Artificial Intelligence In Medicine 144 (2023) 102638

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Research paper

A holistic AI-based approach for pharmacovigilance optimization from patients behavior on social media☆

Valentin Roche[a], Jean-Philippe Robert[a], Hanan Salam[b,*]

[a] Université Claude Bernard - Lyon 1, Faculté de Pharmacie, Institut des Sciences Pharmaceutiques et Biologiques, 8 Avenue Rockefeller, 69008, Lyon, France
[b] SMART Lab, New York University Abu Dhabi, Saadiyat Island, Abu Dhabi, PO Box 129188, United Arab Emirates

ARTICLE INFO

MSC:
0000
1111

Keywords:
Social network analysis
Drug safety
Pharmacovigilance
AI for healthcare
Natural Language Processing

ABSTRACT

In this paper, we propose a holistic AI-based pharmacovigilance optimization approach using patient's social media data. Instead of focusing on the detection and identification of Adverse Drug Events (ADE) in social media posts in single time points, we propose a holistic approach that looks at the evolution of different user behavior indicators in time. We examine various NLP-based indicators such as word frequency, semantic similarity, Adverse Drug Reactions mentions, and sentiment analysis. We introduce a classification approach to identify normal vs. abnormal time periods based on patient comments. This approach, along with user behavior indicators, can optimize the pharmacovigilance process by flagging the need for immediate attention and further investigation. We specifically focus on the Levothyrox® case in France, which sparked media attention due to changes in the medication formula and affected patient behavior on medical forums. For classification, we propose a deep learning architecture called Word Cloud Convolutional Neural Network (WC-CNN), trained on word clouds from patient comments. We evaluate different temporal resolutions and NLP pre-processing techniques, finding that monthly resolution and the proposed indicators can effectively detect new safety signals, with an accuracy of 75%. We have made the code open source, available via github.

1. Introduction

Pharmacovigilance involves collecting, detecting, assessing, monitoring, and preventing Adverse Events (AEs) associated with pharmaceutical products [1]. Safety signals are information about increasing or decreasing frequency of AEs caused by medication, requiring attention and further investigation [2,3]. These signals can be detected from various sources, often reported by primary care teams. However, patients commonly hesitate to express their feelings to healthcare professionals due to the "white coat effect" [4]". Real-life data, derived from the daily lives of patients, is frequently exchanged within social circles through various channels such as oral communication, messaging, and social networks. Analyzing real-life data is crucial for improving healthcare quality and regulating health systems. It serves multiple purposes [5], including medico-economic evaluations and post-marketing surveillance in pharmacovigilance. For instance, social media platforms (e.g. specialized social networks like Doctissimo®) can be valuable sources for identifying new signals, enabling spontaneous and honest communication between patients and healthcare professionals regarding Adverse Events (AEs) and Adverse Drug Reactions (ADRs) [6].

The rise of Artificial Intelligence (AI) and big data has enabled the collection, analysis, and cross-referencing of large amounts of data, including patients' social media data. Natural Language Processing (NLP) and Machine Learning (ML) techniques have been successfully utilized in medical forums for various applications such as identifying and extracting Adverse Drug Reactions (ADRs) [7–12], retrieving targeted medical information [13], detecting drug non-compliance [14], and conducting sentiment analysis on patients' opinions [15]. Deep learning techniques have also emerged [16], offering improved performance in tasks like mining patients' reviews [17] and detecting ADRs [10–12,18–20] compared to traditional ML approaches.

Previous research on pharmacovigilance using social media data has primarily focused on supervised extraction of Adverse Drug Events (ADE) using labeled datasets. This process involves detecting posts

☆ The work of Hanan Salam was supported by Tamkeen.
* Corresponding author.
E-mail addresses: valentin.roche@outlook.fr (V. Roche), robert.jeanph@gmail.com (J.-P. Robert), hanan.salam@nyu.edu (H. Salam).

https://doi.org/10.1016/j.artmed.2023.102638
Received 22 December 2021; Received in revised form 11 August 2023; Accepted 14 August 2023
Available online 21 August 2023

mentioning ADEs, identifying specific ADE mentions, and standardizing them [12]. However, it is unclear how these ADE extraction pipelines can effectively detect safety signals that require immediate attention and investigation. Isolating ADE identification without considering the temporal aspect is insufficient for signaling potential problems. The proposed approach shifts focus from ADE identification and normalization to detecting abnormal periods in patients' posting behavior on social media, which could indicate potential safety signals. Additionally, annotating data can be cumbersome, especially with large datasets. Therefore, exploring pattern recognition methods for unsupervised analysis of the task can be valuable.

In this paper, we propose a method for early detection of safety signals by analyzing the frequency of adverse events reported by patients on social networks and examining their posting behavior. Our focus is on the Levothyrox® case in France [21–23], which received significant media coverage following changes in the drug's formulation. Many patients experienced adverse effects, and we aim to determine if analyzing data from the Doctissimo® forum can improve pharmacovigilance and facilitate more proactive responses from authorities and laboratories. The Levothyrox® scandal provides distinct time points and periods related to safety signal identification, such as formula changes and media coverage, allowing us to explore unsupervised methods for detecting signals based on patient behaviors rather than relying solely on ADR identification in posts.

The work is in this paper is two-fold. Firstly, we employ NLP techniques to extract relevant indices for potential safety signals during the Levothyrox® scandal. This includes analyzing word frequency, semantic similarity, sentiment, and Adverse Drug Reactions. Secondly, we propose a new deep CNN architecture [24] called Word Cloud CNN (WC-CNN) that utilizes word clouds to detect abnormal periods. We investigate different temporal resolutions and NLP pre-processing techniques to enhance the model's performance. The contributions of our work are as follows: (1) A novel holistic non-supervised approach for pharmacovigilance optimization based on the analysis of patients behavior on social media, (2) a novel Deep Learning approach using word clouds for early detection of safety signals, (3) NLP-based indicators from patient reviews for safety signal detection, and (4) the first analysis of the Levothyrox® scandal utilizing NLP and deep learning tools. Moreover, we have made the extracted data focusing on the Levothyrox scandal used in this work[1] and the source code[2] available for research purposes.

The rest of this paper is organized as follows: Section 2 reviews the literature on AI and NLP approaches in analyzing medical forums. Section 3 describes the materials and methods employed. Section 4 presents the study's findings. Finally, Sections 5 and 6 provide the discussion and conclusion of the paper, respectively.

2. State of the art

Natural Language Processing (NLP) techniques have been successfully exploited for the analysis of medical forums in the medical field for various applications. However, despite the advances in data science and social networks analysis, major challenges remain to overcome to operationalize the analysis of patient posts, and efficiently support the pharmacovigilance process [2,25]. Challenges include (1) inconsistent, unstructured and region-specific data (2) variable quality of information on social media, (3) data privacy, (4) meeting expectations of pharmacovigilance experts, (5) relevant information identification on the internet, and (6) robust and evolutive architecture. In the following, we review relevant literature on data-driven pharmacovigilance research.

[1] Data is accessible via https://zenodo.org/record/7397895#.Y42mPy8RoUE.

[2] Source code is accessible via https://github.com/SMART-Lab-NYU/EarlyDetection_PharmaceuticalScandals.git.

2.1. Patients' reviews sentiment analysis

Various works have focused on the study of patients' sentiment in their reviews on medical forums. Typically the polarity (positive vs. negative) of the patient's comment is detected in an attempt to understand the patients satisfaction about various aspects of their healthcare. For instance, [15] studied how people express their opinions about doctors and drugs in medical forums by exploring lexicon-based and supervised learning based sentiment analysis. Models to detect sentiment polarity (positive/negative) in reviews were trained separately on drugs reviews and doctors reviews, respectively. It was found that drug reviews are more difficult to classify than those about physicians. The use of an informal language was found to characterize reviews about physicians. On the other hand, a combination of informal language with specific terminology (e.g. adverse effects, drug names) with greater lexical diversity was found to characterize reviews about drugs. Similarly, using sentiment analysis, the approach of [26] attempted to categorize the polarity of patients' online comments regarding their hospital health care (recommend a hospital, hospital was clean, good patient treatment).

The polarity of patients expressed sentiment (e.g. negative) on social media can be an indicator of possible issues that might require further attention. This was demonstrated in the state of the art, for instance in the work of [27] which showed that patients posts which mention adverse drug reactions are associated with negative sentiments. Previous work has focused on the study of patients' sentiment with the aim of understanding the patients satisfaction about various aspects of their healthcare. However, it was not investigated in the context of safety signal detection. In this work, we investigate the evolution in time of patients sentiment polarity as a possible indicator for safety signal detection.

2.2. Targeted medical information retrieval

Some works tackled targeted information retrieval using clinical messages or patients reviews. For instance, an approach for relevant clinical messages filtering was proposed in [13]. NLP was used to identify clinical phrases and keywords in the messages posted to an internet mailing list. The selected phrases and keywords were then used as search strings to identify, filter and store clinically relevant messages for further analysis. Pre-processing techniques included stopwords removal, upper to lower case conversion, removal of words less than 3 characters, selection of tokens with length greater than 5 and frequency greater than 7, analysis of the 300 most frequent bi-grams and tri-grams, and sorting the messages by the number of n-gram/keywords they contain followed by computing their occurrence per year. The work of [28] also proposed a semi-automatic update system for new candidate terms identification in live datasets for inclusion in the open access and collaborative consumer health vocabulary. The system consisted of three main parts: a Web crawler and an HTML parser, a candidate term filter that utilizes NLP techniques such as term recognition methods, and a human review interface.

In this work, we perform an in-depth frequency analysis of words and bi-grams in the patients reviews. We investigate if the evolution in time of this frequency can be indicative of a safety signal.

2.3. ADE identification and extraction

Adverse Drug Events identification and extraction from social media data has also gained attention from the pharmacovigilance research community. Example approaches include the study of drug non-compliance or use misbehavior in health online forums using supervised classification [14]. The proposed approach in [14] employed tokenization, POS-tagging, and Lemmatisation as pre-processing techniques, and NaiveBayes, Random Forest and Simple Logistic as classification methods. A manual analysis of the messages content has

revealed that the detected misbehavior in relation to non-compliance constitutes 28% under-use, 27% over-use and 6% misuse.

A vast amount of work has also focused on Adverse Drug Reaction (ADR) identification and extraction from social media data [7,10–12,29,30]. The end goal of detection and identification of ADRs and ADEs is to inform public policy [12]. However, the identification of ADRs in isolation of the temporal aspect is not sufficient to signal any problem. It is the frequency within a specific period of time that might signal a problem that requires attention. Existing approaches in ADR identification and extraction from social media data rely primarily on NLP and Machine Learning (ML) including traditional ML and Deep Learning (DL) approaches.

Existing works in this area include [9] who proposed a standardized protocol for the evaluation of an NLP-based software for the extraction of adverse drug reactions (ADR) from health forums messages. ADR information extraction was performed by extracting the relation between the drug and adverse events entities, and then tested against a gold standard (manually made by two persons experienced in medical terminology). The approach of [8] entailed a weighted online recurrent extreme learning machine for the extraction of ADR mentions. Features were obtained by concatenating character-level (obtained using a modified online recurrent extreme learning machine) and word-level embeddings (obtained from a pre-trained model). An F-score of 87.5% was obtained with this method. Similarly, [31] proposed a method for ADR detection from social media based on SVM and skip-gram lexical patterns. In [29], an ensemble model architecture which combines a domain-specific pre-processing pipeline, Bidirectional Encoder Representations from Transformers (BERT) [32], and Logistic Regression classification was exploited for ADR presence detection. Methods for filtering disorder terms that do not correspond to adverse events in order to optimize the identification of ADR from social media were also exploited in the literature. The approach of [33] exploited a distance-based approach where a distance (as number of words) between the drug term and the disorder or symptom term in the post was computed. An analysis of a corpus of drug-disorder pairs from 5 French forums using Gaussian mixture models and an expectation–maximization (EM) algorithm was performed. The results show that distance between terms can be used for identifying false positives, thereby improving ADR detection in social media.

Recently, deep learning has attracted researchers to propose approaches for ADR detection and extraction from social media [16]. Among the used architectures, we can find Recurrent Neural Network (RNN) [18] approaches. For instance, the RNN-based approach of [18] labels words in an input sequence with ADR membership tags. Word-embedding vectors were used as input features. Bi-LSTM was also proposed by [20] for the detection and identification of professionally unreported drug side effects. The approach made use of Bidirectional Encoder Representations from Transformers (BERT) [32] sentence embeddings which outperformed standard deep learning architectures. Similarly, [12] proposed an ADE resolution pipeline composed of an ADE classifier for filtering tweets with ADE mentions, a Named Entity Recognition (NER) module for ADE mentions extraction, and an ADE normalizer for mapping the entities mentions to their respective MedDRA concepts. For the tweets containing ADE filter, a binary classifier was trained using the transformer model RoBERTa [34] with undersampling and varying loss weights training to tackle class imbalance. The ADE mentions extraction was performed using a bidirectional gated RNN-based architecture. For ADE normalization, supervised, semi-supervised and unsupervised approaches were investigated. The outcomes of this work outlined that optimal performance of pipeline architectures of ADE resolution tasks requires training and fine-tuning the different components of the pipeline architecture based on input data imbalance. In [30], different versions of BERT combined with an ensemble of neural networks was investigated for ADR extraction and normalization. For the normalization task, three approaches were proposed including a classification approach, a metric-learning neural model, and an approach combining both. In [35] mining ADR mentions was done using sequence labeling with word embedding cluster features. The authors introduce ADRMine, a machine learning-based concept extraction system that uses conditional random fields (CRFs) and a different features, including a novel feature for modeling words' semantic similarities. The words' similarities are modeled by clustering words based on unsupervised, pre-trained word embeddings generated from unlabeled user posts in social media using a deep learning technique. Although the task of ADR extraction and identification is directly correlated with safety signals detection. However, the detection of an ADR mention in a particular social media post is not sufficient to flag a safety signal that requires further investigation. It is the frequency or dis-proportionality of reporting of a certain ADR and its variation in time that actually constitute a safety signal. Till today, the focus of existing approaches was on atomistic approaches such ADR extraction and normalization, with no further analysis of the frequency or trends over time. Consequently, the literature can benefit from a holistic approach to the problem. Moreover, other than the ADR mentions, other indicators can be useful to better inform safety signal detection. For instance, the expressed sentiment in the posts, the words frequency, the semantic similarity, etc.

While the end goal of ADE detection and identification is to detect safety signals that require immediate attention and further investigation, it is unclear how the current ADE extraction pipelines can be used for this purpose. For instance, the identification of ADEs in isolation of the temporal aspect is not sufficient to signal any problem. It is the altered patients behavior on social media, and the increased frequency of typical behaviors (e.g. increased mentions of ADE) within a specific period of time that might signal a problem that requires attention. Consequently, the proposed approach does not focus on the task of ADE identification and normalization in text but rather on the detection of abnormal periods of patients posting behavior on social media, which might indicate a potential safety signal. On the other hand, data annotation can be a cumbersome task, especially when the size of the dataset is large. Consequently, investigating pattern recognition methods that allow an unsupervised analysis of the task can be useful. Moreover, non of the existing approaches explored the use of word cloud images as inputs to deep CNN architecture.

2.4. Thyroid hormone replacement therapy

Very few works have tackled the data analysis concerning thyroid hormone replacement therapy. The work of [36] used NLP to identify the themes of patient medication concerns regarding thyroid hormone replacement therapy in a dataset collected from WebMD in the United States. They used multiple regression analyses to examine the predictability of the primary medication concerns on patient treatment satisfaction. Their study has found six distinctive themes of patient medication concerns related to Levothyroxine treatment and that treatment satisfaction on levothyroxine was highly dependent on the primary medication concerns of the patient. As Pre-processing techniques, the approach applied stopwords removal, tokenization, stemming, and words frequency. Latent Dirichlet allocation was used to detect the topics of concerns. A very recent approach is that of [37] who proposed a dynamic co-clustering method, named the dynamic latent block model (dLBM), as a tool for automatic safety signal detection. The dLBM operates on ADR count data matrices evolving over a time period and recognizes clusters in a meaningful way that identified safety events.

We focus on a famous drug use case: the Levothyrox® case. The new formula of Levothyrox® was marketed in France in March 2017 by the Merck® laboratory. An increase in the frequency of Adverse Drug Reactions due to taking the drug was identified and reported in the media from July 2017. In response to this incident, the Merck® laboratory had to withdraw the new formula from the market and reintroduce in France the old formula of Levothyrox® in October 2017.

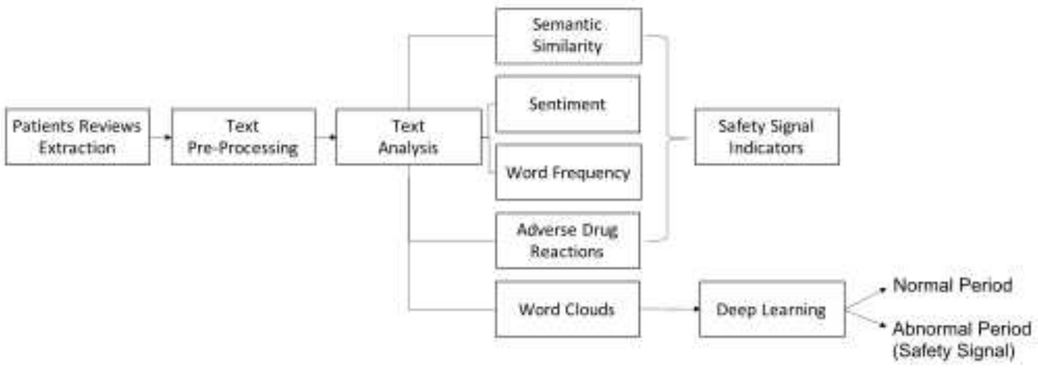

Fig. 1. Workflow of the proposed approach.

The Levothyrox® case presents an important real world use case in pharmacovigilance studies allowing to investigate AI and data science methods for the development of data-driven pharmacovigilance approaches. Compared to the above reviewed approaches, our work falls in the category of safety signal detection in pharmacovigilance. We analyze the patients comments to extract possible indicators of a safety signal. We propose a CNN deep architecture that takes as input word cloud images extracted from patients' reviews. We explore various NLP pre-processing techniques and their effect on the performance of the deep model. The proposed approach is holistic in the sense that we look at the global behavior of the patients' posts and its evolution in time.

3. Materials and methods

In this section, we present the workflow of the proposed work (cf. Fig. 1). The proposed approach is composed of five steps. First in a data extraction step, a data corpus of patients reviews from an endocrinology forum is collected. NLP text pre-processing techniques are then applied to standardize the data and make it suitable for further analysis. Following, various NLP-based data analysis techniques are explored for detecting potential safety signals. Finally, we propose a non-supervised deep learning approach named the Word Cloud CNN (WC-CNN) deep architecture for classifying a period as normal or abnormal. An abnormal period indicates the occurrence of a safety signal which requires further investigation.

3.1. Data extraction

The data corpus was collected from the French health forum Doctissimo using a web scraping algorithm.[3] Doctissimo® was chosen since it is the most used health forum in France by drug consuming patients (ranking first with 61% of users). Other sources of information could have been chosen such as Twitter which is the most used website in the world by drug consuming patients. Twitter brings together 52% of these patients against 27% for all discussion forums combined. However, access to Twitter data is chargeable, which is why Doctissimo® has been selected.

The extraction was performed on the "Thyroïde et Problèmes Endocriniens" (Thyroid and Endocrine Problems) sub-forum with the keyword "levothyrox". The choice of extracting information from this forum using the particular relevant keyword "levothyrox" was to limit the amount of extracted data. Indeed, during data extraction, Doctissimo® blocks the scraping task when reaching a limit of 8,000 extracted discussion threads, since it detects an automatic machine activity.

We collected the messages written between years 2000 and 2020. This resulted in a total of 110,260 comments written by a total of 7650 subjects. For each of the comments, we extract the date, pseudo of the person who wrote the comment, the comment's text, and URL link.

[3] http://forum.doctissimo.fr

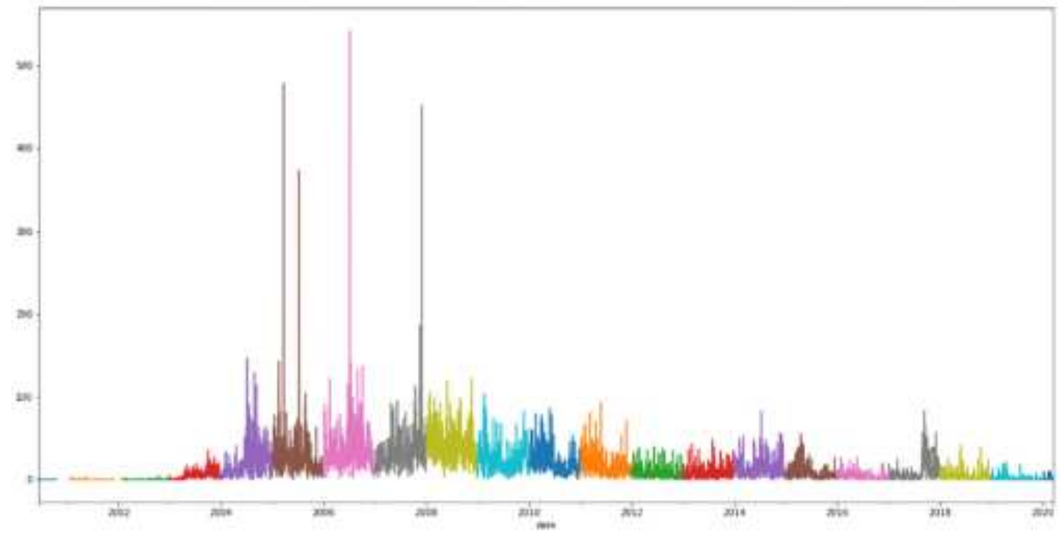

Fig. 2. Number of comments posted on the Doctissimo® forum from year 2000 to 2020.

Fig. 2 shows a plot of the number of comments posted on the forum between years 2000 and 2020. The figure shows that the forum was not used a lot by patients between the years 2000 and 2003. An increase of comments on the forum is event starting from year 2003. Between the years 2004 and 2012, we can see that the forum is more frequented between the years 2004 and 2012 than after 2012. It is difficult to explain the causes of this decrease in attendance except for the hypothesis that some of Doctissimo® users have migrated to other communication channels such as general social networks. Between 2012 and 2016, more comments were observed than between 2016 and 2020 but less important than before 2012. The Levothyrox® affair having taken place in 2017, it was decided to restrict the working database to the period of 2016–2020 during which the occurrence of comments is stable.

3.2. Text pre-processing

Following the data extraction phase, a set of pre-processing techniques were performed on the extracted comments in order to standardize the data for further analysis and clean the dataset from noisy information that might affect the accuracy of the NLP-based analysis negatively. These include:

1. ***Text cleaning***: removal of apostrophes, accents, images, emoticons, particular tags (br, span, table, strong, div, etc.), special characters (i.e. any character other than letters from A to Z and numbers from 0 to 9), digits or numbers contained in words, isolated characters, tabulations and line breaks, digits and numbers except dates, multiple spaces, unwanted patterns (those that have been identified in the database are "# 034" and "# 039"), inserted links (identified using the "http" or "https" pattern), and animated images (identifiable by their link ending with the pattern ".gif."). Additionally rows where at least one cell is empty were also removed from the dataset. These comments can be identified by the pattern NaN (Not a Number) or NaT (Not a Timstamp);
2. ***Uppercase to lowercase conversion***: All uppercase letters are converted to lowercase;
3. ***Spelling correction***: correction of the spelling of the most frequent words in the database. First word clouds by month, week and day visualization was performed. This allowed to visually identify the different used spellings in the database. A word cloud allows to visually represent the most frequent words in the corpus. It represents them in different sizes according to their frequency of occurrence. Thus, the bigger the word, the more it is used in the corresponding part of the dataset (month, week, day). Then, a dictionary was manually created with the most frequent words and their synonyms. For example the synonyms of "levothyrox" were identified to be "levo", "levothyro", and "levotyrox". All occurrences of these words were replaced by the common synonym "levothyrox".

4. ***Irrelevant words exclusion***: a list of irrelevant words were identified and removed from the dataset. These words were qualified as undesirable/irrelevant because they are very recurrent in the comments and are neutral. They do not add any added value in terms of signal detection. These terms have been identified manually in the text. Examples of such words include "actuellement (actually)", "bonsoir (good evening)", "bisous (kisses)".
5. ***Stopwords removal***: stopwords are words so frequent that they bring a lot of noise to the analysis. A word is qualified as empty when it is not discriminating and does not distinguish the comments from one another. A list of stopwords is created by crossing several lists freely accessible on the internet and by analyzing the database and the word clouds. The most frequent stopwords in French are "le", "la", "les" (the), "de", "du" (of), "ce" (this), etc.
6. ***Lemmatization***: lemmatization consists of reducing words to their common lemma to decrease spelling variations between words that have a similar meaning. As an example, the lemmatization transforms the words "petit", "petite", "petits", "petites" into their common lemma which is "petit" (small). We use Spacy library[4] to perform the lemmatization of the words in the corpus.
7. **Short comments removal**: since a sentence has at least three words (a subject, a verb and a complement), comments that are not sentences are deleted to avoid spurious comments.
8. ***Duplicate comments removal***: sometimes users post the same message more than once. Consequently, duplicate comments were removed from the dataset.

3.3. Data analysis approach

Various NLP analysis techniques were performed on the dataset. The goal is to investigate via various means if there is any detectable difference between the normal and the abnormal period. The abnormal period is defined to be the period where the Levothyrox® scandal happened in France. The following analyses were performed: (1) words and n-grams frequency analysis, (2) semantic similarity analysis, (3) sentiment analysis. In case a difference was detected, these analyses methods can be used as indicators of a potential safety signal and can be used as part of a pharmacovigilance surveillance system.

Words and N-grams frequency analysis. The first explored avenue for a safety signal indicator is a frequency analysis of words in the patients' comments. The frequency analysis is carried out on single words, as well as on words sequences referred to in NLP research as n-grams [38]. N-grams are contiguous sequences of N elements in a sentence. N can be any positive integer. Often, N does not exceed 3 because it is rare to frequently see more than 3 adjacent words in different sentences. In this work, bi-grams ($N = 2$) are used to know the most frequent word associations according to the different periods of the study. The purpose is to understand what are the most frequent words or words sequences occurring in the corpus, and whether there is a difference of occurring words during the different periods of the analysis.

A correlation analysis is also performed between the yearly occurrence of the most frequent words, as well as between the most frequent bi-grams . The purpose of this analysis is to investigate whether the words or bi-grams significant to the studied use case are more correlated with each other than the others.

Semantic similarity analysis. Semantic similarity measure is a metric that allows to determine the similarity between various terms such as words, sentences, documents, concepts or instances. It allows to find the degree of relevance between items that are conceptually similar but not necessarily lexicographically similar [39,40].

In order to compute the semantic similarity of two words, first the words should be converted into numerical vectors, a process which is referred to in NLP research as word embedding. The semantic similarity is then computed using a distance measure. The most commonly used distance measure is the cosine similarity.

We employ the Fasttext algorithm [41] for learning word embeddings and computing the semantic similarity. FastText supports both Continuous Bag of Words and Skip-Gram models which are the most commonly used model architectures for learning word embeddings. We implement the skip-gram model to learn vector representation of relevant words from our corpus. The following parameters were used. The size of the embedding vector is set to 60. The window size is set to 20. The minimum word number is set to 3. The down-sampling ratio is set to $1e^{-}2$. The number of iterations of 2000 is used.

Our goal from this analysis is to investigate whether learning the embeddings from the data corresponding to each year, would result in a higher semantic similarity between the words relevant to the studied use case. For this, we train 5 word representation models, trained on the data samples from the periods (1) years 2016 to 2020, (2) year 2016, (3) year 2017, (4) year 2018, (5) year 2019, and (6) year 2020. We then compute the semantic similarity between the identified relevant word to the Levothyrox® scandal.

Sentiment analysis. Also using the FastText library, we perform sentiment analysis of the patient's comments. We train a classifier to detect the polarity of a comment (positive/negative). Our aim is to study the evolution of the patients' sentiment in function of time and examine if there is an evident increase of the negative sentiments expressed in the patients' comments during the period of the scandal.

The training of the sentiment detection algorithm is carried out on a database of French tweets dataset from Kaggle already labeled positive or negative.[5] Since French social media datasets are scarce, this dataset was generated by translating an existing English tweets dataset to French. The polarity of each forum post is then predicted by running the trained algorithm.

For the sentiment detection model to be as efficient as possible, it must be trained on a database in the same language and on a subject most similar to the database to be labeled. Ideally, the tweets would be in French and touch on medical topics. However, no database meeting these two criteria was found. It was therefore decided to focus on the general French tweets database.

Adverse drug reactions analysis. We study the evolution of Adverse Drug Reactions occurrence in the patients comments. For this we develop a method for the detection of Adverse Drug Reactions. The method is based on regular expressions. Regular expressions (regex or regexp) [42] are sequences of characters specifying a search pattern. First we define a list of possible Adverse Drug Reactions related to hypothyriodism. Then using a string-searching algorithm, we detect and count the occurrence of the defined Adverse Drug Reactions in the patients comments.

3.4. Deep learning approach

The safety signal prediction problematic is formulated as a classification problem with two classes: normal and abnormal. A normal period is defined to be a period where there was no event triggering an increase in the frequency of adverse drug reactions normally reported by patients. In other words, it refers to a period where the behavior of patients is considered as normal. The abnormal period on the other hand, is defined to be a period where an important event has happened and triggered an alteration of the reporting behavior of patients on the medical forum. Such behavior is an indication that a safety signal

[4] https://spacy.io

[5] https://www.kaggle.com/hbaflast/french-twitter-sentiment-analysis

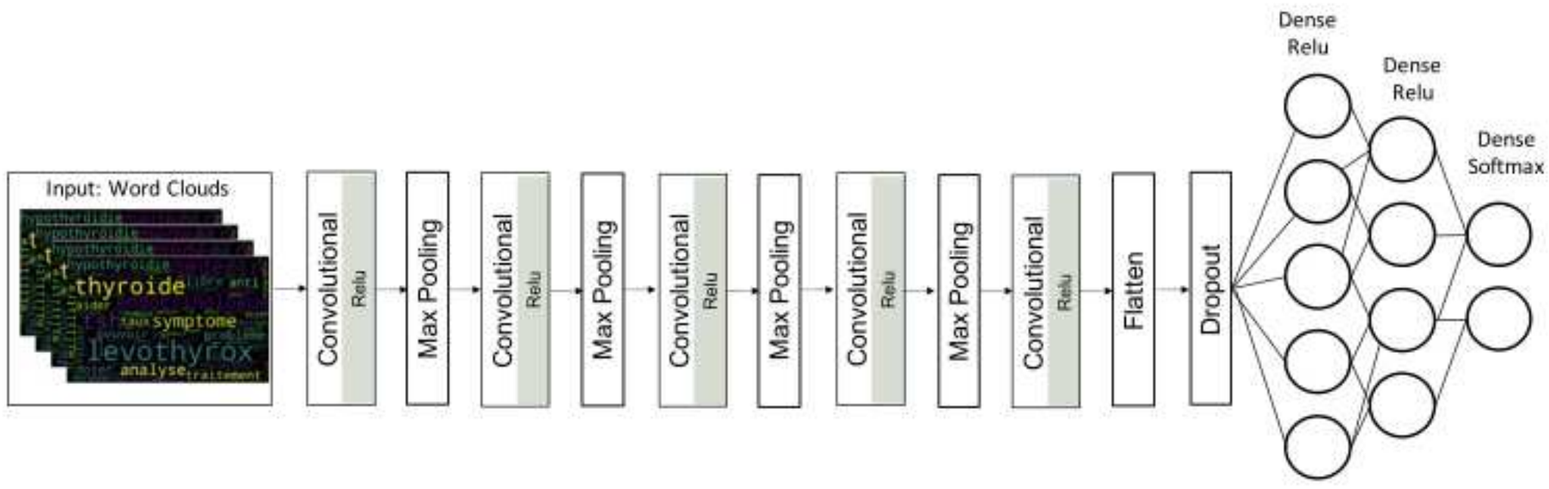

Fig. 3. Proposed WC-CNN architecture. The architecture takes as input word clouds extracted from patients comments at specified temporal resolutions.

should be flagged, to incite attention and consequently a verification or other type of action.

The proposed deep neural network, referred to as Word Cloud CNN (WC-CNN) is an architecture allowing to learn high level features from word clouds extracted from the patients' reviews at different temporal resolutions. Word clouds being visualizations of the most frequent words in a given text with their size reflecting their importance and frequency, represent an avenue to explore Convolutional Neural Networks [24] for classification. The output of the WC-CNN is a binary label (0 for abnormal, and 1 for normal period). Fig. 3 presents an overview of the proposed architecture and the following details the word clouds extraction and the proposed architecture.

3.4.1. Word clouds extraction

A word cloud is one of the most popular data visualization techniques used to represent textual data. It allows to visually represent the most frequent words where the size of a word indicates its frequency or importance in the text. We generate word clouds of size 800 × 500 with a maximum number of words of 200. The maximum font size for the largest word is set to 110. The minimum font size is set to 4. A Relative Scaling (RS) of 0.5 is used. The Font Size (FS) is computed in function of the frequency as follows:

$$FS = (RS * (\frac{frequency}{lastfreq}) + (1 - RS)) * FS \quad (1)$$

3.4.2. WC-CNN architecture

The word cloud features extracted from the patients' comments are fed to a Convolutional Neural Network [24] (named: Word Cloud CNN) to extract high-level deep features. The deep architecture is composed of four two-dimensional (2D) convolutional layers each followed by a ReLU activation function and a max pooling layer, a fifth convolutional layer, a flatten layer, and a dropout layer.

Network implementation details. The number of convolution filters for all convolutional layers is 128. The first 3 convolutional layers have filter sizes of 6 × 6. the fourth and fifth convolutional layers have filter sizes of 3 × 3 and 2 × 2 respectively. A pool size of 2 × 2 was used for all pooling layers. RELU is used as activation function for all convolutional and fully connected layers. The output layer is a dense layer of size 2 with a Softmax activation function. The proposed models are trained with the categorical cross entropy as loss function and the Adam optimizer. The batch size is set to 50 samples. The number of epochs for training is set to 100. An early stopping is performed when the loss function stops improving after 10 epochs.

4. Results

In this section, we present the results of the performed data analysis presented in the above section. Moreover, the results of the proposed deep learning model are also presented and discussed.

4.1. Data analysis results

In this section we present the results of the proposed data analysis approach for investigating relevant indicators of potential safety signal. Section 4.1.1 presents the results of words and n-grams frequency analysis, including the evaluation of the relevance of the n-grams to the studied problematic, and the study of the correlation between the terms of the corpus. Sections 4.1.2, 4.1.3, and 4.1.4 present the results of the semantic similarity, sentiment evolution, and ADR analysis respectively.

4.1.1. Words and N-grams frequency analysis

We present the results of the performed frequency analysis over the dataset. The frequency analyzes carried out are as follows:

1. Word frequency analysis over the entire studied period (2016 to 2020).
2. Analysis of bi-grams over the entire studied period (2016 to 2020).
3. Analysis of the correlation between the occurrence of the most frequent words.
4. Analysis of the correlation between the occurrence of the most frequent bi-grams.

Words frequency analysis. Fig. 4 shows the top 10 frequently occurring words in function of their frequency during this period. Among these words, we can find the words "levothyrox", "tsh", "medecin (doctor)", "traitement (treatment)", "thyroide", "resultat (result)", and "endocrinologue (endocrinologist)". This allows to validate that the topics appearing in the patients' comments the most are related to the problems concerning the medication, and treatment.

Fig. 5 shows the most frequently observed words over five years (2016–2019) and their frequency of occurrence in each year. From the figure, it is observed in 2017 (year of Levothyrox® scandal) that the word "dosage" or "doser (dose)" occurs 1.74 times more than in 2016 and 3.49 times more than the years after 2017. Similarly for the words "fatiguer (tired)" (1.63 vs. 2.63), "formule (formula)" (274 vs. 14.22), "levothyrox" (3.23 vs. 5.72), "mal (pain)" (2.86 vs. 4.00), "medecin (doctor)" (2.20 vs. 2.45), "symptome (symptom)" (1.70 vs. 2.12), and "traitement (treatment)" (1.53 vs. 2.11). The observation of these results clearly indicates strong user activity in connection with Levothyrox® in 2017. Many of the identified words have an over-representation of their occurrences during 2017; starting with the example of a "formula" which was only used twice in 2016 and 548 times the year after. With these results, it was indeed possible to deepen this work by using the undesirable effects and their related significant terms (e.g. "dosage", "doser (dose)", "fatiguer (tired)", "mal (pain)", "symptome (symptom)", "traitement (treatment)").

Table 1
The 10 most frequent bi-grams in years 2016 to 2020.

2016	2017	2018	2019	2020
prise sang	nouveau formule	prise sang	prise sang	hormone thyroidienne
hormone thyroidienne	prise sang	norme labo	norme labo	test synacthene
anti tpo	effet secondaire	hormone thyroidienne	auto immun	tsh bas
pris sang	ancien formule	nouveau formule	maladie auto	anti tpo
norme labo	norme labo	anti tpo	tsh bas	pg milliliter
doser levothyrox	mal tete	ancien formule	anti tpo	tsh mui
auto immun	ancien levothyro	thyroxin henning	resultat tsh	ac anti
prise poids	formule levothyrox	effet secondaire	arret traitement	tsh haute
anti thyroperoxydase	hormone thyroidienne	auto immun	probleme thyroide	unite fourchette
bas norme	doser levothyrox	pris sang	tsh norm	arret cytomel

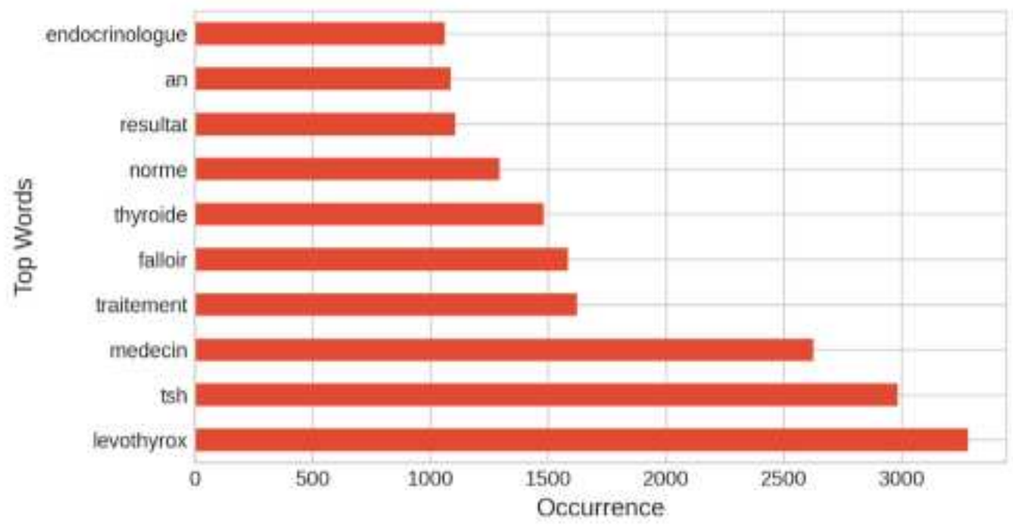

Fig. 4. Top word occurrence (period 2016–2020).

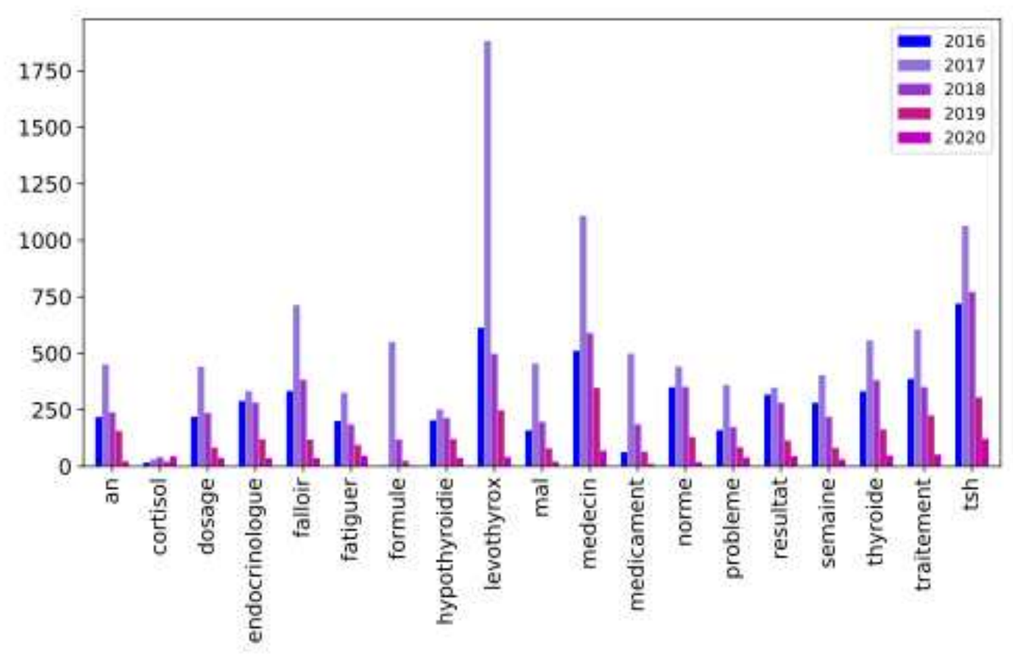

Fig. 5. Top word occurrence per year (period 2016–2020).

Bi-grams frequency analysis. Fig. 6 shows the 10 most frequent bi-grams over the whole studied period. Among these we can notice that "nouveau formule (new formula)", "effet secondaire (secondary effect)", and "ancien formule (old formula)" are among the bi-grams appearing during this period. Fig. 7 shows the most frequently observed bi-grams over five years (2016–2019) and their frequency of occurrence. The analysis made through the bi-grams is more relevant than the over the individual words occurrence presented in the previous section. The terms displayed are consistent and more in line with what is expected. Certain bi-grams are particularly relevant to the studied scandal such as "ancien formule (old formula)", "effet secondaire (side effect)", "formule levothyrox (levothyrox formula)", and "nouveau formule (new formula)". The figure clearly shows that in the year of the scandal (2017) the number of occurrence of these bi-grams is much higher than their occurrence during the years. Table 1 shows the 10 most frequent bi-grams in years 2016 to 2020. We can notice that even more convincingly, bi-grams that did not exist in 2016 become some of the most observed bi-grams of the year of the formula change. These include bi-grams that are directly related to the studied scandal, such as "nouveau formule (new formula)", "ancien formule (old formula)",

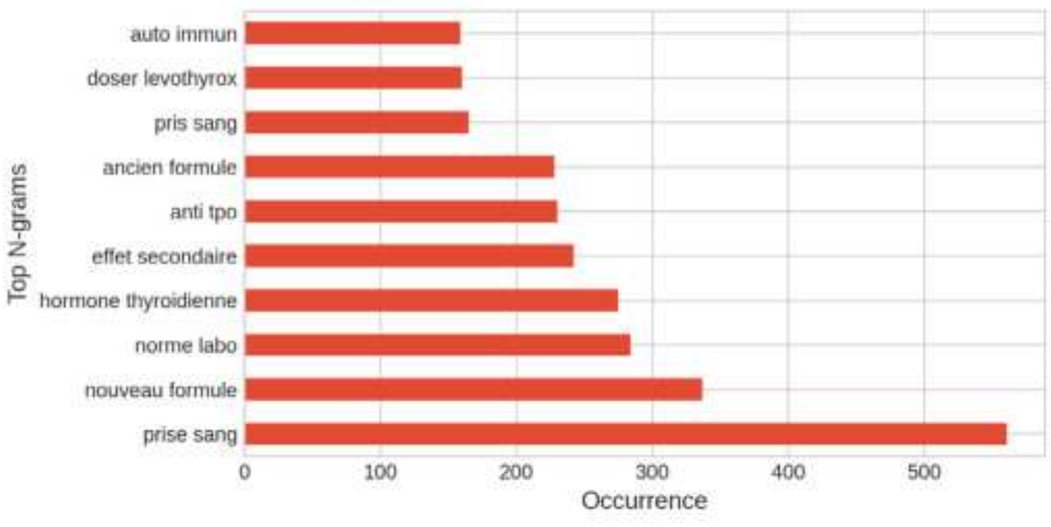

Fig. 6. Top n-gram occurrence (period 2016–2020).

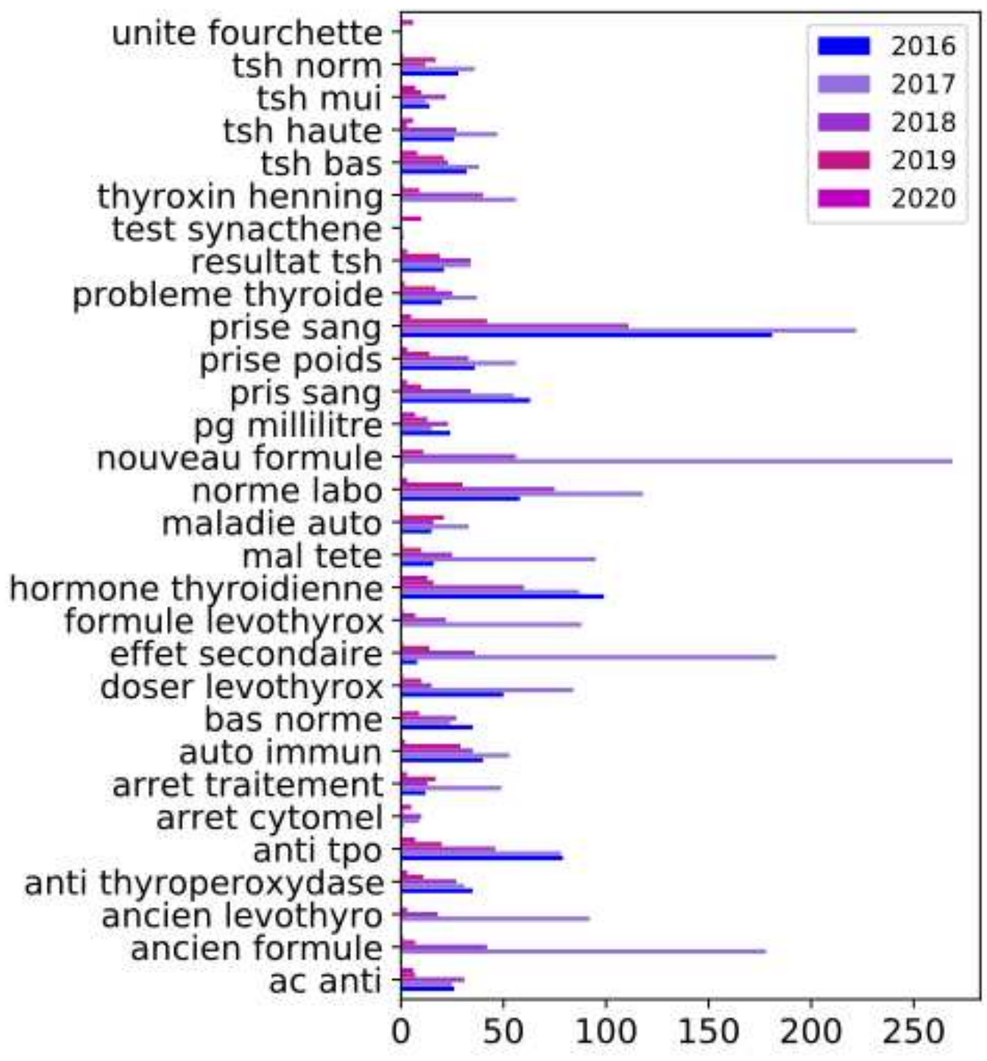

Fig. 7. Top n-gram per year (period 2016–2020).

"ancien levothyrox (old levothyrox)" and "levothyrox formula (formule levothyrox)".

Correlation analysis. Table 2 displays the highest correlations between the 2 bigrams among the top identified bi-grams during the years 2016–2020.

The 10 strongest correlations between two bi-grams are between the identified most occurring bi-grams (cf. Fig. 7) and concern the change in formula of Levothyrox® .

This is undeniable proof of what could be speculated: something new happened in 2017, to the point of dominating the trends over a four-year period. In addition, looking at a monthly breakdown, it can be

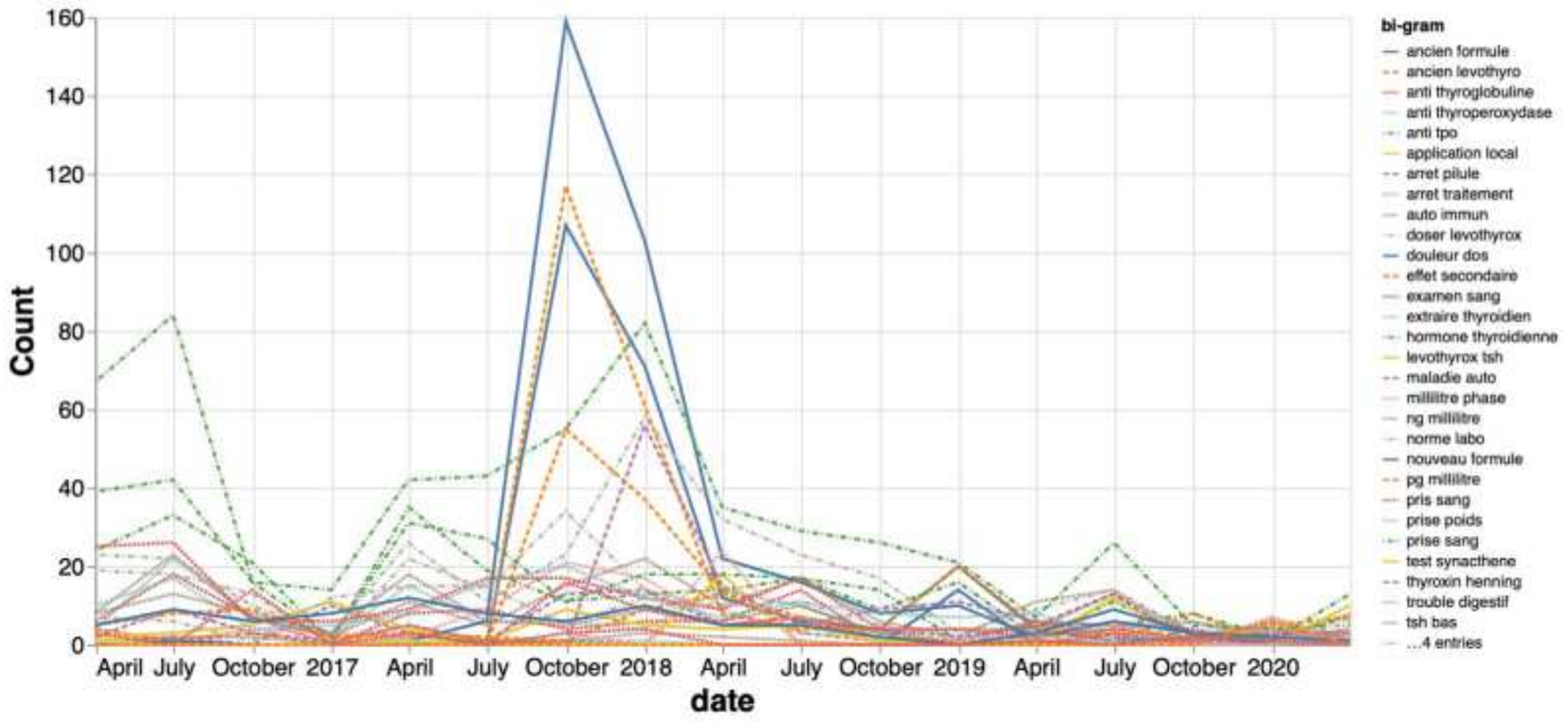

Fig. 8. Histogram reflecting the frequency of appearance of "top n-gram" for the period 2016–2020.

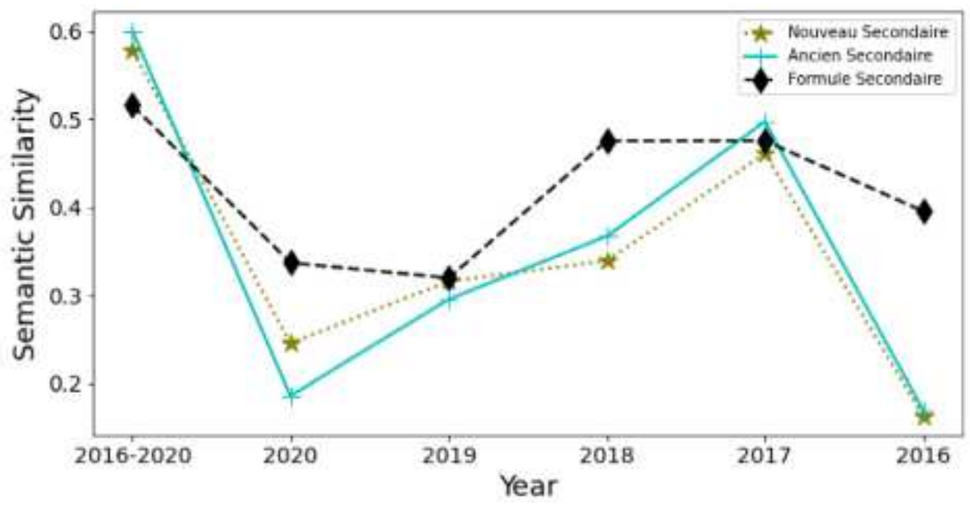

Fig. 9. Semantic similarity analysis (period 2016–2020).

seen in Fig. 8 that this phenomenon started with media announcements, in July 2017.

The study of this histogram shows that using a correlation study of bi-gram occurrences, we can perform an early detection of possible pharmaceutical safety signals.

4.1.2. Semantic similarity analysis

Fig. 9 shows the evolution of the semantic similarity between relevant words to the Levothyrox® scandal. We compute and plot the semantic similarity between the words "Nouveau (new)" and "Secondaire (secondary)", "Ancien (old)" and "Secondaire", "Formule (formula)" and "Secondaire". The word representations were learned on the data corresponding to all years (2016–2020), and only the data corresponding to each year. The figure shows that year of the scandal (2017) shows the highest semantic similarity score with the words of interest.

4.1.3. Sentiment evolution analysis

We analyze the evolution of the perceived sentiments in the user's comments. Fig. 10 shows the evolution of negative and positive sentiments in the patients' comments per year from 2016 to 2020 (top), per month from 2016 to 2020 (middle), and per month during the year 2017. The figure clearly shows an increase of negative sentiments in the patients' comments during the year of the scandal (2017). Zooming on the number of sentiments taking into account a monthly resolution confirms that patients comments were negative during the period of Levothyrox formula change.

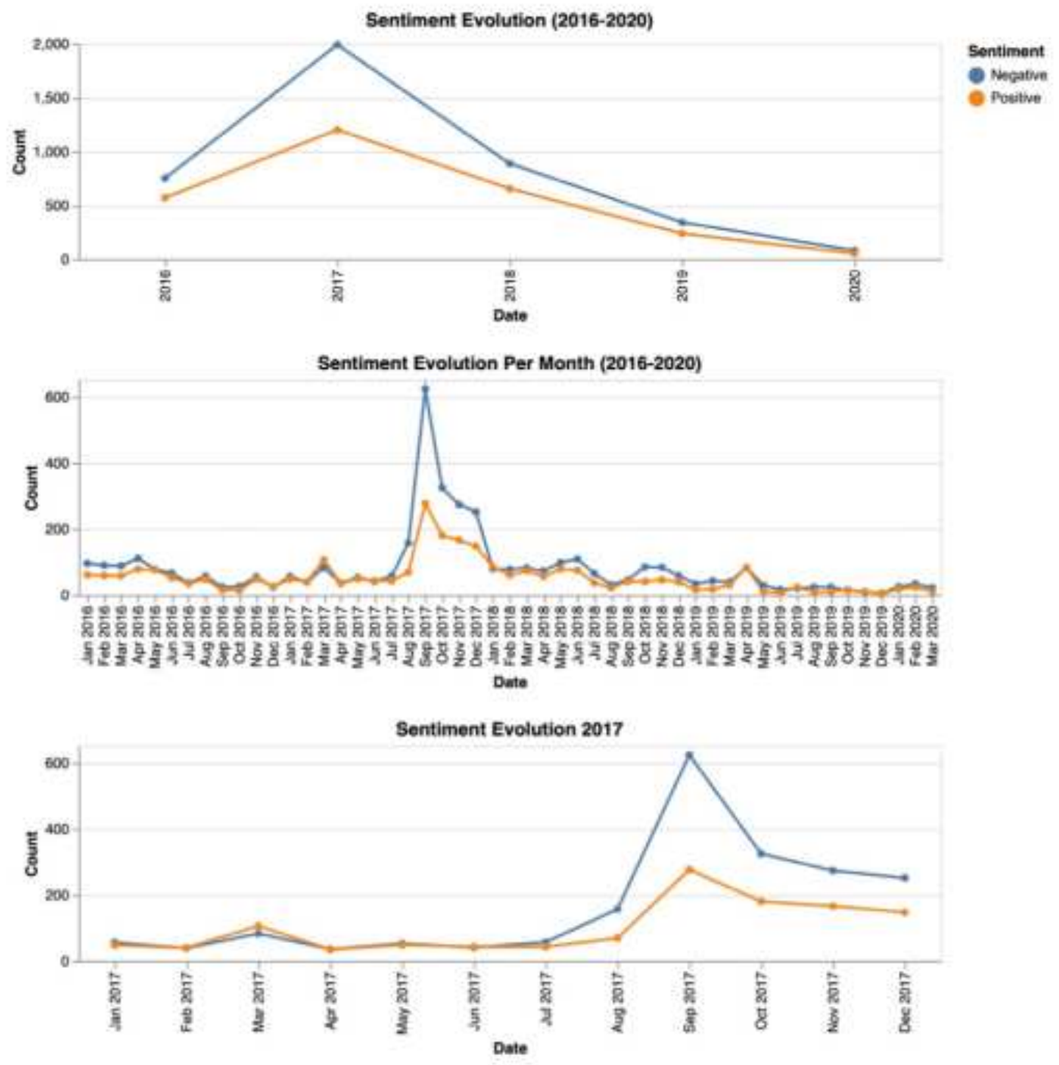

Fig. 10. Evolution of positive and negative sentiments in the patients' comments per year from 2016 to 2020 (top), per month from 2016 to 2020 (middle), and per month during the year 2017.

Table 2
The highest correlations between the 2 bigrams among the top identified bi-grams during the years 2016–2020.

Bi-gram 1	Bi-gram 2	Correlation
ancien levothyro	nouveau formule	0.9999287212
ancien formule	nouveau formule	0.9995480603
ancien formule	formule levothyrox	0.9993103979
ancien formule	ancien levothyro	0.9992348913
ancien levothyro	effet secondaire	0.9989443185
effet secondaire	nouveau formule	0.9988471039
formule levothyrox	nouveau formule	0.9988416757
ancien levothyro	formule levothyrox	0.9982914213
ancien formule	effet secondaire	0.9971094253
effet secondaire	formule levothyrox	0.9966794925

4.1.4. Adverse drug reactions analysis

Fig. 11 shows the Adverse Drug Reactions occurrence count in the patients comments between 2016 and 2020 (upper plot) and during the year 2017 (lower plot). The plot of the Adverse Drug Reactions count using a yearly resolution shows that there are some periods that witness an increase of the Adverse Drug Reactions occurrence in the patients reviews. This increase might be due to several factors depending on the

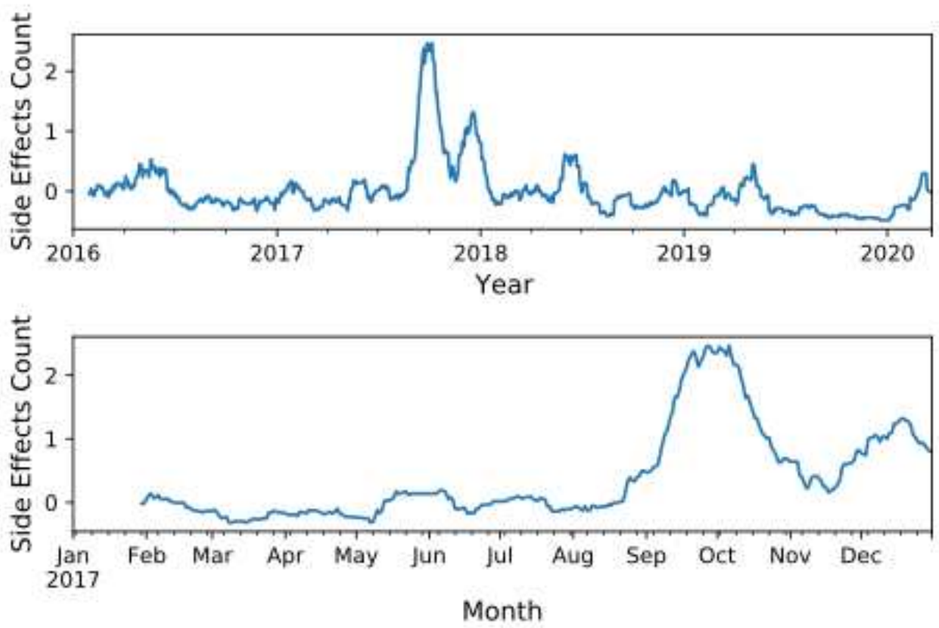

Fig. 11. Adverse Drug Reactions occurrence in the patients comments between 2016 and 2020 (upper plot) and during the year 2017 (lower plot).

Fig. 12. Visualization of the most common Adverse Drug Reactions detected from the corpus (period 2016–2020).

event that triggered the patients to complain about the Adverse Drug Reactions that they are experiencing. One notable peak happened at the end of year 2017 which coincides with the Levothyrox® scandal. Zooming on the Adverse Drug Reactions occurrence during this year. The plot shows that the number of Adverse Drug Reactions started to increase from the middle of August and reached it peak beginning of September, lasting till October where we start to see a decrease of the count of Adverse Drug Reactions.

We detect the most common Adverse Drug Reactions and perform the same analysis as above. Fig. 12 shows the word cloud of the most common Adverse Drug Reactions detected from the corpus. Fig. 14 also shows the top n-gram Adverse Drug Reactions (period 2016–2020). Fig. 13 shows the plots of the counts of the most common Adverse Drug Reactions. The plots show the same trend detected in Fig. 11.

4.2. WC-CNN performance evaluation

In order to choose the best CNN architecture for the abnormal period detection problematic, we have built and evaluated 11 CNN architectures by varying the number of layers. Due to space constraints, we do not report the quantitative results of these models. From this evaluation, it is observed that the prediction performance increases with increasing number of convolution layers. Consequently, we only report the results corresponding to the architecture described in Section 3.4.2 which represent the architecture resulting in the highest performance.

In the following, we report the performance evaluation of the proposed Word Cloud CNN (WC-CNN) for abnormal period classification. We evaluate the performance of the model against three variables: (1) the defined abnormal period from which the abnormal samples are sampled from, (2) the text pre-processing techniques used to preprocess the corpus, and (3) the temporal resolution used for extracting the word clouds (daily, weekly, monthly, all). The aim is to explore

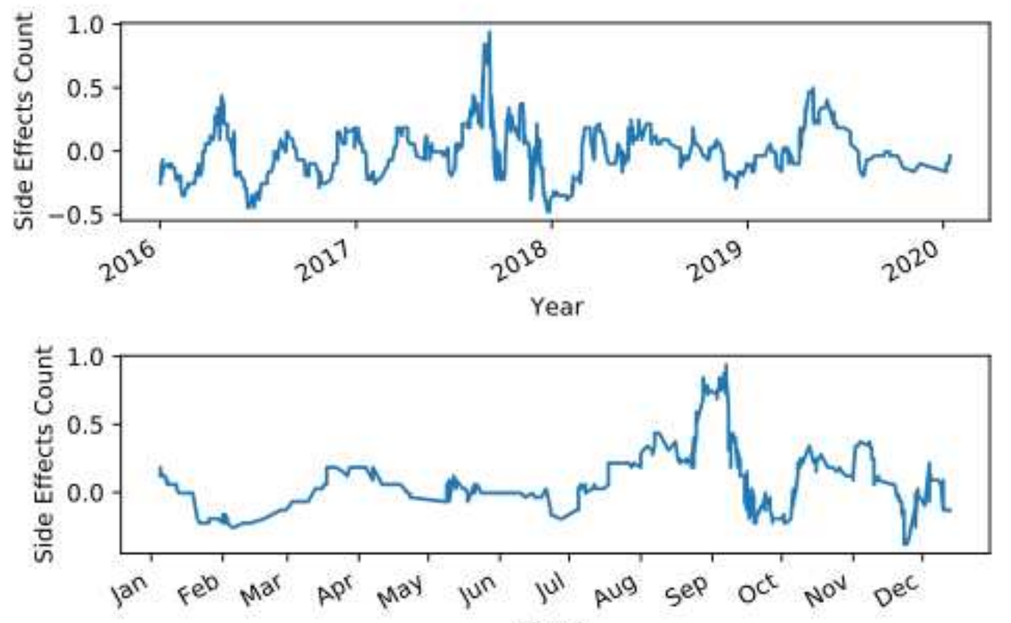

Fig. 13. Most common Adverse Drug Reactions occurrence in the patients comments between 2016 and 2020 (upper plot) and during the year 2017 (lower plot).

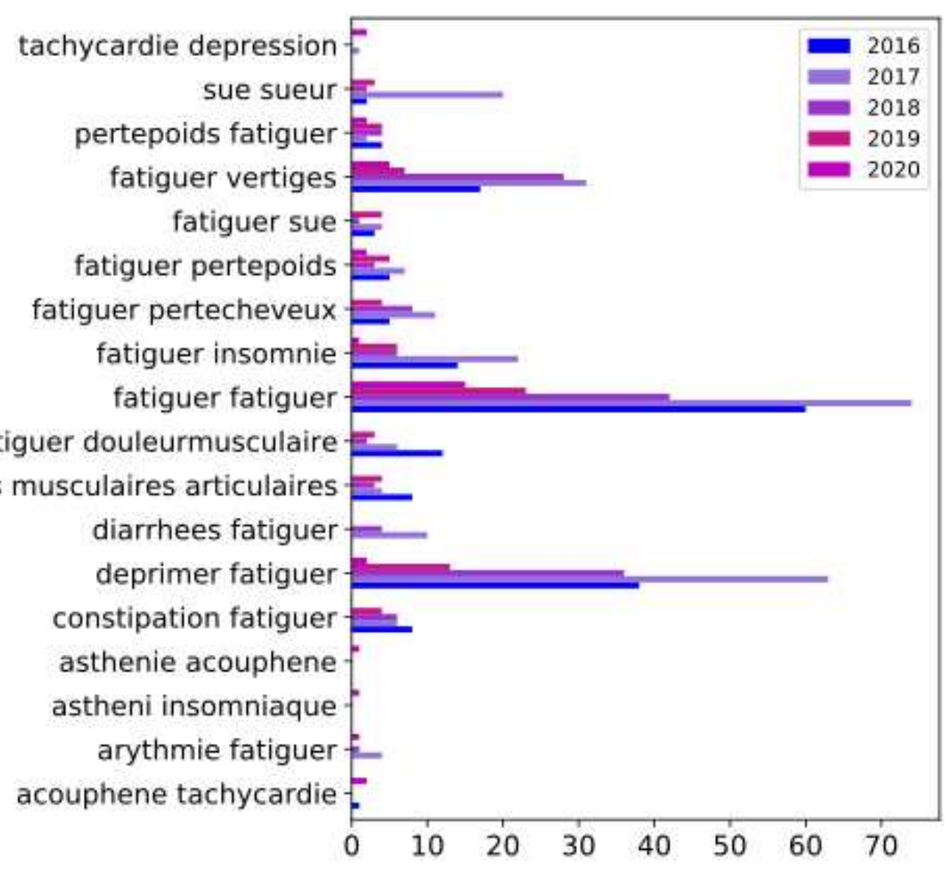

Fig. 14. Top n-gram Adverse Drug Reactions (period 2016–2020).

different combinations of text pre-processing techniques and word clouds extracted at different temporal resolutions to determine which combination results in the highest predictive accuracy for abnormal period detection.

Abnormal Period Definition – In order to test the performance of the neural network, the word clouds were labeled normal or abnormal according to the following three periods of abnormality. Fig. 15 is an illustration of these periods.

1. *July to December 2017*: information is relayed in the media.
2. *May 2017 to February 2018*: we increased the abnormal period duration by 2 months from each side (before and after the information was relayed in the media).
3. *March 2017 to April 2018*: we increased the abnormal period duration by 2 months period around the second period.

In order to obtain balanced datasets, the normal period is selected taking into account the number of data samples in the abnormal period. We select the same number of abnormal data sample from the period before and after the defined abnormal period.

We hypothesize that an early detection can be confirmed if:

- The performance of the neural network is good over the period July to December 2017. This means that there is indeed

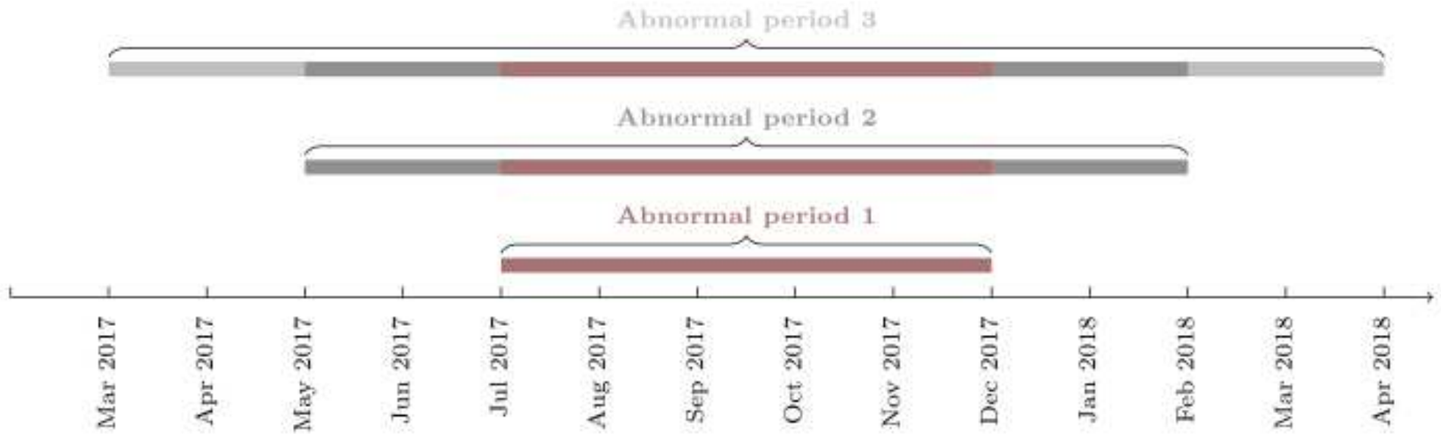

Fig. 15. Illustration of the three defined periods of abnormality used to evaluate the WC-CNN model. The word cloud samples assigned the negative label (abnormal) are extracted from the defined abnormal period. We select the same number of abnormal data sample from the period before and after the defined abnormal period.

Table 3
WC-CNN performance for the detection of abnormal periods (safety signals) using different combinations of pre-processing techniques and different Training periods data. Abnormal Period: July to December 2017.

Abnormal Period: July to December 2017				
Pre-processing	Training period			
	Day	Week	Month	All
$R_{W<3}$ R_{SW} *Lem* *Lem*$^{+}$ *CW*	0.557	0.519	0.250	0.572
$R_{W<3}$ R_{SW} *Lem* *Lem*$^{+}$	0.554	**0.611**	0.500	0.527
$R_{W<3}$ R_{SW} *Lem* *CW*	0.548	0.593	**0.667**	0.493
$R_{W<3}$ R_{SW} *CW*	0.533	0.463	0.333	0.527

Table 4
WC-CNN performance for the detection of abnormal periods (safety signals) using different combinations of pre-processing techniques and different Training periods data. Abnormal Period: May 2017 to February 2018.

Abnormal Period: May 2017 to February 2018				
Pre-processing	Training period			
	Day	Week	Month	All
$R_{W<3}$ R_{SW} *Lem* *Lem*$^{+}$ *CW*	0.531	**0.625**	**0.750**	0.566
$R_{W<3}$ R_{SW} *Lem* *Lem*$^{+}$	0.547	0.523	0.550	0.537
$R_{W<3}$ R_{SW} *Lem* *CW*	0.494	0.455	0.400	0.519

a difference between the clouds of words labeled normal and abnormal.

- The results observed over the period May 2017 to February 2018 and/or March 2017 to April 2018 are as good as over the period July to December 2017. This confirms that the word clouds of two or three periods of abnormality are similar (as the three abnormal periods include the period July to December 2017)

If the two previous conditions are met, it can be concluded that a difference with the period of normality is identifiable and that early detection is possible.

Text Pre-processing Techniques – For the above defined three abnormal periods, five different word cloud pre-processing techniques are applied to test whether the variations in the pre-processing have an effect on the prediction performance. This would inform the best combination of pre-processing for the deep learning model to learn to differentiate between normal and abnormal periods.

Temporal Resolution – Word clouds are produced at three different frequencies: daily, weekly and monthly, in order to know the most relevant data to use to obtain the best prediction results. A combination of the data extracted at these three frequencies is also performed. The advantage of defining three periods of abnormality is to identify whether an early detection of the problem related to the change in formula of Levothyrox® is possible.

Table 3 summarizes the results of the model trained on data from the period July to December 2017. Table 4 summarizes the results of the model trained on data from the period May 2017 to February 2018. Table 5 summarizes the results of the model trained on data from the period March 2017 to April 2018.

4.2.1. Comparison of the three defined abnormal periods

We compare the performance of the model in function of the defined period of abnormality. Looking at Tables 3 and 4, one can notice that the highest accuracy is obtained when the model is trained on data from the abnormal period of May 2017 to February 2018 (accuracy = 0.750). This is followed by the abnormal period of July 2017 to December 2018, with a maximum accuracy of 0.667. For the period of March 2017 to April 2018, Table 5 shows that the results are worse than the models trained on data samples from the other abnormality periods with accuracy levels less than 0.6 (highest obtained accuracy is 0.589). During tests with other neural networks, the period March 2017 to April 2018 frequently shows results a little below the other two. This poorer result can be explained because the new formula arrived on the market in March and was still little consumed by patients. Also, the end of the period of abnormality extending until April 2018, the patients exchanged less on the subject after December 2017. The word clouds are less close to those observed between July and December 2017 which is the period at during which an explosion in the frequency of comments is observed on Doctissimo® . The results of the periods of abnormality July to December 2017 and May 2017 to February 2018 being just as good, it is possible to conclude that an early detection of signals indicating abnormal events is identifiable from the month of May 2017.

4.2.2. Temporal resolution effect

It is important to determine which temporal resolution is the optimal one for extracting the word clouds and classifying an abnormal period. For instance, a word cloud extracted at a daily/weekly/monthly temporal resolution represent a summary of the most frequent words in all the posted comments during a certain day/week/month. The question is whether monitoring patients commenting behavior on a daily basis is sufficient to detect an abnormal period and signal investigatory action, or whether a weekly or a monthly temporal resolution is more informative.

From Table 3, one can see that for the July to December 2017 abnormality period, the highest prediction is obtained via monthly word clouds (accuracy 0.667), followed by weekly word clouds with an accuracy of 0.611.

Looking at Table 4, the same conclusions are drawn for the period of abnormality of May 2017 to February 2018 with three very good results of 0.75 and 0.65 for monthly word clouds and 0.625 for weekly word clouds.

It is observed that the prediction performance is better on weekly and monthly word clouds. It is not observable on this neural network but during the various tests carried out, it is identified that the monthly word clouds present better results than the weekly ones. When all the word clouds are combined, the results are worse and almost identical to those observed with daily word clouds (still around 0.5). This observation is explained by the very large number of daily word clouds compared to weekly and monthly which greatly reduce the performance of prediction.

Table 5
WC-CNN performance for the detection of abnormal periods (safety signals) using different combinations of pre-processing techniques and different Training periods data. Abnormal Period: March 2017 to April 2018.

Abnormal Period: March 2017 to April 2018				
Pre-processing	Training period			
	Day	Week	Month	All
$R_{W<3}$ R_{SW} Lem Lem^{+} CW	0.566	0.476	0.500	0.574
$R_{W<3}$ R_{SW} Lem Lem^{+}	0.547	0.444	0.571	0.553
$R_{W<3}$ R_{SW} Lem CW	0.531	0.532	0.464	0.541
$R_{W<3}$ R_{SW} CW	**0.584**	**0.589**	0.429	0.555

4.2.3. Data pre-processing effect

We analyze the effect of text pre-processing techniques on the performance of the proposed WC-CNN model for safety signal detection. The pre-processing steps are the following.

1. $R_{W<3}$: Removal of comments containing less than three words. Typically, a complete sentence is composed of at least three components: a subject, a verb, and a complement. Consequently, removing comments containing less than three words might improve the model's performance as it would be trained on less noisy data.
2. R_{SW}: Removal of "stopwords".
3. Lem: Lemmatization of words in the comments.
4. Lem^{+}: Improvement of lemmatization by creating several lists to correct the spelling of words observed in word clouds.
5. CW: Creation of a list called "$wordstodelete$", intended to clean the word clouds of parasitic words unrelated to the medical context.

Data pre-processing has little impact on the prediction performance of neural networks. Indeed, results greater than 0.6 are observed with all the different cleaning operations. These observations are also found during tests carried out by varying the parameters of the neural networks.

5. Discussion

In this section, we discuss the findings of the work as well as the strengths and limitations.

5.1. Results discussion

The first phase of the work consisted of extracting and then cleaning the data. The second phase which analyzed the best bi-grams over the period 2016 to 2020 revealed a single frequency peak between August 2017 and January 2018. The most frequent bi-gram ("Old Formula"), has a rate of appearance 160 times greater in August 2017 than in March 2017 and 6 times greater than in January 2018. "Side effects" appears 8 times in 2016, 14 times in 2019 and 183 times in 2017. The same profile is observed for other bi-grams ("New Formula", "Old Formula", "Doser Levothyrox"). The third phase of algorithmic processing was aimed at extracting and analyzing undesirable effects. In mid-September 2017, the comments showed an average frequency of occurrence of 25 per day against 6 per day between 2016 and 2020. The application of a normalization function to the curve of occurrence of adverse effects as a function of time concluded that an unusual and significant event did indeed occur in 2017. The fourth phase made it possible, through the use of convolutional neural networks, to identify similarities between the terms used during the period of May 2017 to February 2018. Also, the algorithm confirmed that the terms cited during the period of abnormality are different from those of the period of normality. Thus, it was concluded that detection of early signals, indicators of abnormal events was possible from May 2017.

5.2. Strengths and limitations

The work carried out in this paper is a first step towards a holistic approach for pharmacovigilance process optimization, and consequently, safety signals detection. The approach was built iteratively, in the absence of a reference. It is based on the use of medical knowledge and data processing. The approach is holistic in the sense that it does not focus on ADR extraction, but rather investigates other signals pertaining to the users' behaviors in function of time during a real world pharmacovigilance use case.

The study is mainly based on the Levothyrox® case in France which is a real-world pharmacovogilance case study that presents an important test-bed for pharmacovigilance research. If we take a closer look at the timeline of the Levothyrox® case, we can assert the following. The change of the medication formula took place in March 2017. In August 2017, the case received a massive media coverage due to an increased posting behavior of patients, characterized by an increased reporting of ADR on different channels, which has led the manufacturing laboratory to remove the new formula from the Market. The analysis of the users behavior during the time gap between the change of formula and the new formula withdrawal allows to explore whether the early detection of safety signals would have been possible. Moreover, considering that this time period is a period representing abnormal posting behaviors of patients, unsupervised methods for obtaining data annotations were explored, circumventing the necessity of manual data annotations, which can be a cumbersome task.

Instead of relying solely on the identification of ADE in posts for safety signal detection [12], the proposed approach uses data analysis techniques to analyze different indicators of safety signals in addition to a deep learning approach that classifies time periods as normal vs. abnormal to indicate an abnormal activity of patients on health forums. The interest of this approach lies in identifying possible unexpected phenomena pertaining to the users' posting behavior on medical forums, which can be used to conduct in-depth pharmacovigilance investigation.

The emergence of a problem around Levothyrox® change of formula in France is visible starting from the end of summer 2017 and this work confirms it by presenting findings of the only existing analysis of user comments on a single platform. Based on the statistical analysis of time series representing the frequency of words or n-grams, it turns out that many common vocabulary words of little or no relevance pollute statistical processing. It therefore becomes essential to use appropriate and codified terminologies, in order to efficiently create an extremely precise medical reference. This can help build dictionaries of common words aimed at eliminating background noise and spotting unwanted effects in a reproducible manner across all drug classes and all medical data. Classifications such as ICD-10, ATC classification should then be used. It is then a matter of bringing together under a well-designated entity all the words relating to a specific symptom, including spelling errors and the different ways of describing that symptom.

The use of basic statistical tools does not allow the early detection of abnormal events to be demonstrated. Indeed, the application of a normalization function to the daily occurrence of the Adverse Drug Reactions reported in the messages is the clear proof of the existence of an abnormal and significant event during the period when the "scandal" erupted in the messages. However, it does not make it possible to highlight the occurrence of abnormal events before July 2017. It is through the use of more sophisticated computer tools that early detection of adverse events is possible. The convolutional neural network, which is a deep learning tool, effectively makes it possible to detect the first abnormalities from May 2017. It is these tools that will improve the detection of weak signals and therefore anticipate drug problems. The results obtained are very encouraging because there is real potential to improve them. Indeed, in this work, biases limit the ability to obtain better results and to deploy and generalize these methods. Only one drug has been studied, through messages from users

of a single discussion forum. The drug in question was the subject of a case, not because of the harmful action of the active principle (by its toxic nature or its mechanism of action) but because of a change of formulation which involved communication failures. The iatrogenic effect, real for some patients but quickly resolved by dose adjustment, is not the only cause. It could be interesting and very relevant to apply this methodology on a minimum number of information sources (social networks) and specialties to compare the results obtained.

6. Conclusion

In this paper, we propose a new pharmacovigilance optimization approach by considering the use of data science and machine learning to collect and analyze real world data from patients. We focus on a famous drug use case: the Levothyrox® case. We develop an AI-based method for the early prediction of the undesirable effects of the drug mentioned on the endocrinology sub-forum of the Doctissimo® site. The proposed approach is holistic, and differs from existing works by analyzing the patients behaviors on medical forums at different time resolutions. This pioneering work in the field of pharmacovigilance presents very encouraging results. It demonstrates a real capacity for innovation in the use of data science and artificial intelligence for statistical processing of real-life patient data. It makes it possible to consider, after adjusting for biases and setting up avenues for improvement, the extrapolation of the model to other data sources and other similar events or scandals. Possible avenues of research include the use of clustering approaches on the extracted temporal data.

Declaration of competing interest

- All authors have participated in (a) conception and design, or analysis and interpretation of the data; (b) drafting the article or revising it critically for important intellectual content; and (c) approval of the final version.
- This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue.
- The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript.

References

[1] Organization WH, et al. The importance of pharmacovigilance. World Health Organization; 2002.

[2] Le Covec E, Biopharma K, Lasne BLR, Care E. Adverse drug reactions on social media: Bias and limitation. In: PhUSE EU connect 2018. 2018, p. 1–13.

[3] Hauben M, Aronson JK. Defining 'signal' and its subtypes in pharmacovigilance based on a systematic review of previous definitions. Drug saf 2009;32(2):99–110.

[4] Pickering TG, Gerin W, Schwartz AR. What is the white-coat effect and how should it be measured? Blood Press Monitor 2002;7(6):293–300.

[5] Bate A, Reynolds RF, Caubel P. The hope, hype and reality of Big Data for pharmacovigilance. Ther Adv Drug Saf 2018;9(1):5–11.

[6] Agency EM. Guideline on good pharmacovigilance practices (GVP). Annex I—Defin (Rev 4) 2017.

[7] Lardon J, Abdellaoui R, Bellet F, Asfari H, Souvignet J, Texier N, et al. Adverse drug reaction identification and extraction in social media: a scoping review. J Med Internet Res 2015;17(7):e171.

[8] El-Allaly E, Sarrouti M, En-Nahnahi N, Ouatik SEA. An adverse drug effect mentions extraction method based on weighted online recurrent extreme learning machine. Comput Methods Programs Biomed 2019;176:33–41.

[9] Arnoux-Guenegou A, Girardeau Y, Chen X, Deldossi M, Aboukhamis R, Faviez C, et al. The adverse drug reactions from patient reports in social media project: Protocol for an evaluation against a gold standard. JMIR Res Protocols 2019;8(5):e11448.

[10] Klein A, Alimova I, Flores I, Magge A, Miftahutdinov Z, Minard A-L, et al. Overview of the fifth social media mining for health applications (#SMM4H) shared tasks at COLING 2020. In: Proceedings of the fifth social media mining for health applications workshop & shared task. 2020, p. 27–36.

[11] Magge A, Klein A, Miranda-Escalada A, Ali Al-Garadi M, Alimova I, Miftahutdinov Z, et al. Overview of the sixth social media mining for health applications (#SMM4h) shared tasks at NAACL 2021. In: Proceedings of the sixth social media mining for health (#SMM4H) workshop and shared task. 2021, p. 21–32.

[12] Magge A, Tutubalina E, Miftahutdinov Z, Alimova I, Dirkson A, Verberne S, et al. DeepADEMiner: a deep learning pharmacovigilance pipeline for extraction and normalization of adverse drug event mentions on Twitter. J Am Med Inform Assoc 2021;28(10):2184–92.

[13] Bekhuis T, Kreinacke M, Spallek H, Song M, O'Donnell JA. Using natural language processing to enable in-depth analysis of clinical messages posted to an internet mailing list: a feasibility study. J Med Internet Res 2011;13(4):e98.

[14] Bigeard É, Grabar N. Detection and analysis of medical misbehavior in online forums. In: 2019 sixth international conference on social networks analysis, management and security (SNAMS). IEEE; 2019, p. 7–12.

[15] Jiménez-Zafra SM, Martín-Valdivia MT, Molina-González MD, Ureña-López LA. How do we talk about doctors and drugs? Sentiment analysis in forums expressing opinions for medical domain. Artif Intell Med 2019;93:50–7.

[16] Lee CY, Chen Y-PP. Prediction of drug adverse events using deep learning in pharmaceutical discovery. Brief Bioinform 2021;22(2):1884–901.

[17] Rivas R, Montazeri N, Le NX, Hristidis V. Automatic classification of online doctor reviews: evaluation of text classifier algorithms. J Med Internet Res 2018;20(11):e11141.

[18] Cocos A, Fiks AG, Masino AJ. Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in Twitter posts. J Am Med Inform Assoc 2017;24(4):813–21.

[19] Lee CY, Chen Y. Machine learning on adverse drug reactions for pharmacovigilance.. Drug discovery today 2019;24(7):1332–43.

[20] Fan B, Fan W, Smith C, et al. Adverse drug event detection and extraction from open data: A deep learning approach. Inf Process Manage 2020;57(1):102131.

[21] Casassus B. Risks of reformulation: French patients complain after merck modifies levothyroxine pills. Br Med J 2018;360. [Online].

[22] Concordet D, Gandia P, Montastruc J-L, Bousquet-Mélou A, Lees P, Ferran AA, et al. Why were more than 200 subjects required to demonstrate the bioequivalence of a new formulation of levothyroxine with an old one? Clin Pharmacokinet 2020;59(1):1–5.

[23] Nicolas P. Comment on:"Why were more than 200 subjects required to demonstrate the bioequivalence of a new formulation of levothyroxine with an old one?". Clin Pharmacokinet 2020;59(2):273–5.

[24] Gu J, Wang Z, Kuen J, Ma L, Shahroudy A, Shuai B, et al. Recent advances in convolutional neural networks. Pattern Recognit 2018;77:354–77.

[25] Bousquet C, Dahamna B, Guillemin-Lanne S, Darmoni SJ, Faviez C, Huot C, et al. The adverse drug reactions from patient reports in social media project: five major challenges to overcome to operationalize analysis and efficiently support pharmacovigilance process. JMIR Res Protocols 2017;6(9):e6463.

[26] Greaves F, Ramirez-Cano D, Millett C, Darzi A, Donaldson L. Use of sentiment analysis for capturing patient experience from free-text comments posted online. J Med Internet Res 2013;15(11):e239.

[27] Korkontzelos I, Nikfarjam A, Shardlow M, Sarker A, Ananiadou S, Gonzalez GH. Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts. J Biomed Inform 2016;62:148–58.

[28] Doing-Harris KM, Zeng-Treitler Q. Computer-assisted update of a consumer health vocabulary through mining of social network data. J Med Internet Res 2011;13(2):e37.

[29] Gusev A, Kuznetsova A, Polyanskaya A, Yatsishin E. BERT implementation for detecting adverse drug effects mentions in Russian. In: Proceedings of the fifth social media mining for health applications workshop & shared task. 2020, p. 46–50.

[30] Miftahutdinov Z, Sakhovskiy A, Tutubalina E. Kfu nlp team at smm4h 2020 tasks: Cross-lingual transfer learning with pretrained language models for drug reactions. In: Proceedings of the fifth social media mining for health applications workshop & shared task. 2020, p. 51–6.

[31] Bollegala D, Maskell S, Sloane R, Hajne J, Pirmohamed M, et al. Causality patterns for detecting adverse drug reactions from social media: text mining approach. JMIR Public Health Surv 2018;4(2):e8214.

[32] Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018, arXiv preprint arXiv:1810.04805.

[33] Abdellaoui R, Schück S, Texier N, Burgun A. Filtering entities to optimize identification of adverse drug reaction from social media: how can the number of words between entities in the messages help? JMIR Public Health Surv 2017;3(2):e6577.

[34] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. Roberta: A robustly optimized bert pretraining approach. 2019, arXiv preprint arXiv:1907.11692.

[35] Nikfarjam A, Sarker A, O'connor K, Ginn R, Gonzalez G. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. J Am Med Inform Assoc 2015;22(3):671–81.

[36] Park SH, Hong SH. Identification of primary medication concerns regarding thyroid hormone replacement therapy from online patient medication reviews: text mining of social network data. J Med Internet Res 2018;20(10):e11085.

[37] Marchello G, Fresse A, Corneli M, Bouveyron C. Co-clustering of evolving count matrices with the dynamic latent block model: application to pharmacovigilance. Stat Comput 2022;32(3):1–22.

[38] Sidorov G, Velasquez F, Stamatatos E, Gelbukh A, Chanona-Hernández L. Syntactic n-grams as machine learning features for natural language processing. Expert Syst Appl 2014;41(3):853–60.

[39] Ali A, Alfaycz F, Alquhayz H. Semantic similarity measures between words: A brief survey. Sci Int(Lahore) 2018;30(6):907–14.

[40] Harispe S, Ranwez S, Janaqi S, Montmain J. Semantic similarity from natural language and ontology analysis. Synth Lect Hum Lang Technol 2015;8(1):1–254.

[41] Joulin A, Grave E, Bojanowski P, Douze M, Jégou H, Mikolov T. FastText.zip: Compressing text classification models. 2016, arXiv preprint arXiv:1612.03651.

[42] Friedl JE. Mastering regular expressions. O'Reilly Media, Inc.; 2006.