systems to make quick decisions. Me and my teammates tried to come up with a product that analyzes the news in an end-to-end way. If it is needed to be explained the product we want to make from the eyes of the users, when the user entered the site, he/she can easily access the title of the news entered in the last few days, the sentiment analysis result, and the NER labels when they click on the news to read. In this way, the user will be able to perform news scanning and analysis within a few minutes, which may take a long time, and will be more active when making investment decisions. After seeing whether the news is negative or positive at the first stage, the user will receive a good service when accessing the information they want, as the data in the content such as name, commodity, country are determined by NER analysis.

Named Entity Recognition, or NER, is a component of the NLP discipline of information extraction. The entities (information) in the text are detected, recognized, and categorized into specified groupings like person, time, organization, etc. It is used to pinpoint the object affected by the action or circumstance in the news using text data that has been scraped. Because of this, we required categories for things like people, money, organizations, events, and products.

Sentiment analysis is a technique for determining and categorizing opinions expressed in a text, especially to determine whether the author has a positive, negative, or neutral attitude towards a specific issue. It is utilized to analyze the impact of the news on the product that is retrieved from the NER engine in the text data. Due to the project's constrained timetable, trained models are utilized for both the NER and Sentiment Analysis portions of the project.

My task in this project was to develop the news crawler module and take responsibility for the design of the system. Initially, I created a database by scraping news from some news sites and tagged a small data for the training set. This data, which will be used in NER and sentiment analysis modules, was useful for us. In previous studies, there were already developed news-analysis models, but the crawler and frontend parts were not. In an example project we found; simple libraries were used rather than high-performance models.

The criteria that the company expects from us in terms of success is that we achieve a high success rate by fine-tuning the NER/sentiment models. At the same time, it was important for us to offer the service and interface where the user can easily access the right information.

3.2.Tools and Techniques Used

In this project, we aimed to develop four different modules: configurable-crawler, sentiment-analysis, named entity recognition-analysis, user-interface. After our meeting with six trainee supervisors in the team, he was assigned to develop these modules. While I was responsible for the configurable-crawler part, Yağmur and Bengi worked for sentiment-analysis, Oğuz and Erkam for NER-analysis, and Berk for user-interface. Our supervisor met with us on a weekly basis during the development of the project and listened to the developments from us. Daily sync meetings were held for the project for the team members working together to learn about the developments in other modules.

I was responsible for the development of the configurable-crawler module. To talk about my willingness to take on the task, I've done some bot development scraping using selenium before, but these were amateur works that I did for hobby purposes. As a result of the research, I did at the beginning, I chose scrapy, which is a library that I can use comfortably. Also, I used the Python language because I felt more comfortable with big data management and because it hosts the scrapy library. This application, which easily extracts the information on the site by creating a spider, offered more advantages than other tools due to its speed and scalability. It didn't take long for me to become proficient as a result of following the tutorials and I got to work.

One of the biggest challenges I faced was the inability to crawl dynamic sites. As you scrolled down on such sites, more and more news were being loaded and the scrapy tool could not detect the news. That's why I used news sites that work with static and pagination logic in the crawling process. In this way, after defining the classes on the news site, the spider would scan page by page and write the news in JSON format and collect it to create a large data set. Also at this stage, due to the problem that long news contents may cause in sentiment analysis, we decided that the summary should be included in the dataset while scraping the news with the nltk library. It also made sense to provide the user with a brief overview as an alternative. The accumulated news was labeled by me so that other models could be developed later. Thus, the machine learning models found were tested by the team.

Later, in order not to overload news sites and not to collect illegal news, RSS Feeds provided within the legal framework were used. Although I encountered many problems while developing the crawler module, which is intended to be generic and configurable, RSS Feed has made our job easier. Two different options emerged with the generic module that performs data mining on pagination-based news sites. The development I mentioned at first was more productive for creating the database. The alternative module using RSS Feed, on the other hand, was used in the project itself, offering a faster and more efficient solution. At the end of the crawler process is done, data is stored in the JSON format categorized like author, text, summary, etc. to pass them to other modules.

The news content in the dataset from the crawler part in JSON format included information such as name, entity, non-governmental organization, and this information had to be highlighted on the user-interface after labeling. The named-entity-recognition module that Oğuz and Erkam worked on was developed with spaCy. Labels in the news content, such as name, country, non-governmental organization, language, date, which will provide easy readability to the investor, were created with spaCy. The en_core_web_sm, a model trained with a large dataset, was fine-tuned, and made suitable for the project. One of the most common problems here was entity redundancy. After filtering these, 18 different labels were created. In addition, a more effective solution was produced for the user interface during labeling with the 'visualizer' in spaCy. Tests on documents satisfied my teammates and supervisor. Finally, the module encapsulated with a small and fast pipeline, job was ready.

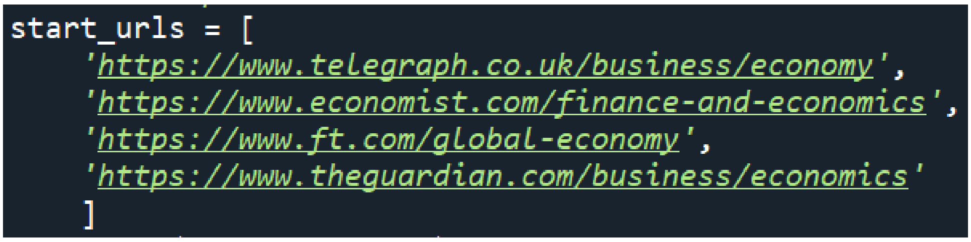
Thanks to sentiment analysis, mood in long news texts would be labeled as positive, negative, and neutral. For this, it was recommended to us by the supervisor that the models that have already been developed should be investigated. In the sentiment analysis module developed by Yağmur and Bengi, the mrm8488/distillroberta-finetuned-financial-news-sentiment-analysis natural language processing model was used. They also used models on servers that include high GPU-providing harware so that the models can run quickly. Financial_phrasebank consisting of 2264 sentences and 3 labels and nickmuchi/financial-classification containing 5057 sentences and 3 labels were used as dataset. He also helped develop the different finbert-based NLP models ProsusAI/finbert and ahmedrachid/FinancialBERT-Sentiment-Analysis. However, it continued with Nickmuchi's model, which obtained a better score later on. With the news tagged by me during the crawler development mentioned earlier, the fine_tuned process was performed at this stage.

At the last stage, it was aimed to combine the sentiment and NER outputs of the dataset that emerged after being crawled by combining these modules, and to create a user interface by creating an API. Here Berk and I worked on connecting these modules. The process of establishing and encapsulating the libraries used in the studies was easily completed. Later, Berk created the final version of the application using node, express, and react to develop the frontend part. Then, Berk stated that Node.js is used because of containing better tools like React.js and it is easy to handle writing pipeline and middleware. With the provision of daily news coverage, 100 titles will be listed on the site and sentiment analysis will be featured on the main page. In addition, when these headlines are clicked, long news content tagged with NER and a news summary easily created with the nltk library are presented to the user.

3.3.Detailed Explanation

In this section, which will include technical details, I will explain how I developed the configurable-crawler module, which I was responsible for in this project, step by step. Before technical details, to briefly explain web scraping workload used in project from beginning to end, the process of obtaining and extracting data from websites is known as web scraping. A collaborative and open source web scraping framework is called Scrapy. It is an open-source framework with thorough instructions and examples. To gather data comprising information about news, such as title, author, date, link, content, and summary, Scrapy employs RSS Feed (a file that provides information about the contents of the website). The nltk library is used to generate a summary from the news content if the news does not already contain one so that the user does not need to read the entire news item in order to comprehend the subject. Finally, a JSON file containing all the obtained data is created and delivered to the named entity recognition and sentiment analysis algorithms for forecasting. The websites Wallstreet, Bloomberg, and The Guardian were chosen as the sources for the financial news since they offer reliable information on financial accounts. Daily JSON files are created by converting scraped financial news from close to 100 sources.

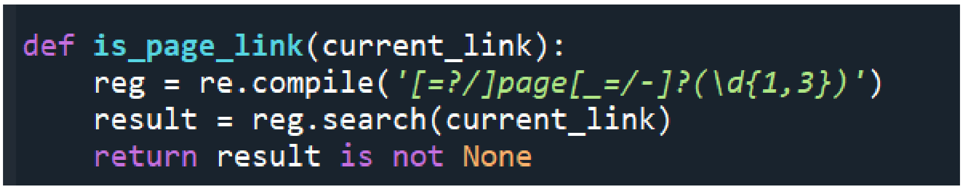
With the scrapy library I mentioned in the previous paragraph, I learned how to access and benefit from the HTML codes on the pages with tutorials. First, I determined what we needed by examining the news sites during crawling, and after discussing this issue with my teammates, we decided that the title, author, date and text information was sufficient, as in Figure 1. In order to retrieve the information in Figure 1, we needed the category definition in the HTML code and this varied from page to page. That's why we used the Article object of the newspaper library to reach a generic result, and I could easily scrape this information from the site where the news was located. But the main problem was to reach these links on the news site. First, I collected pagination-based category sites that statically loaded news and accumulated them. In Figure 2, you can see the links of the sites that lead to the economics and finance categories we are looking for.



```
start_urls = [
    'https://www.telegraph.co.uk/business/economy',
    'https://www.economist.com/finance-and-economics',
    'https://www.ft.com/global-economy',
    'https://www.theguardian.com/business/economics'
    ]
```

Figure 2 – Start URL’s to crawl

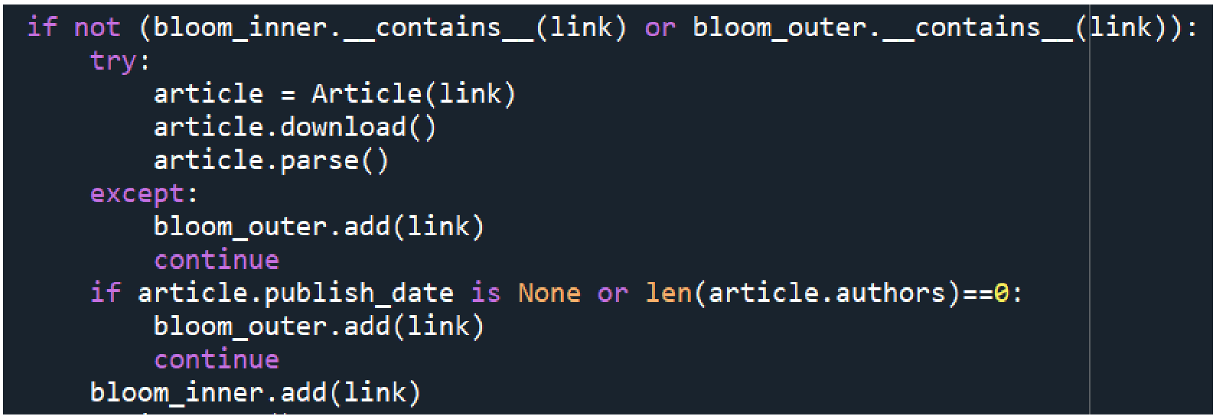
Based on selected categories of news sites and newspapers, it was necessary to collect links from HTML codes for pagination-based crawling. At this point, the difficulty of reaching a generic result made me spend a few days. Lastly, together with my supervisor, we decided to pull out all the links and do filtering. There were two different problems here: defining the link with the news text and the link to the next page. In order to detect the news link, the Article object I mentioned before is created with the link and put in the exception handling block, and newspaper library is used. In order to define the page links, it is questioned in Figure 3 by defining regex depending on the url configuration on most sites.



```python
def is_page_link(current_link):
    reg = re.compile('[=?/]page[_=/-]?(\d{1,3})')
    result = reg.search(current_link)
    return result is not None
```

Figure 3 – Function to determine page link

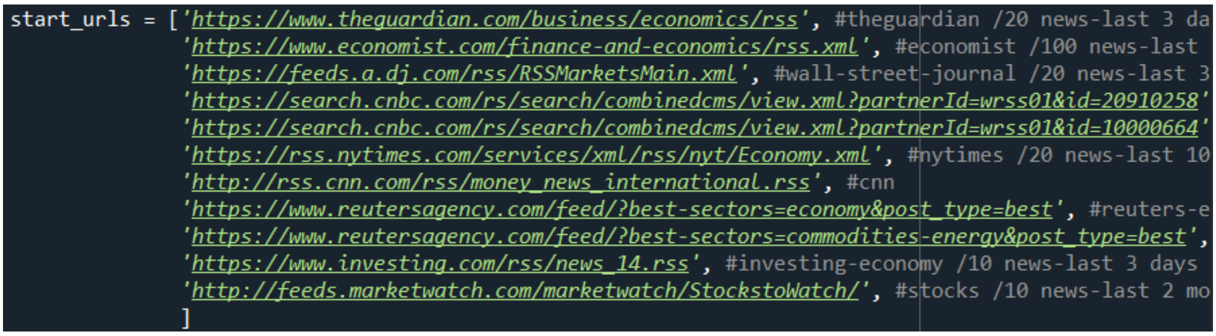
After all the links are collected on the category page, the URL indicating the next page is determined and assigned to the variable. At the same time, if all links are news texts, collecting and saving information by querying the newspaper library is done in this try-catch block. Of course, since all links on the news site will be queried during this process, many URLs out of necessity reduce performance. During this study, about 400 links were collected, while only 20 of them were news links, only 1 of them was the link of the button that leads to the next page. The rest of the links can be given as home page, different categories, advertisements, images. Therefore, in order to increase performance, an Article object should not be called by remembering links such as advertisements that were previously excluded from the news link in every category URL containing many news. Therefore, when the crawler moves to the second page, it will no longer query news such as homepage and advertisements again. At the same time, one of the problems encountered with the experience was the repetitive storage of news. Therefore, taking the confirmation that the instant link has not been queried before will increase the performance. So I started integrating the Bloom Filter implementation into the code as the fastest solution.



```python
if not (bloom_inner.__contains__(link) or bloom_outer.__contains__(link)):
    try:
        article = Article(link)
        article.download()
        article.parse()
    except:
        bloom_outer.add(link)
        continue
    if article.publish_date is None or len(article.authors)==0:
        bloom_outer.add(link)
        continue
    bloom_inner.add(link)
```

Figure 4 – Bloom filter implementation

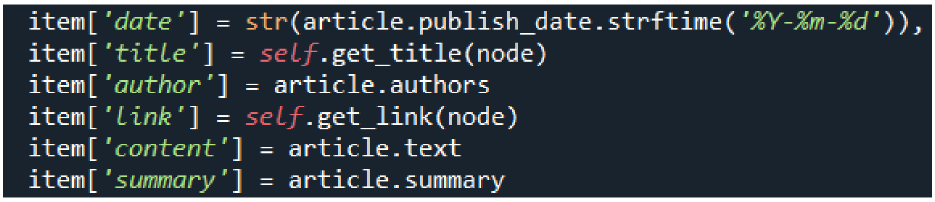
The main purpose of the integration of filtering and Bloom Filter was to be able to crawl more information in the fastest way by increasing performance. For this, I decided to keep the links to unwanted sites such as homepage, category, and advertisement in a global bloomfilter object, valid for the entire news site. Next, I created the local bloomfilter object to get the news texts on each category page only once. In this way, I got the performance increase running 4 times faster than the list implementation. In the code in Figure 4, if the link variable is not included in bloom_inner and bloom_outer, the Article object is used to query whether the link contains news text and if it does not, it is added to bloom_outer. If it does, the scraping process starts after it is added to bloom_inner. This spider, which I originally developed for Crawler, helped us to create a large dataset. The dynamic sites here are not crawled, my supervisor told me that they should be ignored. In the continuation of this title, I will explain the new spider created with RSS Feeds.



```
start_urls = ['https://www.theguardian.com/business/economics/rss', #theguardian /20 news-last 3 da
              'https://www.economist.com/finance-and-economics/rss.xml', #economist /100 news-last
              'https://feeds.a.dj.com/rss/RSSMarketsMain.xml', #wall-street-journal /20 news-last 3
              'https://search.cnbc.com/rs/search/combinedcms/view.xml?partnerId=wrss01&id=20910258'
              'https://search.cnbc.com/rs/search/combinedcms/view.xml?partnerId=wrss01&id=10000664'
              'https://rss.nytimes.com/services/xml/rss/nyt/Economy.xml', #nytimes /20 news-last 10
              'http://rss.cnn.com/rss/money_news_international.rss', #cnn
              'https://www.reutersagency.com/feed/?best-sectors=economy&post_type=best', #reuters-e
              'https://www.reutersagency.com/feed/?best-sectors=commodities-energy&post_type=best',
              'https://www.investing.com/rss/news_14.rss', #investing-economy /10 news-last 3 days
              'http://feeds.marketwatch.com/marketwatch/StockstoWatch/', #stocks /10 news-last 2 mo
              ]
```

Figure 5 – Start URLs of RSS Feed

As a team, in one of the meetings we held with the supervisor, we decided that it would be easier to scrape daily news with RSS Feed. One of the problems affecting this decision was that we had to load less on the domain in the process of data mining from news sites. Another is that most news sites impose a news reading limit and then make it mandatory to subscribe to paid subscriptions. Although the subscription barrier was overcome, it was not a method that we wanted to reach information in an unethical structure. That's why I focused on RSS Feed services provided by news sites. Most news sites were providing users with 20 recent news free of charge via the feed. A wide range could be created by using this service of more than one news site. The links of the RSS Feeds we used during development are in Figure 5. Configurable addition and removal operations can be changed according to the wishes of the user. The page layout that appears in the browser when The Guardian is clicked on the links is in Figure 6. The title under the Item object was scraped directly and the information was extracted, and the URL was passed to the Article object. In this example, the Article object is needed because the news text in the description attribute is missing. With the newly created spider, the site was easily crawled and the information was saved in JSON format in Figure 7.



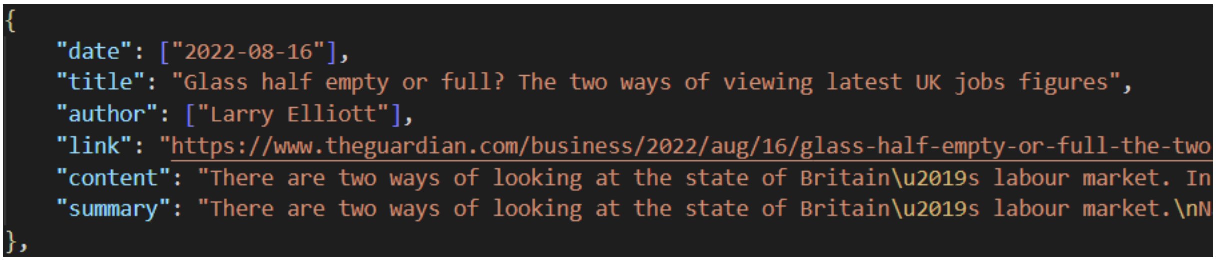
```
item['date'] = str(article.publish_date.strftime('%Y-%m-%d')),
item['title'] = self.get_title(node)
item['author'] = article.authors
item['link'] = self.get_link(node)
item['content'] = article.text
item['summary'] = article.summary
```

Figure 7 – Format of the JSON extraction

It really made my job easier to use Spider, which is included in Anaconda, which I used during the development phase. It was comfortable for me to be able to develop from the terminal after I had easily accessed the libraries. Being able to edit middleware and pipeline with ScraPy, as well as creating spiders in XML format were some of the important features for me. In summary, large datasets were created with the first developed spider and used during development as needed. The RSS Feed-based spider, which was developed later, would provide news capture with periodic flow and scrape the news according to the time it was run. In addition, with these stored news, a large dataset will be formed in the process.

3.4.Results

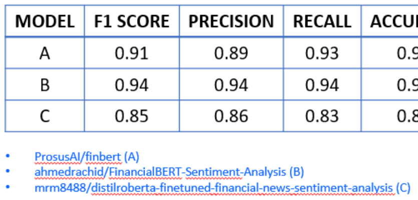
After the modules were completed by me and my teammates, the modules were combined and the end-to-end product emerged. If I need to start from my own module, 100 outputs like the news dictionary in Figure 8 were sent to NER and sentiment analysis in JSON format. As can be seen in Figure 8, the date, title, author, link, content, and summary of the news are stored. One of the most recent problems with the configurable-crawler module was the inconsistency of the RSS Feed sources with each other. For example, a news site that offers 20 news was doing this by choosing from the last 1 week, and another news site was providing the service of 30 news in the last 2 days. However, this situation was not perceived as a problem by the supervisor and it was requested to stay that way.



```
{
    "date": ["2022-08-16"],
    "title": "Glass half empty or full? The two ways of viewing latest UK jobs figures",
    "author": ["Larry Elliott"],
    "link": "https://www.theguardian.com/business/2022/aug/16/glass-half-empty-or-full-the-two
    "content": "There are two ways of looking at the state of Britain\u2019s labour market. In
    "summary": "There are two ways of looking at the state of Britain\u2019s labour market.\nN
},
```

Figure 8 – Output data unit fromcrawler module

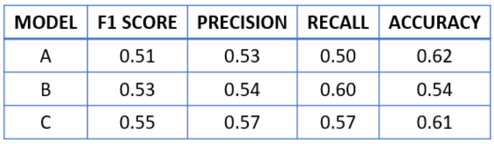
On the Sentiment analysis side, Yağmur and Bengi first tried to find the best model by running the models on 500 test datasets and comparing their scores. Figure 9 shows the model scores on the financial classification dataset as well as the model scores on the team-labeled dataset. Figure 9 shows that Model B performed better than the other two models in the financial categorization dataset since it was customized to it, while Model A's results are comparable to Model B's. Additionally, model A's accuracy is higher than both models, as shown in Figure 9, so they chose the model ProsusAI/finbert. This model provides the likelihood of making favorable, negative, or neutral predictions about the provided text. However, the most important problem encountered here was that the news content that the model was working on had a limit of 512 words. When using the model, there were certain decisions to be made. Only 512 tokens (words separated into separate tokens by the tokenizer; typically, each token consists of one word, but some words may contain two or more tokens) could be predicted by the model at once since the complexity increased exponentially with token length. As most news stories began with a summary section and the final paragraphs of the texts discussed how something similar had occurred in the past, we made the decision to truncate any text that was longer than 512 tokens. Following the truncation process, we found that the news's title also affected the sentiment analysis of the news. In order to get the maximum probability value of the sentiment, we do a computation that uses the prediction values of the text and title, giving the text a 60 percent prediction value and the title a 40 percent prediction value. They also arrived at the results in Figure 10 when they tested the model on data we labeled ourselves. But the reason why the scores here dropped was the possibility that we might have mislabeled the data. That's why we didn't focus too much on these scores and our continued use of the model was approved by the supervisor. It was very difficult for us to write and integrate the artificial intelligence model that summarizes in a limited time. This part was one of the missing parts of the project that could be improved.



MODEL	F1 SCORE	PRECISION	RECALL	ACCURACY
A	0.91	0.89	0.93	0.90
B	0.94	0.94	0.94	0.94
C	0.85	0.86	0.83	0.85

- ProsusAI/finbert (A)
- ahmedrachid/FinancialBERT-Sentiment-Analysis (B)
- mrm8488/distilroberta-finetuned-financial-news-sentiment-analysis (C)

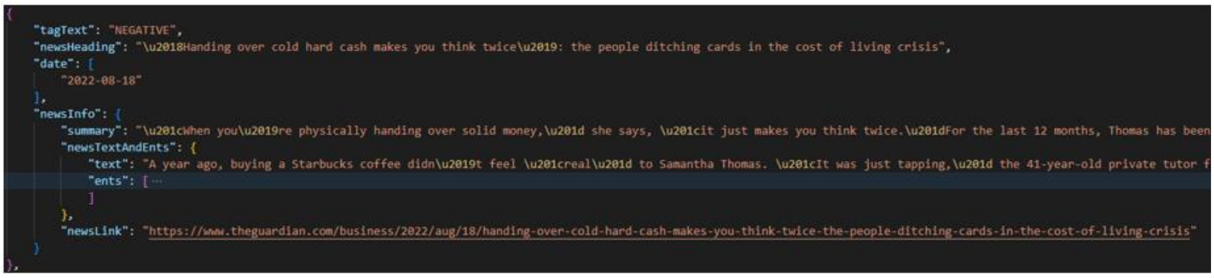
Figure 9 – Score of the models



MODEL	F1 SCORE	PRECISION	RECALL	ACCURACY
A	0.51	0.53	0.50	0.62
B	0.53	0.54	0.60	0.54
C	0.55	0.57	0.57	0.61

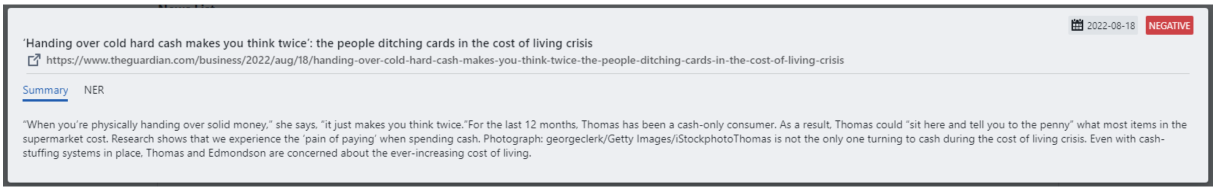
Figure 10 – Finetuned version

NER analysis was developed by the model with an f-score of 0.85, and a common JSON format emerged with the results from the sentiment part. In Figure 10, there is how a news that goes through sentiment and NER processes looks like. The use of NER tags here was highly appreciated by the supervisor. Later, Berk visualized the obtained information with frontend development and revealed the final product. Screenshots of the final product can be found in Figure 12, Figure 13, Figure 14.



```
{
    "tagText": "NEGATIVE",
    "newsHeading": "\u2018Handing over cold hard cash makes you think twice\u2019: the people ditching cards in the cost of living crisis",
    "date": [
        "2022-08-18"
    ],
    "newsInfo": {
        "summary": "\u201cWhen you\u2019re physically handing over solid money,\u201d she says, \u201cit just makes you think twice.\u201dFor the last 12 months, Thomas has been
        "newsTextAndEnts": {
            "text": "A year ago, buying a Starbucks coffee didn\u2019t feel \u201creal\u201d to Samantha Thomas. \u201cIt was just tapping,\u201d the 41-year-old private tutor f
            "ents": [ ···
            ]
        },
        "newsLink": "https://www.theguardian.com/business/2022/aug/18/handing-over-cold-hard-cash-makes-you-think-twice-the-people-ditching-cards-in-the-cost-of-living-crisis"
    }
},
```

Figure 11 – API's unit data



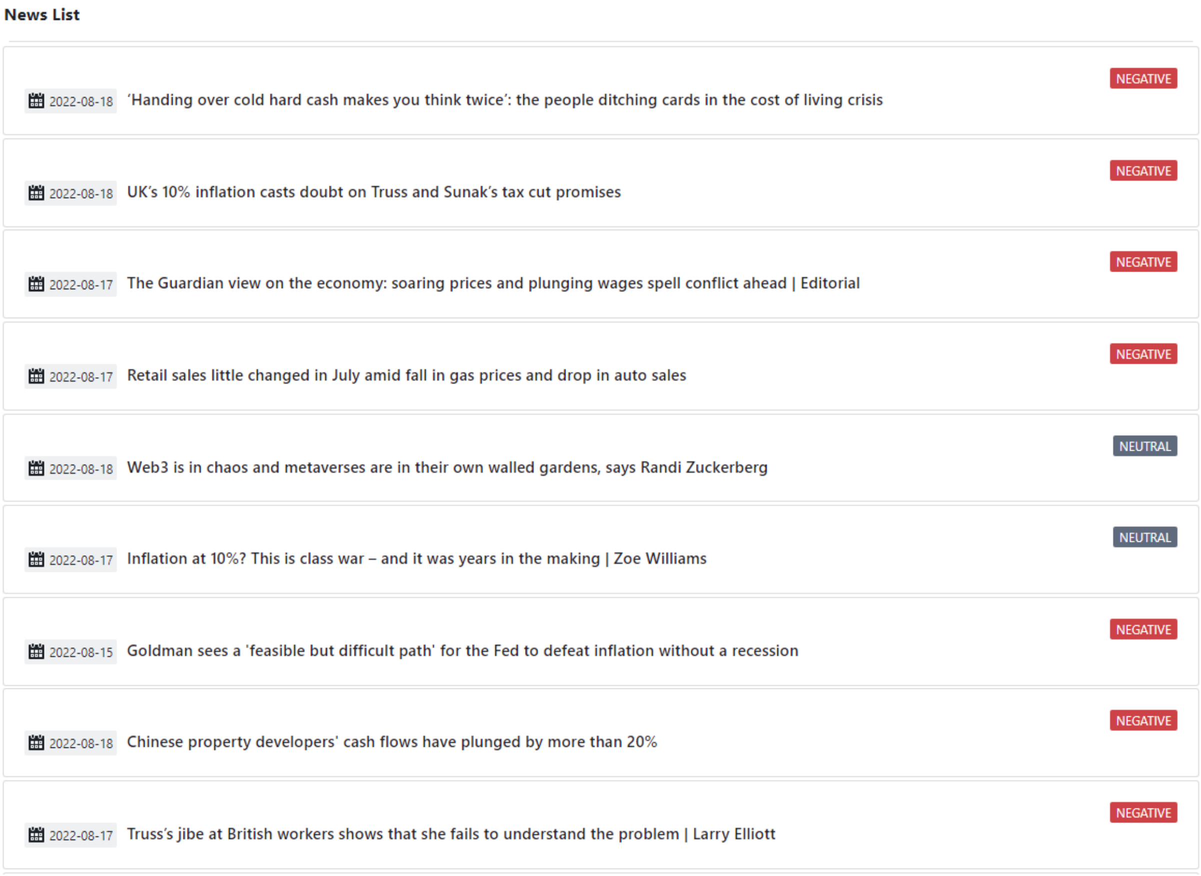
2022-08-18 NEGATIVE

'Handing over cold hard cash makes you think twice': the people ditching cards in the cost of living crisis

https://www.theguardian.com/business/2022/aug/18/handing-over-cold-hard-cash-makes-you-think-twice-the-people-ditching-cards-in-the-cost-of-living-crisis

"When you're physically handing over solid money," she says, "it just makes you think twice."For the last 12 months, Thomas has been a cash-only consumer. As a result, Thomas could "sit here and tell you to the penny" what most items in the supermarket cost. Research shows that we experience the 'pain of paying' when spending cash. Photograph: georgeclerk/Getty Images/iStockphotoThomas is not the only one turning to cash during the cost of living crisis. Even with cash-stuffing systems in place, Thomas and Edmondson are concerned about the ever-increasing cost of living.

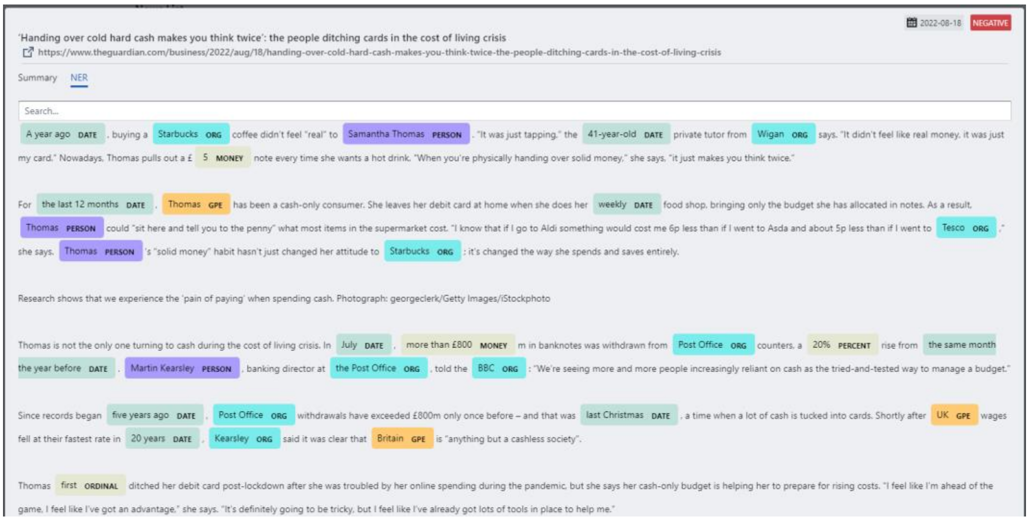
Figure 12 - User interface of project, summary page



News List

2022-08-18	'Handing over cold hard cash makes you think twice': the people ditching cards in the cost of living crisis	NEGATIVE
2022-08-18	UK's 10% inflation casts doubt on Truss and Sunak's tax cut promises	NEGATIVE
2022-08-17	The Guardian view on the economy: soaring prices and plunging wages spell conflict ahead | Editorial	NEGATIVE
2022-08-17	Retail sales little changed in July amid fall in gas prices and drop in auto sales	NEGATIVE
2022-08-18	Web3 is in chaos and metaverses are in their own walled gardens, says Randi Zuckerberg	NEUTRAL
2022-08-17	Inflation at 10%? This is class war – and it was years in the making | Zoe Williams	NEUTRAL
2022-08-15	Goldman sees a 'feasible but difficult path' for the Fed to defeat inflation without a recession	NEGATIVE
2022-08-18	Chinese property developers' cash flows have plunged by more than 20%	NEGATIVE
2022-08-17	Truss's jibe at British workers shows that she fails to understand the problem | Larry Elliott	NEGATIVE

Figure 13 – User interface of project, main page



2022-08-18 NEGATIVE

'Handing over cold hard cash makes you think twice': the people ditching cards in the cost of living crisis

https://www.theguardian.com/business/2022/aug/18/handing-over-cold-hard-cash-makes-you-think-twice-the-people-ditching-cards-in-the-cost-of-living-crisis

A year ago DATE , buying a Starbucks ORG coffee didn't feel "real" to Samantha Thomas PERSON . "It was just tapping," the 41-year-old DATE private tutor from Wigan ORG says. "It didn't feel like real money, it was just my card." Nowadays, Thomas pulls out a £ 5 MONEY note every time she wants a hot drink. "When you're physically handing over solid money," she says, "it just makes you think twice."

For the last 12 months DATE , Thomas GPE has been a cash-only consumer. She leaves her debit card at home when she does her weekly DATE food shop, bringing only the budget she has allocated in notes. As a result, Thomas PERSON could "sit here and tell you to the penny" what most items in the supermarket cost. "I know that if I go to Aldi something would cost me 6p less than if I went to Asda and about 5p less than if I went to Tesco ORG ," she says. Thomas PERSON 's "solid money" habit hasn't just changed her attitude to Starbucks ORG : it's changed the way she spends and saves entirely.

Research shows that we experience the 'pain of paying' when spending cash. Photograph: georgeclerk/Getty Images/iStockphoto

Thomas is not the only one turning to cash during the cost of living crisis. In July DATE , more than £800 MONEY m in banknotes was withdrawn from Post Office ORG counters, a 20% PERCENT rise from the same month the year before DATE . Martin Kearsley PERSON , banking director at the Post Office ORG , told the BBC ORG : "We're seeing more and more people increasingly reliant on cash as the tried-and-tested way to manage a budget."

Since records began five years ago DATE , Post Office ORG withdrawals have exceeded £800m only once before – and that was last Christmas DATE , a time when a lot of cash is tucked into cards. Shortly after UK GPE wages fell at their fastest rate in 20 years DATE , Kearsley ORG said it was clear that Britain GPE is "anything but a cashless society".

Thomas first ORDINAL ditched her debit card post-lockdown after she was troubled by her online spending during the pandemic, but she says her cash-only budget is helping her to prepare for rising costs. "I feel like I'm ahead of the game, I feel like I've got an advantage," she says. "It's definitely going to be tricky, but I feel like I've already got lots of tools in place to help me."

Figure 14 - User interface of project, NER page