ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

QUANTIFYING ANXIETY DURING THE SECOND INDUSTRIAL REVOLUTION (1870-1915)

SEMESTER'S PROJECT

Thomas Benchetrit, supervised by Elena Fernandez Fernandez and Jérôme Baudry

10th June 2022

1 Abstract

Using newspapers issued during the Second industrial revolution, this study aims to quantify the anxiety of the people in four different languages: French, English, German, and Spanish. By performing word ratio for OCR assessment, the quality of each newspaper was assessed and robustly compared to the others. Then, using state-of-the-art Transformers models for date detection, it assesses the performance of these models on noisy text datasets in French and English, and use them to perform date detection on the French newspaper Le Figaro. From it, this study tries to tie together the number of dates expressed in a newspaper to the anxiety of the population at that time.

2 Introduction

2.1 Context

Innovations create rhythm in the life of an economy, much so in an inter-dependant one that has flourished in our era. The first Industrial Revolutions settles itself in Great Britain from 1760 to 1815 with the appearance and the democratization of steel, weaving machines, and the steam engine [1]. Those new technologies enabled the beginning of an economic boom, allowing faster trade routes and agriculture to be much more efficient. It allowed western countries to go from a traditionally agriculture-based economy to an economy that thrives on large-scale mechanically manufactured goods.

The second Industrial Revolution is often referred as the "Technological Revolution", with the harnessing of electrical power and the widespread of new communication technologies such as the telegram. From the late 19th to early 20th centuries[4], all the innovation and new technologies radically changed the way people lived, travelled and had leisure. These technology-induced structural changes have led several researchers to investigate on how those new technologies induced anxiety in the different populations subject to those innovations.

According to other studies[2], the western economies are today in the midst of a third industrial revolution. The changes brought by internet and the computational power that comes with it has drastically refashioned how we communicate, and has made a lot of jobs that were thought essential obsolete.Typists, which were a highly required job in companies a few decades ago are today a relic from the past. Therefore, as all industrial revolutions before it , the question of innovation-induced anxiety can be asked. Our era sees an important amount of deep cultural changes, which are related to those new means of communications, and are also a consequences of the way we profit from our environment, such as climate change[3].

2.2 Motivation

When doing historical research, the information source quality is of paramount importance. Primary sources are documents, papers, objects, recordings or any other source of information that were made by people with a direct connection with it, and that was issued at the time the research is focusing on. One of the documents that can be used as primary sources of information are newspaper issued at the time of the study. Therefore analysing the articles of those newspapers is of tremendous interest to researchers to discover how people lived, interacted or work. Moreover, as newspapers are issued at a fined-grained temporal step(daily newspapers are issued everyday), studies can trace back and analyse shifts in the "spirit of the age" with great precision.

A great amount of effort has been done in the recent years to digitize and make accessible those newspapers to the academic community. However, as paper degrades itself through time, the Optical Character

Recognition (OCR) models that were used yield hard to process noisy dataset that can be challenging to parse. Moreover, the sheer amount of newspapers to analyse and unpack can be discouraging for researchers with little or no aptitudes in Computer Science or Natural Language Processing (NLP).

In the context of our project, we will try to use OCR-digitized newspapers from the second industrial revolution in order to algorithmically detect anxiety during this era. The dataset is composed of a multilingual set of newspapers, dated from 1875 to 1920, namely :

- The New York Herald, a daily American newspaper founded in 1835, in English
- Le Figaro, a daily French newspaper founded in 1826, in French
- El Imparcial, a daily Spanish newspaper founded in 1867, in Spanish
- Berliner Tageblatt, a daily German newspaper founded in 1872, in German

2.3 Goals

In this project, several goals were pursued:

1. First, this study tries to find an uniform method to assess the quality of the OCR that is consistent in each language in order to determine how this could affect the results in the future studies
2. Then, in order to be usable easily by academic researchers without a deep computer science background, we tried to create an easy-to-use and expandable python library that could be use to treat files for the same newspapers used in this study, and even new ones.
3. Moreover, in order to detect anxiety, a Named Entity Recognition (NER) pipeline was run on the processed newspapers to detect dates occurrences, which are thought to correlate with anxiety
4. Also, these pipelines results were manually assessed on our specific noisy dataset in French and English, to determine how well the different pipelines behaved on sub-quality documents

3 Methods

3.1 OCR quality assessment

Optical character recognition is an electronic set of methods to translate written documents and images into machine-readable text files. In our case, the written documents are the newspapers described in Part.2.2. OCR assessment can be done through several methods. One of the best ways would be to comapre the number of optically well read characters compared to a list of manually anotated corpus. However, in our case such corpus do not exist, and the available dataset does not contain the first-hand scanned documents. Another way of assessing OCR quality is to count the word ratio [6]. The word ratio can be defined as the number of tokens belonging to a reference corpora over the number of digitized token recognized by the OCR method used.

> *COOLIDGE 1 HARDING GOING TOGANALZUNE I Will Study Conditions to Huvo A in orient i inj Ko liovod of Tolls Tltll OF THKEK WEEKS With Party Will Make Stay First at Point Isabel Tex r v HaHinjr Happy Hit Not Exultant at Victpry Special Drnpatrh to Tur Nkw Yf i I ftjARION Ohio Nov At Senator Hardinr rat down in his library and penciled the following state ment that the r y are wholly dependable I do not to say that I 0 oleased and of course hmny fritter my gratitude Hut I am not oxultnnt It is not a Demons It is a renewed cxpres ion of*

”

Example of text block OCR reading for the New York Herald
3.11.1920

The crucial detail of this method resides in the choice of the reference corpus used to determine if a token belongs to a correct word representation in the targeted language. As the dataset studied in this project is multilingual, the chosen corpus must also be robust and not differ in term of a valid token definition from one language to another. For example, in some corpora, single letters such as c, d could be recognize as valid tokens whereas in others those tokens would be discarded. The final corpora chosen to perform the word ratio assessment bases itself on the HunSpell [7] open-source spell-checking library. This library provides multi-lingual spell-checking dictionaries that can then be used to validate the OCR-obtained tokens. Then, the `pyenchant` library [8] for python was used as a wrapper around the `Enchant` library [9] in order to use the HunSpell dictionaries.

3.2 Date detection using NER models for anxiety measuring

Anxiety can express itself through different ways. The one studied in this paper is the mention of dates. The hypothesis is that the more anxious an era is, the more likely it is to either mention dates that recalls ancient idealised pasts, or mention future moments that brings the anxiety. Date detection can be thought as a token classification task. In the past, such tasks was handled by rule-based models, such as the `date parser`[15] python library . However, as dates lemmas can take different forms (for a text written on the 1st of January, *tomorrow*, *the 2nd of January*, *02/01*, *the next day* all refer to the same date), such methods are not robust to date format. Moreover, if the date lemma is misspelt then the lemma would not be recognized as a date. Therefore, in this study,state-of-the-art NER Transformer-based model were used to infer dates.

Named Entity Recognition regroups diverse methods to locate and classify named entities in unstructured text. Even though these methods were mainly used to detect persons and organisation, some models can also detect dates. However, as date detection was not the main focus of those models, wide-spread production-ready models are not equally available across languages. Out of the four languages available in our dataset, only French and English languages were covered by sufficiently documented models in wide-spreads model databases such as `spaCy` [10] or `Hugging Face` [11]. Transformers are a category of deep learning models that are mainly use to treat sequential data, such as time series, or sentences for examples. They base themselves on a self-attention mechanism [12] and are widely used mainly in Computer Vision and NLP. The models we used during this part of the study are :

- `en_core_web_trf`[13], a transformer model developed by `Explosion Ai`, the company which created spaCy. It is trained on the OntoNotes 5 dataset [14] and on the WordNet 3 dataset[16]. Its architecture bases itself on the RoBERTa model, developed by Facebook AI[17]. It is used to detect dates for the English language.
- `camembert-ner`[19], a transformer model that bases itself on the CamemBERT model[18] which possesses the same architecture as the RoBERTa model with an additional fine-tuning for the NER task.

3.3 NER quality assessment

The models used for date detection were trained on specific datasets that do not include the ones in this study. This implies that their performance may by significantly different that what could be expected for their reported measurements. Therefore, this part of the study aims to determine the performance of those two models on the selected newspapers OCR readings. To do so, a random sample of text blocks

were extracted from the corpus, manually annotated using `ner-annotator`[20] a webApp for text annotation. Common metrics used [21] for tagging assessment are Precision, Recall, and the F1 score. They can be defined as :

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad \text{F1} = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} \tag{1}$$

where TP is the number of True Positives i.e well recognized date, FP is the number of False Positives, i.e misclassified tokens and FN the number of False Negatives, i.e not recognized dates.

4 Results

4.1 OCR quality

In this part are presented the results of the OCR quality assessment using the word-ratio technique described in Part.3.1. Figs.1,2,3,4 present the word ratio of each text block with respect to its length from a year of data for each newspaper. Fig.5 is an histogram displaying the distribution of the word ratio for each text block and for each newspaper.

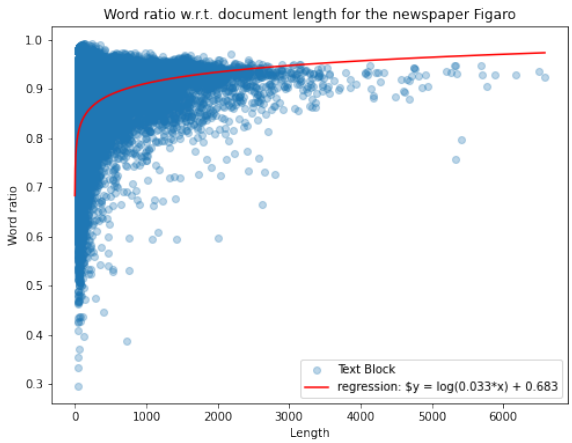

Figure 1
Logarithmic regression on the word ratio computation for the French newspaper Le Figaro, $R^2 = 0.162$

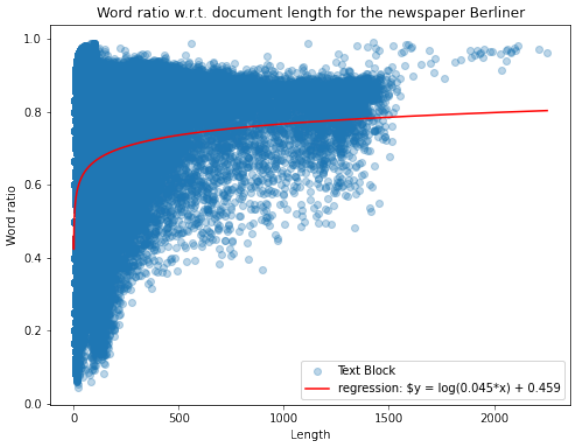

Figure 2
Logarithmic regression on the word ratio computation for the German newspaper Berliner Tageblatt, $R^2 = 0.207$

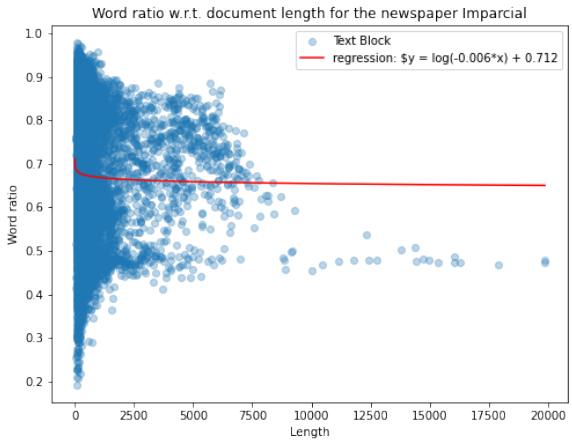

FIGURE 3
Logarithmic regression on the word ratio computation for the Spanish newspaper El Imparcial, $R^2 = 0.002$

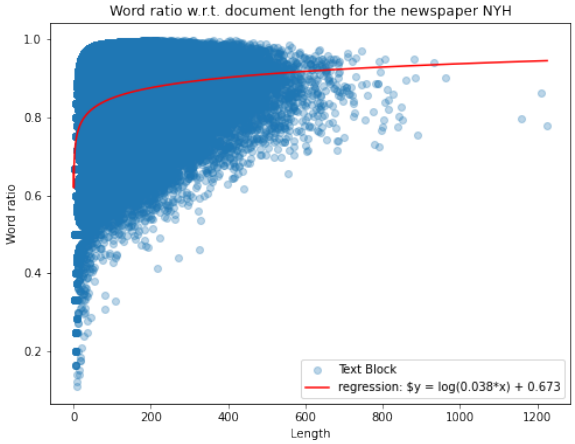

FIGURE 4
Logarithmic regression on the word ratio computation for the American newspaper New York Herald, $R^2 = 0.158$

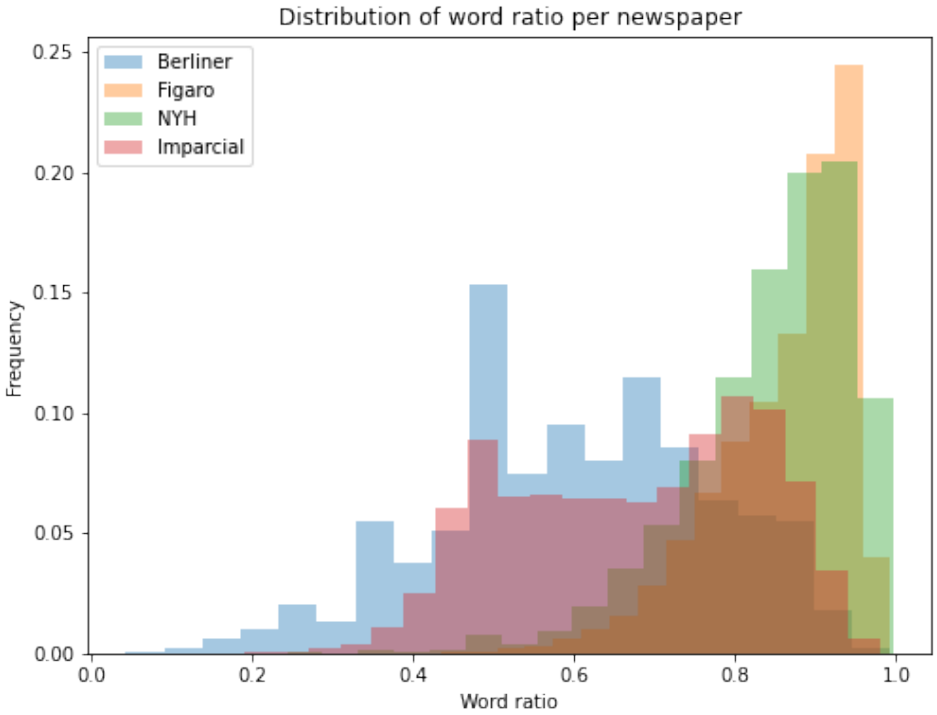

FIGURE 5
Distribution of the text blocks word ratio for the four newspapers

4.2 NER ASSESSMENT

The next two tables Tab.1, 2 compiles the metrics used to assess the two models that were used to do date detection.

	Reported metrics	Assessed metrics
Precision	0.90	0.78
Recall	0.90	0.84
F1 score	0.90	0.80

TABLE 1

Table reporting the metrics for the NER date detection for the English newspaper the New York Herald. It contains both the metrics reported by the model creators ("reported metrics") and the metrics assessed on the newspaper dataset ("assessed metrics")

	Reported metrics	Assessed metrics
Precision	0.90	0.704
Recall	0.90	0.476
F1 score	0.90	0.568

TABLE 2

Table reporting the metrics for the NER date detection for the French newspaper Le Figaro. It contains both the metrics reported by the model creators ("reported metrics") and the metrics assessed on the newspaper dataset ("assessed metrics")

4.3 DATE DETECTION

First, the number of missing files has been computed per month for the French newspaper Le Figaro, in order to assess the incompleteness of the dataset. Then the number of dates per available newspaper issue was computed. In order to normalize the number of dates obtained in each issue, the number obtained was divided by the number of tokens that were present in this newspaper issue, giving the date ratio that is displayed on Fig.7

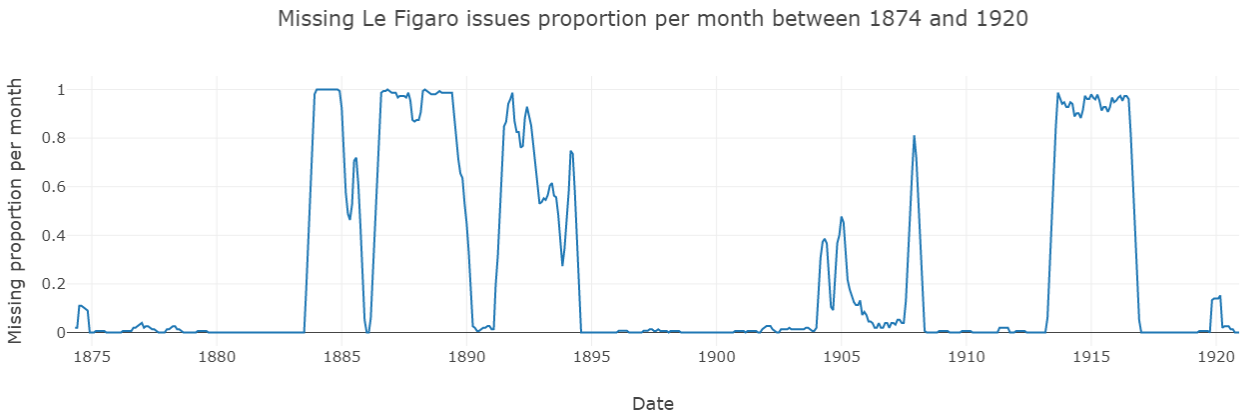

FIGURE 6

Proportion of missing Le Figaro newspaper issue per month

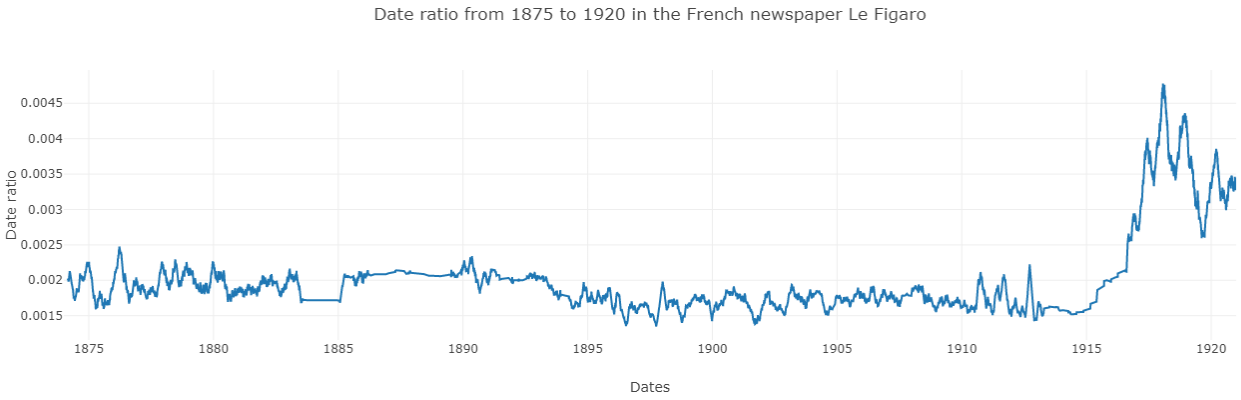

FIGURE 7
Date ratio from 1874 to 1920 in Le Figaro

5 DISCUSSION

5.1 OCR QUALITY

From Fig1 to 4 are presented the word ratio results for the four newspapers presented in this study. The first notable observation is that for three out of four newspaper, namely Le Figaro, the Berliner Tageblatt and the New York Herals, the word ratio grows logarithmically with the block length. A logarithmic regression was therefore performed on each of these newspapers, obtaining a R^2 score ranging from 0.158 to 0.207.


AMERICAN PORTS.
OCEAN STEAMSHIPS.
LOYD SABAUDO
THE FAST ITALIAN MAIL LINE
NEW YORK–NAPLES–GENOA
S.S. "PESARO" . . Jan. 17th
S.S. "REGINA D'ITALIA"..Jan. 22
S.S. "Re D'ITALIA" . . Feb. 4
FURNESS, WITHY & CO., Ltd.
GENERAL AGENTS
34 Whitehall Street, N. Y. C.
GET YOUR
STEAMSHIP
INFORMATION
from General Headquarters
RAYMOND & WITCOMB CO.
OCEAN STEAMSHIPS.

FIGURE 8
Screenshot of a digitized New York Herald issue, dated on January 2, 1920[22]

This can be explained by the fact that bigger blocks usually are articles, which are well formatted, as can be observed on Fig.8, do not contain any images and do not significantly vary in font size, font, layout, etc... Therefore, it is easier for an algorithm to correctly infer characters and text structure from the digital image. However, as the length of the block grows after a certain threshold, the gain is marginal, explaining the choice of a logarithmic regression. For the Spanish newspaper however, even the longer blocks were not well recognized. This can be explain through several factors. First, as English is the most studied language in Natural Language Processing, models are trained to perform well on it,and therefore are less efficient on other languages. Moreover, Spanish is linguistically further away from English than French and German[24] which may increase the effect mentioned earlier.

This is further corroborated by Fig.5 which compiles the distribution of text block per word ratio, per newspaper. One can see that the Spanish and German newspapers yield significantly worse OCR quality than French and English. Partly explained by the limited growth of models for those two languages compared to English and French, another explanation may also resides in the the morphological characteristics of these languages. Indeed, German and Spanish both have longer aver word length [25], which increases the probability of an error on one of the character composing the word.

5.2 NER ASSESSMENT

The NER assessment has been done in two languages : French and English. For both of these languages, the manually assessed metrics computed are lower than the ones reported by the models creators. Moreover, were only selected the text blocks with a minimal length and word ratio to filter out the noisiest blocks. In this study, the word ratio threshold was set at 75% and the minimum block length at 15 tokens. Whereas the precision, recall and f1 score of the date detection method in English yield quite close results from the ones reported (-0.1 for the f1 score) , the French model performs significantly worse on the studied dataset (-0.34), with a sharp difference between precision and recall, with precision being the less affected.

The noisy nature of these dataset can explain a part of this drop. Indeed, models were trained on cleaned dataset and therefore are not well suited for these kind of work. In order to reduce the gap in the precision and recall, one method would have been to first parse the text blocks with an error corrector to first de-noise the dataset, then run the chosen pipeline.

After manually checking which dates lemma were not well recognized by the French model, one key observation was that dates that were at the end of longer text blocks tended to be less recognized, which may indicate that the model used for French do not benefits sufficiently from the self-attention head. Finally, the fact that the french model was less robust to noise can also be attributed to the age of model. Indeed, the English model used for date detection has been profusely studied and fine-tuned on extended datasets, whereas the date detection part of the French model has only been trained on the wikiner dataset.

5.3 DATE DETECTION AND ANXIETY

Date detection has only been done in French using the model described in Part.3.2. First, by analysing the dataset, we observed that several issues of the newspapers were missing, which led to Fig.6. We can clearly see the gap induced by the First World War between 1914 and 1918. The repeated gaps from 1883 to 1894 may however stay unexplained.

The date ratio obtained for the French newspaper Le Figaro are presented on Fig.7. From it, one can see a doubling in the date ratio during and after the First World War. This observation goes in the same direction as our hypothesis. Indeed, war times are stress-inducing[26], and the increased date ratio extracted from the newspaper tend to confirm that.

REFERENCES

[1] The British Industrial Revolution in Global Perspective, by Robert C Allen, 2009

[2] The Rise of the Network Society: The Information Age: Economy, Society, and Culture, by Manuel Castells, 2011

[3] Sixth assessment cycle of the IPCC, https://www.ipcc.ch/report/sixth-assessment-report-cycle/ , consulted on 08.06.2022

[4] The Second Industrial Revolution, 1870-1914, by Joel Mokyr and Robert H. Strotz https://faculty.wcas.northwestern.edu/jmokyr/castronovo.pdf , consulted on 10.06.2022

[5] The History of Technological Anxiety, Joel Mokyr, Chris Vickers, and Nicolas L. Ziebarth https://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.29.3.31

[6] Finding Names in Trove: Named Entity Recognition for Australian Historical Newspapers, by Sunghwan Mac Kim Steve Cassidy

[7] Hunspell library, http://hunspell.github.io/ , consulted on 08.06.2022

[8] PyEnchant library, a wrapper around the enchant library,used with the hunspell dictionaries, https://pyenchant.github.io/pyenchant/ , consulted on 08.06.2022

[9] Enchant library for spell checking, https://abiword.github.io/enchant/, consulted on 08.06.2022

[10] Spacy Library, https://spacy.io/ , consulted on 08.06.2022

[11] Hugging Face model HUb, https://huggingface.co/ , consulted on 08.06.2022

[12] Attention is All You Need, by Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin, https://arxiv.org/pdf/1706.03762.pdf consulted on 08.06.2022

[13] Spacy english model, https://spacy.io/models/enen_core_web_trf consulted on 08.06.2022

[14] OntoNotes 5 (Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, Ann Houston), https://catalog.ldc.upenn.edu/LDC2013T19 , consulted on 08.06.2022

[15] DateParser python library https://dateparser.readthedocs.io/ , consulted on 09.06.2022

[16] WordNet 3 dataset , https://wordnet.princeton.edu/ , consulted on 10.06.2022

[17] RoBERTa model, https://arxiv.org/abs/1907.11692 , consulted on 08.02.2022

[18] CamemBERT mdodel, by Facebook IA, https://camembert-model.fr/ , consulted on 8.06.2022

[19] CamemBERT ner with date detection, https://huggingface.co/Jean-Baptiste/camembert-ner-with-dates , consulted on 09.06.2022

[20] NER annotator web application https://github.com/tecoholic/ner-annotator , consulted on 08.06.2022

[21] Part of speech tagging: a systematic review of deep learning and machine learning approaches, by Alebachew Chiche and Betselot Yitagesu https://journalofbigdata.springeropen.com/articles/10.1186/s40537-022-00561-y , consulted on 08.06.2022

[22] New York Herald screenshot, https://www.loc.gov/resource/sn83030313/1920-01-02/ed-1/?sp=2r=0.376,0.325,0.813,0.459,0 , consulted on 10.06.2022

[23] English is the most studied language in NLP https://towardsdatascience.com/the-importance-of-natural-language-processing-for-non-english-languages-ada463697b9d , consulted on 09.06.2022

[24] Lexical Distance between languages in Europe, by Steinbach, Elms, and Tishchenko , https://alternativetransport.files.wordpress.com/2015/05/lexical-distance-among-the-languages-of-europe-2-1-mid-size.png , consulted on 09.06.2022

[25] Distribution of ortographic word lengths for several languages, https://www.researchgate.net/publication/230724353_CLEARPOND_Cross-Linguistic_Easy-Access_Resource_for_Phonological_and_Orthographic_Neighborhood_Densities

[26] Future Anxiety Among Young People Affected by War and Armed Conflict: Indicators for Social Work Practice, by Nouf M. Alotaibi https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8666412/ , consulted on 09.06.2022