

Patient-patient interactions visualization for drug side effects in patients' reviews

Zaher Salah¹, Esraa Elsoud², Kamal Salah³, Waleed T. Al-Sit^{4,5}, Manal Maaya'a¹,
Ahmad Al Khawaldeh⁶

¹Department of Information Technology, Faculty of Prince Al-Hussein Bin Abdullah II for Information Technology, The Hashemite University, Zarqa, Jordan

²Department of Computer Science, Faculty of Information Technology, Zarqa University, Zarqa, Jordan

³Deanship of Preparatory Year and Supporting Studies, Imam Abdulrahman Bin Faisal University, Dammam, Saudi Arabia

⁴Department of Computer Engineering, Faculty of Engineering, Mutah University, Al-Karak, Jordan

⁵Department of Information Technology, Faculty of Computer Information Science, Higher Colleges of Technology, Dubai, UAE

⁶Department of Pharmaceutical Chemistry, College of Pharmacy, Jerash University, Jerash, Jordan

Article Info

Article history:

Received Feb 13, 2024

Revised Feb 28, 2024

Accepted Mar 10, 2024

Keywords:

Artificial intelligence

Data mining

Interactions visualization

Medical data analysis

Natural language processing

Opinion mining

Sentiment analysis

ABSTRACT

This paper describes the patient-patient interactions (PPIs) graph extraction framework from patient's review transcripts. The concept is to visualise patients as nodes and interactions representing links. Links are made based on review text similarity. Nodes are categorized as positive or negative according to the patient's attitude toward a given drug. Attitudes are then utilized to categorize the links as supporting or opposing the use of a certain drug. If both patients share the same attitude: negative (severe side effect) or positive (moderate side effect), the relationship is considered supportive; if not, the link is considered opposed. Resulting graph represent a drug as a dispute between two factions arguing on related drug. The framework is explained and evaluated using a dataset included 3,763 patients' reviews linked to 255 different drugs, -predictive-value (0.37). We argue that, this is caused by derogatory jargon that is an expected feature of patient's review. The true-negative-recognition-rate is 0.70 and true-positive-recognition-rate is 0.54. Total-average-accuracy, which is independent of class priors, is 0.66. Results show that, it is possible to use text proximity measures and sentiment analysis to capture PPIs structure.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Zaher Salah

Department of Information Technology, Faculty of Prince Al-Hussein Bin Abdullah II for Information Technology, The Hashemite University

Zarqa, Jordan

Email: zaher@hu.edu.jo

1. INTRODUCTION

Text visualization is a method for turning text into simplified graphs that make texts easier to read and save our time when we are looking for certain tasks related to our information needs more quickly and effectively. Sentiment analysis is the use of natural language processing (NLP) to follow up public opinion around a specific product, issue, or trend. Sentiment analysis, commonly referred to as opinion mining, is the process for recognizing positive and negative feelings, opinions, viewpoints, attitudes, and sentiments using data mining approaches [1]. Typically, this process analyzes opinions on an object of interest that is incorporated in a text. This object of interest could be a drug (medication), a person, a movie, an occasion, or a product [2]. Consequently, opinion mining aims to automatically extract subjective information rather than factual or objective information from various forms of textual data. The degree of subjectivity incorporated in

the textual data varies, therefore texts can be classified as emotionally poor or emotionally rich based on the amount of positive and negative subjective terms that are used throughout the text. One particular type of opinion mining that focuses on the medical field is called medical opinion mining. This research paper focuses on using sentiment analysis techniques to extract argument (discussion like) graphs describing patient review transcripts from medical textual data. The purpose of the graphs is to effectively visualize certain complex structure of the arguments, such as who discusses related topics (and to what degree) and who is opposed to whom (and to what degree).

More specifically, this research paper describes the patient-patient interactions (PPIs) graph extraction framework in more detail, which enables the representation of patient reviews as sets of connected nodes, with nodes standing in for patients and links for significant interactions between patients. If there is a sufficient degree of similarity (semantic similarity) between the patient reviews that each individual patient has made, then the interaction between the patients seems significant. Next, sentiments are used to label nodes and links between nodes. Depending on whether a patient is in favor of or against the side effects of the medication, nodes are labeled with patient's attitude, which can be either positive (mild side effect) or negative (severe side effect). Following the identification of the patients' (nodes') attitudes, the links may be labeled correspondingly. A link (relationship) is labelled as "supporting" and is colored green if the two nodes it connects share the same label for attitude. In the case where the two nodes have distinct labels for attitude then the label for link will be opposing and its color will be red. When two patients are linked, the thickness of the link represents the semantic similarity, which is determined by accumulating all the terms (words) with non-zero term frequency-inverse document frequency (TF-IDF) weights in both patients' reviews that seem to be about the same related topic. The generated graphs provide a high-level structural visualization of the interaction between the two sides of opposing fractions. The textual patient review transcripts (focusing on drug side effects) dataset, which includes patient reviews, ratings, and comments submitted by patients or caregivers and published online in HTML format (at www.druglib.com), acted as the main focus of the research work described in this paper. Graph generating process was conducted on this dataset. To the best of our knowledge, there is no one has conducted similar experiments on this type of datasets before. The structure of this paper is as follows. In section 1, previous studies on lexicon-based sentiment analysis is briefly reviewed and some background is provided. The application domain and the dataset are covered in section 2. The PPIs graph extraction framework is introduced in section 3 along with an example for explanation. In section 4, an assessment of the proposed framework is discussed. After that, section 5 introduces some findings and presents potential future extensions for the proposed research work.

As mentioned before, the analysis of patients' review transcripts is the focus of the research work presented in this paper. More precisely, it is focused on two specific objectives: (i) using sentiment analysis methods and tools to discover patients' "attitudes"; and (ii) visualizing the PPIs structure as graphs. An overview of previous studies that is relevant to the rest of the paper is given in this section. We begin by giving context for our research by introducing sentiment lexicons and then discussing prior work that has utilised opinion mining methods and sentiment analysis in drug side effects graph analysis. The distinct words inside a text act as sentiment indicators for determining whether it should be categorized as positive or negative. However, due to the complexity and richness (e.g. synonyms, homonyms, and irony) of natural languages, subjective word recognition is a difficult task. One option is to utilize sentiment lexicons to search up words to recognize independent (as different to objective) terms and then assess the degree of sentimentality and variation of attitude (positive or negative) related to the independent words detected. This data may then be utilised to disclose an opinion of the text's overall sentiment. The core concept is to aggregate independent word level sentiment scores to get a document level sentiment score. More specifically, the primitive word level sentiment scores may be merged to generate more complex sentiment scores for various levels of text inclusion like sentence level or paragraph level, allowing for a determination of the polarity of these complex levels.

Sentiment lexicons are a kind of lexicon used to classify sentiments. They give individual words an emotion (sentiment) score, and an orientation. An emotion score is a numerical measurement that indicates the degree to which something is subjective. The orientation of a word indicates whether it indicates approval or disapproval of an item or notion. As a result, the polarity of a text may be determined by computing the quantity of positive and negative phrases and accumulating their sentiment ratings. The result indicates the document's polarity (whether positive or negative). Manually constructed sentiment lexicons of relatively modest size may be expanded by beginning with a core collection of positive and negative seed words. This list is augmented via lexical initiation approaches that use the semantic links between words and their substitutes and antonyms or through term similarity measures in big corpora [3].

The research work described in this paper uses a generic sentiment lexicon, SentiWordNet 3.0, which encompasses the previous SentiWordNet 1.0 [4]. SentiWordNet links to each synset (s) found in WordNet a group of three scores: Pos(s) (positivity), Neg(s) (negativity), Obj(s) (neutrality or objectivity).

The range of each score is from 0 to 1, and for each synset s : $\text{Pos}(s)+\text{Neg}(s)+\text{Obj}(s)=1$. From the results obtained from the experiments conducted in this research work, SentiWordNet has an important benefits, over other obtainable lexicons, of containing the majority of words (SentiWordNet 3.0 covers 117659 words). The majority of published research on sentiment analysis (see for example [5]–[14]) concentrated on what might be described as typical forms of subjective text available on social networks, blogs, or dedicated websites such as news articles, movie reviews, product or service reviews. Since there is an abundant amount of literature on these conventional approaches, we restrict our focusing in this research work on methods that are directly relevant to medical sentiment analysis, which is the subject of interest for the work described in this paper.

Mining significantly improve the quality of prediction and classification of diseases when appropriately used. Many healthcare informatics fields apply data mining techniques for different purposes, such as clinical care and Administration of health services. With the use of decision support systems for clinical applications, doctors can access more relevant information rapidly, leading to faster and more effective diagnosis and treatment recommendations [15]. Recently, the diagnosis Many researchers exploited data mining approaches like classification, clustering, regression, forecasting and association rules mining in healthcare. The algorithms of data of adverse side effects depended on the data submitted by the users on social media networks. Therefore, the information submitted by users regarding drugs and their side effects on social media is an essential resource for monitoring drugs. Many studies have been conducted to identify drug side effects employing medical case reports and relevant information from social media [16]–[19]. Many users are on social media, which is why social media gives us ample opportunities as the users can interact there through other comments on drug relations. Biomedical research can use social media resources because they hold their vast importance in analyzing the interactions between products and drugs [20]. The literature shows there is not enough study to explore the natural products' connections with the drugs, but new opportunities have been explored because of social media. Social media has a wide range of tools that can be used to study the interactions and users' opinions on a particular matter. Studying certain serious situations and pathologies associated with social neglect can also be helpful. Other sources can be used, but there is a chance that they are underreported [20]. Diverse online media stages offer the extraordinary potential to screen general well-being in examining the effects of drugs. The capability of online media in pharmacovigilance (identifying, evaluating, comprehending, and preventing side effects or any other issue related to medications or vaccinations) [21] has appeared via the investigation of utilizing resources like Twitter [22], [23]. The platform of social media had just been utilized to consider other medical problems, for example, flu and Ebola infection [24], [25]. Carbonell *et al.* [26] examined the mentions of drugs on Twitter to investigate the capability of online media in the identification of drug-drug interactions (DDIs). A total of 1,456,961 mentions were downloaded, which translates to 2,406 names and 946 medications across 53 languages. English was the most frequently used language, accounting for almost 30% of the messages. The time frame for this study was three weeks, from October 6, 2014, to October 27, 2014. On Monday, October 14th, the most tweets (86,969) were posted, whereas on Saturdays, the lowest rates within the studied time were recorded. Following the filtering procedures outlined in methods described they research work, 99,485 tweets were retained for analysis. 390 categories were linked to drug names. The recorded results indicates that the most compounds cited were anti-bacterial agents (85), followed by anti-inflammatory agents (80) and antineoplastic drugs (75). The number of tweets associated with these three categories, throughout the analysis period, was of 1,140, 2,938 and 8,906, respectively.

In earlier studies, patients' review and opinion forums retain crucial information concerning preferences and experiences of users across various product domains. Using data mining techniques like sentiment analysis, this information can give the doctor valuable resources for acute side effects. Gräßer *et al.* [27] collects online user reviews by crawling online pharmaceutical sites. This data contains information on drugs effectiveness in addition to side effects. The researcher performs preprocessing tasks over drug reviews acquired from pharmaceutical review sites. To forecast the sentiments of user overall satisfaction, side effects, and efficacy of user reviews on particular drugs, the authors first performed a sentiment analysis. The authors also investigated the transferability of trained classification models between domains in an effort to address the challenge of missing annotated data. In their work, the authors demonstrated that transferring learning techniques can be used to manipulate similarities towards domains and it was a good technique for sentiment analysis in the context of cross-domain.

Even though the current DDI extraction techniques can improve performance and yield additional knowledge by utilizing external resources like ontologies or biomedical databases, their updating is challenging, which causes delays. In order to increase the amount of information available for the deep learning DDI extraction approach, Xu *et al.* [28] made use of user-generated content resources. Using a full-attention technique, they combine content generated by individual users with local contextual information to deliver new and rich knowledge. The deep learning classification algorithm of this method uses attention

outputs merged with concept embeddings and offset embeddings for entity as input. The outcomes of the experiment demonstrate that the method improved evaluation metric scores.

Medical lexicons provided by the food and drug administration's (FDA) COSTART corpus or the unified medical language system (UMLS) were employed in traditional medical NLP research. Nevertheless, conversational side effect expressions that are frequently found in submitted patient assessments (reviews) are frequently underrepresented in these official lexicons. Consequently, rather of utilizing these constrictive lexicons, we extract side effect terms from the patients' reviews. The basic method of comparing word frequency distributions between two datasets was employed in order to identify statistically significant phrase patterns. The focus was on statin drugs, which are often prescribed pharmaceuticals with a wide range of side effects. The standard statistical log-likelihood ratio calculation revealed a high correlation between statin drugs use and weakening and discomfort in the muscles. Additionally, there is a statistically significant correlation between statin drugs and a number of debilitating diseases, including heart failure, rhabdomyolysis, Parkinson's disease, and amyotrophic lateral sclerosis (ALS). The scientific literature on statins supports several findings in [29].

Data mining applications are exponentially growing in the context of development of information data sets and techniques to identify common drug side effects and can save time and resources [30]. Such automated tools are essential in generating a knowledge database. The main frameworks to separate connections from text incorporate co-event-based, rule-based, and artificial intelligent (AI) approaches [31]. Co-occurrence-based strategies are less complex and set up a connection between two substances dependent on co-occurrence. Rule-based strategies utilize semantics to comprehend the importance of a specific relationship. In order to extract phrases containing side effect and causal medication pairs, Sohn *et al.* [32] developed an integration of machine learning employing side effect keywords and pattern identification rules. This allowed the system to find the majority of side effect occurrences. Compared to individual side effect extraction, the hybrid method for detecting side effect phrases included more side effects and causal drugs combinations. In order to identify a specific side effect and the drug that is causing it, the researcher in this study manually created materials and methods for pattern matching criteria by looking at keywords and side effect patterns. The researcher gave an example of how this system could be trained to recognize side effects in complexly described sentences.

Drug target identification and repositioning is an active research area that involves approaches for researching drug side effects associations. Xu and Wang [33] automatically identified phrases and abstracts linked to side effects using prior knowledge derived from FDA drug labels regarding known drug side effects correlations. Using integrative methodology, the researcher shows how low level genetic and chemical drug processes are reflected in the higher level phenotypic drug side effects connections. An innovative knowledge-driven strategy to retrieving numerous drug side effects combinations from published biomedical literature is efficiently shown in this study. Furthermore, it demonstrates that drug repositioning can directly utilize the derived drug side effects combinations. The large-scale, higher-level drug phenotype connection knowledge that automatically created has a lot of potential applications in computational drug discovery.

One element of efficient data analytics techniques is data visualization. Using interactive visual technologies, decision-makers may quickly recognize the new trend and instantly generate real-time reports [34], [35]. Researchers studying clinical information are attracted to the visualization of clinical data because it can save time when looking up specific drug reactions or going over a medication's most frequent bad side effects. A decision assistance system that uses information visualization to speed up the inspection of possible negative reactions were created by Duke *et al.* [36] and Li *et al.* [37]. The results suggest that information visualization can greatly accelerate the evaluation of possible adverse drug occurrences. The system creates a database with 16,340 distinct drug and its side effect pairs, comprising 250 typical medications.

Oprea *et al.* [38] contrasted their tool's speed and accuracy of side effect retrieving with UpToDate[®]'s indicated a 60% decrease in query completion time (61 s vs. 155 s, $p < 0.0001$). UpToDate[®] is an electronic clinical decision support tool that is evidence-based and available at the point of care to assist medical professionals in choosing the appropriate decisions of care and producing better outcomes (<https://www.wolterskluwer.com>). Patients who take multiple drugs are more vulnerable to negative drug reactions. Although doctors can lower this risk by looking over their patients' medication side effect profiles where this process is time consuming. Regarding these scores the system creates visual maps of adverse reaction for each drug combination that the user chooses. The intricacy of biomedical data makes relationship extraction a challenging task. The association of side effects data with target information can therefore be achieved using biological network analysis. In this research work, the authors collected the result from text mining and deep mining associations of approved drugs and target information. This data comprises 7,684 approved drug labels, 988 unique drugs and 174 side effect. It then clustered based on deep component analysis into a 5×5 self-organizing map. Then a comprehensive network is generated containing drug side

effects clusters using the Cytoscape tool (<https://cytoscape.org/>). The resulted biomedical network of drug side effects facilitates drug repositioning and indicates optimal drug actions [38].

In another work described in [39], a disease side effect network was constructed based on 3,175 associations. Diseases sharing identical side effects tend to be clustered together. Disease side effects associations were visualized using Cytoscape. The generated network includes three clusters of diseases grouped by circulatory system, neuropsychiatric and neoplasms diseases. Each cluster shares diseases with the most relevant side effects, which simplify the suggestion of suitable drugs [39]. In the setting of topic based domains, sentiment analysis using general purpose sentiment lexical resources is a challenge [40]. In such cases, it is better to employ sentiment lexicons that are specialized to a given domain. The main problem with employing these specialized lexicons is that they are often not publicly available and hence need to be deliberately developed, which can be a costly and error-prone procedure. There are two methods for creating specialized lexicons for domain specific analysis of sentiments. The first one is creating a new specialized lexicon and the second is modifying an already existing general purpose lexicon. Both approaches rely on labelled corpora from a certain domain. One example of the first approach is extracting domain specific tourism or health terms from noisy textual corpora in order to generate a domain specific vocabulary [41]. By examining how terms from the general purpose lexicon are utilized with respect of a specific domain and giving these phrases new polarity, a straightforward method for adapting a general purpose sentiment lexicon to that domain was provided [6]. A general purpose sentiment lexicon can be converted into a dedicated lexicon by merging the relations between terms and opinion phrases to determine the most likely polarity of a term as positive, negative, or neutral in the given domain [42]. Two methodologies have also been published in other study [3], which involves adding new domain terms to the seed lexicon and extending it by changing the sentiment ratings of the phrases in it. By crowdsourcing assessment of sentiment phrases and extending the initial seed vocabulary automatically by bootstrapping to integrate new sentimental indicators and concepts, they constructed a domain-specific sentiment lexicon [43]. The lexicon is then modified based on a specific domain. An evaluation of the created lexicon showed that it performs better than the general inquirer sentiment lexicon, a generic sentiment lexicon. Additionally, work on developing domain-specific lexicons using the “dual approach” was disclosed [44].

There are three methods available for calculating term sentiment score: the methods include: i) examining the term's biased occurrence with respect to the positive or negative class label of a document, ii) applying statistical, contextual, or semantic links between terms in a specific domain or iii) training a classifier to predict a sentiment polarity of each term. In the process of generating sentiment scores with respect to specific domain, the sentiment score intensity is calculated for each term or phrase in various domains [45]. Sentiment indicators are the subjective terms in a fresh text that we want to classify as representing a positive or negative viewpoint. However, due to the richness of natural languages, subjective word recognition is a challenging task. Using sentiment lexicons is one approach, where words may be looked up to identify the level of sentiment and polarity connected to the found subjective phrases. Afterwards, an assessment of the overall sentiment of the text can be made using these data. The basic concept is to calculate the overall sentiment value for the textual content by accumulating the subjective word-level sentiment scores. As mentioned before, sentiment lexicons can be used for sentiment mining in two ways: we can use a standard general purpose sentiment lexicon or a specialized sentiment lexicon.

2. DRUGLIB.COM DATASET

A dataset comprising patients' reviews, ratings, and comments made by patients and caregivers and published online in HTML format (www.druglib.com homepage) was used as the main focus for the study presented in this research. Figure 1 shows an extract from accutane (isotretinoin) reviews, ratings, comments made by patients and caregivers. Based on 44 ratings and reviews (drug number 2 in our dataset). This collection's benefit is that each drug's overall, effectiveness, and side effect scores are known. As a result, we can, at least partially, evaluate the veracity of our drug graph construction framework and have somewhat more confident using the method when applying it to short-text reviews of patients (or customers) of all kinds where the outcome is unknown (or not yet known). The scores are on 10 points scale where the best is 10 and the worst is 1. This information is not vetted and should not be considered as clinical evidence.

The authors extracted the reviews, ratings, comments sub- mitted by patients and caregivers associated with 255 drugs from the www.druglib.com web site. QDAMiner4 (<http://provalisresearch.com>) was employed to extract the required textual data out of the HTML drug records. For each drug the reviews, ratings, comments made by patients and caregivers associated with the same drug were collected together. We will refer to this dataset as the DrugLib patient reviews (DPR) dataset. The dataset comprised 3,763 patients reviews associated with 255 distinct drugs. Note that the number of reviews a medication has also corresponds to the number of patients who are participating. Individual reviews can also be referred to as documents because the PPIs graph extraction architecture contains a number of approaches from the

document analysis domain. A review's average word count was (152.5). For evaluation purposes, the patient's total score (on a ten-point scale) was utilized (see section 4).

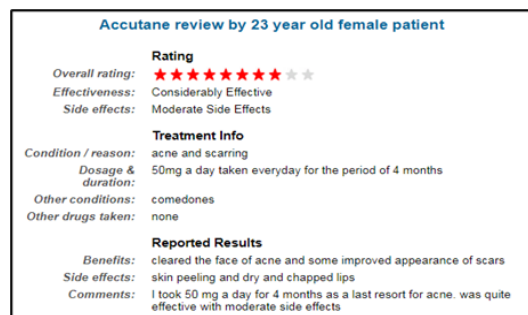


Figure 1. Accutane review, ratings and comments published on the www.druglib.com www site

3. METHOD: THE PROPOSED FRAMEWORK

As input for the PPIs system a set of patients reviews related to a drug (side effects) is used, the output will be a graph depicting the structure of all interactions associated with that drug side effects. The input to the PPIs framework is a set of n patients' reviews $R=r_1, r_2, \dots, r_n$ and the output is a graph of the form $G(V, E, Lv, LE, fmap)$ where: i) V is a set of n nodes (vertices) representing n patients' reviews in the form: $V=v_1, v_2, \dots, v_n$, ii) E is a set of m links (edges) such that $E=e_1, e_2, \dots, e_m$, iii) Lv is a set of two node labels, iv) LE is a set of two link labels, and v) $fmap$ is a function to map node and link labels onto their relevant node and link. The four phases of the PPIs graph extraction framework are: reviews preprocessing, attitude discovery and node labelling, link finding and labelling and finally generating the PPIs graph. The next four sub sections provide a more comprehensive explanation of each of these four phases.

3.1. Preprocessing

Part-of-speech (POS) tagging and text preprocessing is the first phase. Attitude prediction and node labelling is the second phase, while link identification and labelling is the third and PPIs graph construction is the final phase. Part-of-speech tagging (POST) is the process by which every word in a text is given a POS tag based on where it appears in a sentence or in a phrase. To assign a POS tag for a specific occurrence of a word, various tag sets are available. Tag sets can be fine-grained, such as the Penn Treebank POS tag set, which includes a set of 35 distinct part-of-speech tags, or very coarse, such as utilizing a small tag set of the form N, V, Adj, Adv . POS tagging is crucial in sentiment mining since several comparable words with different POS tags will usually have various sentiment values. POST is important for sentiment mining because it makes it possible to do word sense disambiguation, which resolves the polysemy issue (several meanings for one word) by identifying the proper sense (semantic) of a word in a given circumstance or context [46]. A list of terms $\{T = t_1, t_2, \dots, t_m\}$, each coupled with a POS tag $post_j$, will be generated during the POST phase, resulting in a collection of pairs of the type $term_j, post_j$.

A collection of patients' reviews is the PPIs graph extraction framework's input, as was previously mentioned. A patient's review can be considered as a document. In this context, each document represents a patient's review (or simply a patient). First, all uppercase alphabetic letters are converted to lowercase during the pre-processing phase. Next, punctuation marks and numeric digits are eliminated. Stop word removal is the following phase. Words conveying low meanings (e.g. "and" or "the") to which no particular sentiment may be associated are known as stop words [47]–[49]. As a result, stop words are eliminated from the document set too. The next step is to create a bag of words (BOW) representation, $BOW = \{t_1, t_2, \dots, t_{|BOW|}\}$, that contains all of the words that are still present in the document collection (patient reviews). Next, a subset of the BOW will represent each document. Actually, BOW1 and BOW2 two BOWs are formed. As will be seen, BOW1 and BOW2, which are generated slightly differently, are employed for attitude detection and edge identification, respectively. Lemmatization is a process used in BOW1 generation, whereas stemming is a method used in BOW2 generation. In order for "inflated" words that are part of the same stem (root) to be counted together, the process of stemming entails eliminating added affixes from a given word [48]. While many different methods are proposed to accomplish stemming, Snowball stemming was employed in this paper. Words with different POS tags will generally have various sentiment scores when it comes to sentiment analysis. But when stemming is used these words become one word, which is the stem, and share the same sentiment score and loses the actual individual sentiment scores.

Lemmatization is a substitute for stemming that can also be utilized to reduce the forms variety of words. Lemmatization goal is not to reduce words to their roots or stems, but rather to its conventional form. For instance, all nouns would be changed to their singular form and all verbs to their infinitive form [50]. For BOW1 and BOW2, respectively, lemmatization and stemming were therefore applied. Next, two feature spaces are defined using the two bags of words, allowing for the generation of two sets of feature vectors. Apart from the fact that one used lemmatization and the other stemming, the two bags of words differ in that the elements of the feature vector in BOW1 hold frequency counts for terms, whereas the elements in BOW2 hold TF-IDF weights for terms. To put it simply, a document frequency count is the total number of documents (patients' reviews) that include a given word. The most commonly used method for term weighting is the TF-IDF weighting scheme, which is also the one utilized in the context of PPIs graph extraction framework. It aims to "balance out the effect of very rare and very frequent" words in a vocabulary [51]. Because TF-IDF combines local and global word frequency, it reflect the significance of individual terms [52]. The definition of TF-IDF is $W_{ij} = TFIDF(i, j) = tf(i, j) \cdot \left(\log \frac{N}{df(j)}\right)$, where: $tf(i, j)$ is the frequency of term j in document i , N is the total number of documents in the corpus (patients' reviews) and $df(j)$ is the number of documents (patients' reviews) containing term j .

Once the pre-processing stage is complete, allowing each patient's review to be defined by a feature vector. More formally, a review i is represented as a vector $S_i = \{wi1, wi2, \dots, wiz\}$, where w_{ij} is the TF-IDF weight for word j in document (patient's review) i in the case of BOW2 and the frequency count of term j in document i in the case of BOW1. Every element in S_i is associated with a term in BOW1 or BOW2. The notation $T_i = \{ti1, ti2, \dots, tiz\}$ will be used to represent the list of terms associated with feature vector S_i . As a result, there is a one-to-one correspondence between the set of term lists $T = \{T1, T2, \dots, Tz\}$ and the set of feature vectors $S = \{S1, S2, \dots, Sz\}$. After the POST is finished, preprocessing can start. Tokenization and stop word elimination was the initial step in the preprocessing phase. The lexicon-based technique proposed in this research work not used stemming because, as previously mentioned, words such as "suffice", "sufficiency", "sufficient" and "sufficiently" will have various POS tags and, as a result, different sentiment scores. These words will become a single word (stem or root) after stemming is applied, and thus sharing the same sentiment score and therefore maybe losing their more appropriate separate sentiment scores. Alternatively, a lemmatization approach is used. Lemmatization differs from stemming in that the goal is not to reduce a given word to its root, but rather to its lemma (dictionary form). For instance, all nouns would be changed to their singular form and all verbs to their infinitive form [50]. Once tokenization, stop-word elimination, and lemmatization are finished, a BOW representation was once more employed; however, in this case, every word in the BOW is joined to a POS tag. Consequently, every document (the side effect part of the patient's review) is converted into a feature vector form by representing it with a subset of the BOW. The feature vector elements hold word frequency. More formally, a patient's review i is presented as a document vector as: $V_i = w_{i1}, w_{i2}, \dots, w_{im}$, where w_{ij} is the occurrence count of term j in patient's review i . Furthermore, it requires to be mentioned that each element in V_i is associated with a term in the BOW. The list of terms associated with feature vector V_i is indicated using the notation $T_i = \{ti1, ti2, \dots, tim\}$. As a result, a set of feature vectors $V = \{V1, V2, \dots, Vz\}$ and a set of term lists $T = \{T1, T2, \dots, Tz\}$, with a one to one relationship between them.

3.2. Attitude detection and node labelling

The feature vector weights for attitude detection are merely term frequency counts, as can be seen from the foregoing. After that, sentiment analysis is performed on the terms that correspond to every feature vector to identify the node labels. Keep in mind that every patient's review is a node. By searching the terms in the SentiWordNet lexicon, one may determine the sentiment value linked to each term in T_i which is the list of all terms associated with feature vector S_i . Sentiment lexicons assigns orientation and sentiment score to individual words. Sentiment score is a numerical value that represents the subjectivity level. Word's orientation (polarity) can be used to determine whether it conveys agreement or disagreement with a certain thing or idea. As a result, document polarity can be determined by computing the difference between the number of positive and negative words. The outcome reflects the document's polarity, whether positive or negative. SentiWordNet assigns a polarity and a sentiment score by giving each synset (set of items that are semantically comparable) in WordNet a positive and a negative value. In this paper, to generate a list of words from which the relevant score may be retrieved, SentiWordNet synsets have been divided into individual terms. Terms from the same synset are deemed to possess a same sentiment score. However, in the case that two synsets differ and the same term is derived from both, then: (i) if the term has different grammatical tagging, word sense distinction is decided by simply taking into account the different POS tags of the term (as suggested in [46]); if the term has identical grammatical tagging then it is considered as duplicated term and the highest score of the two synsets is adopted. The generic sentiment lexicon (SentiWordNet 3.0) performance in sentiment mining for identifying the sentiment polarity (attitude) of

patient-written reviews using a testing dataset from the DPR dataset (explained in section 2). Sentiment scores assigning is applied to the test data (the side effects part of the patients' reviews) using the SentiWordNet 3.0 sentiment lexicon that is adopted in order to determine the attitude of each individual patient. The predicted patient's attitude can be compared with the known attitude because this attitude is known from the way the patient eventually rank each medicine (side effect score). Ten points are awarded: 10 for the best performance, 1 for the worst. Sentiment analysis was performed on the terms (words) within each generated feature vector representation to identify the attitude reflected by each vector and thus the document (patient's review) it represents. Searching through a sentiment lexicon yields the sentiment score (value) linked to each term t_i in feature vector S_i . A sentiment score is a numerical value that indicates a certain level of subjectivity, as was previously mentioned in section 1. A word's orientation can be used to determine whether it conveys agreement or disagreement with a certain thing or idea. As a result, by counting the positive and negative words and computing the difference, the polarity of the patient's review can be determined. The patient's attitude is then described by the resulting polarity. More formally, the overall sentiment score $score_{ij}$ associated with a patient's review i can be computed using the formula: $score_{ij} = \sum_{j=1}^{j=z} (SWN(term_j) \times w_{ij})$, where: (i) $term_j$ is a term that represents the patient's review i in the feature vector, (ii) SWN is a function that takes the SentiWordNet 3.0 and returns the sentiment score for each $term_j$ ($-1.0 \leq SWN(term_j) \leq +1.0$). The sentiment score is the aggregate of the sentiment values for the $term_j$, (iii) z is the number of terms in the given vector. (iv) w_{ij} is the number of occurrences of $term_j$ in feature vector i . The occurrences count can be a binary value (0 or 1) designating whether the term is absent in document i or present, or it can be an actual frequency count (i.e. number of times $term_j$ appears in document i). The incidence count for $term_j$ is considered in the context of the research work detailed in this paper. Next, $score_{ij}$ is used to determine the attitude label for each document (patient's review). With this in mind, the defined labels for individual attitudes are: {positive, negative, objective, neutral}. Here, positive denotes a text with positive attitude (i.e. moderate side effect in our patient review's side effect part), negative denotes a text with negative attitude (severe side effect), objective denotes that no sentiment scores were discovered, and *neutral* denotes that the sentiment scores cancel each other out. In practice it was discovered that the final two class labels are rarely encountered. The process of attitude identification is explained in Algorithm 1. The algorithm iterates through the patients' reviews input set, represented in terms of the sets S and T (as previously explained in sub section 3.1). Lines 5 through 13 are used to calculate the sentiment score for each patient's review, and lines 14 through 22 are used to determine the corresponding attitude.

Algorithm 1. Attitude identification and node labelling

```

1: INPUT: SentiWordNet3.0,  $S, T \subset BOWI$ 
2: OUTPUT: Attitude labels  $A = \{a_1, a_2, \dots, a_z\}$ 
3: Pos-Count = 0, Neg-Count = 0, Pos-Score = 0, Neg-Score = 0
4: for all  $T_i \in T$  do
5:   for all  $term_j \in T_i$  do
6:     if ( $term_j \in SentiWordNet3.0$ ) then:  $score_{ij} = SWN(term_j) \times w_{ij}$ 
7:     else:  $score_{ij} = 0$ 
8:     end if
9:     if ( $score_{ij} > 0$ ) then: Pos-Count = Pos-Count +  $w_{ij}$ , Pos-Score = Pos-Score +
       $score_{ij}$ 
10:    else if ( $score_{ij} < 0$ ) then Neg-Count = Neg-Count +  $w_{ij}$ , Neg-Score = Neg-
      Score +  $score_{ij}$ 
11:    else if ( $score_{ij} == 0$ ) then: Do Nothing
12:    end if
13:  end for
14:  if (Pos-Score > Neg-Score) then:  $a_i = Positive$ 
15:  else if (Neg-Score > Pos-Score) then:  $a_i = Negative$ 
16:  else if (Pos-Count == 0  $\wedge$  Neg-Count == 0) then:  $a_i = Objective$ 
17:  else if (Pos-Score = Neg-Score) then:
18:    if (Pos-Count > Neg-Count) then:  $a_i = Positive$ 
19:    else if (Neg-Count > Pos-Count) then:  $a_i = Negative$ 
20:    else if (Pos-Count = Neg-Count) then:  $a_i = Neutral$ 
21:    end if
22:  end if
23: end for

```

3.3. Link identification and labelling

As previously mentioned, links between nodes pairs (representing the corresponding patients' reviews) are created when two nodes (reviews) are judged to be semantically (lexically) similar. The similarity between two feature vectors can be computed using a variety of metrics, including the Jaccard, Manhattan, and Euclidean distances [53]. The cosine similarity measure was chosen for the work proposed in

this study due to its widespread acceptance and usage. The cosine similarity measure between two documents (reviews) d_i and d_j can be calculated as $CosSim(d_i, d_j) = \frac{d_i \times d_j}{|d_i| \times |d_j|} = \frac{\sum_{k=1}^{k=z} w_{ik} \times w_{jk}}{\sqrt{\sum_{k=1}^{k=z} w_{ik}^2 \times \sum_{k=1}^{k=z} w_{jk}^2}}$. The range of values

for cosine similarity is 0 to 1. A value of 1 denotes the full similarity between the two documents under evaluation, while a value of 0 denotes their complete lack of similarity. An affinity or similarity matrix (a triangular matrix) is constructed in order to determine similarities between all patient review (node) pairs in the context of the PPIs graph generation framework proposed in this paper. After then, the existence of links between nodes is ascertained using this affinity matrix. If the similarity value between two nodes exceeds the average of all pair-wise similarities, then the two nodes are said to be linked according to the proposed framework. As labels, *support* and *oppose* are mutually used for labelling all individual links. When a pair of two linked nodes have the same attitude (homogeneous), the label *support* is applied and when their attitudes differ (heterogeneous), the label *oppose* is applied. Algorithm 2 explains the process of determining graph links and their associate labels. A list of links (L) is the output, and the collection of feature vectors (S) is the input for the proposed algorithm. Every item in L is in the form: (*start*, *label*, *end*), where *start* and *end* are the starting node ($l_i.start$) and ending node ($l_i.end$) of each link l_i , respectively. The notation $l_i.label$ is used to denote the label associated with a specific link l_i . The affinity matrix is computed in lines 3 through 6 of Algorithm 2, and it is subsequently analyzed in lines 7 through 10 to determine whether links exist. In lines 11 through 15, the labels of individual links are determined.

Algorithm 2. Link establishing and labelling

```

1: INPUT:  $S \subset BOW2$ 
2: OUTPUT: Link labels  $L = \{l_1, l_2, \dots, l_z\}$ 
3: Initialise  $Affinity[z][z]$ 
4: for all review pairs  $\langle s_i, s_{i'} \rangle \in S \wedge i < i'$  do
5:    $Affinity[i][i'] = CosSim(s_i, s_{i'})$ 
6: end for
7: for all components  $Affinity[i][i']$  do
8:   if ( $Affinity[i][i'] > average\_similarity$ ) then: add link  $l_i$  to  $L$ 
9:   end if
10: end for
11: for all  $l_i \in L$  do
12:   if ( $l_i.start == l_i.end$ ) then:  $l_i.label = support$ 
13:   else  $l_i.label = oppose$ 
14:   end if
15: end for

```

3.4. PPIs graph generation

PPIs graph generation is conducted in the last phase of the proposed framework. Using the resulted outcomes from Algorithm 1 and Algorithm 2, the graph generation process is carried out. However, the resulting output can be visualized using any appropriate graph sketching tool. With respect to the research work described in this paper, the authors visualized the extracted PPIs graph data using NetDraw [54] visualizing software tool. One of the evaluated medications from our DPR dataset, accutane (isotretinoin), is used to demonstrate the process supported by the PPIs graph generation framework. When the proposed framework is applied to this drug, the graph shown in Figure 2 is produced. Each patient in Figure 2 is represented by a node that has their age and gender labelled on it. Green node denotes a patient with positive attitude (i.e. moderate side effect) and red node denotes a patient with negative attitude (i.e. severe side effect), from the patient's perspective on the drug under consideration. For a pair of two linked nodes representing two patients, the thickness of the link between them indicates how comparable their semantic content is, which is determined by adding up all the words (terms) in both reviews that appear to be about the same topic and have non-zero weights for TF-IDF (see [55]–[58]). Green links are those that provide support (are in favor) and red links are those that provide opposition (are against).

4. RESULTS AND DISCUSSION

The absence of ground truth data is one of the difficulties in creating PPIs graphs. These graphs can be created naively by hand (manually), but doing so still involves subjectivity and significant resources to the extent that considerable benchmark data cannot be created easily and this process is time consuming. We assessed the attitudes extracted using SentiWordNet 3.0 with the patients' known attitudes, which is defined by how the patient at last ranked each drug (side effect score), in order to assess the PPIs graph extraction framework. There is a ten-point scale: 10 for the best, 1 for the worst. As a result, we are forced to make the assumption that the opinions (attitudes) expressed by the patients in their reviews accurately reflect the patients' opinions of how they will rate each medication. Therefore, it is expected of patients to never change

their opinions while considering a drug. Additionally, the numerical values of individual attitudes determined by Algorithm 1 had to be overlooked for evaluation purposes.

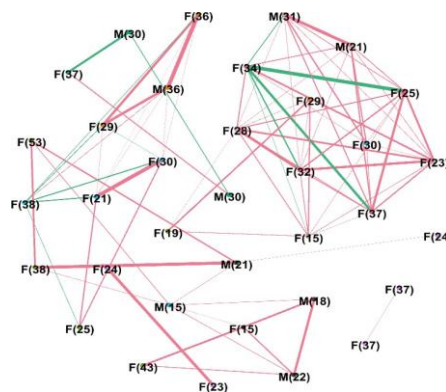


Figure 2. PPIs graph generated from druglib.com patients reviews about accutane drug using the proposed framework

The following commonly used machine prediction performance metrics are utilized: sensitivity (true positive recognition rate)= TP/P , specificity (true negative recognition rate)= TN/N , accuracy (the ratio of correct classification over all classifications)= $(TP+TN)/all$, error rate (the ratio of incorrect classification over all classifications)= $(FP+FN)/all=1- accuracy$, positive predictive value= $TP/(TP+FP)$ and negative predictive value= $TN/(TN+FN)$. The performance of the proposed framework in the context of the chosen measures is displayed in a tabular form containing the machine prediction results. With respect to assessing the propose framework, we considered its performance in classifying individual attitudes, successfully, as moderate side effect (positive attitude) or severe side effect (negative attitude). Table 1 provides the confusion matrix and Table 2 presents the corresponding evaluation using the previously mentioned measures. Based on the results in Table 2, the values demonstrates that the proposed framework adopting sentiment lexicon performs good for prediction attitudes and obtains promising outcomes.

Table 1. Confusion matrix for attitude prediction using sentiment analysis

	SentiWordNet 3.0
True positive (TP)	324
False negative (FN)	279
True negative (TN)	1,272
False positive (FP)	546
Total	2,421

Table 2. Evaluation results for attitude prediction using sentiment analysis

	SentiWordNet 3.0
Sensitivity	0.54
Specificity	0.70
Accuracy	0.66
Error rate	0.34
Positive predictive value	0.37
Negative predictive value	0.82

Examining the results in Table 2 reveals that the framework performs poorly at recognizing positive attitudes (moderate side effects) than at recognizing the negative attitudes (severe side effects). This is manifested in positive predictive value (0.37) and negative predictive value (0.82). We argue that, this is because the text in the side effect part of the individual patients’ reviews normally uses excessively derogatory and domain-specific (dedicated) jargon that is an expected feature of the side effect part of a patient review. This issue at hand can be resolved by utilizing similar data (labelled corpus) that we extracted our test dataset from to create a dedicated sentiment lexicon (using words co-occurrences) for medical language, which can be used as the foundation for the PPIs graph extraction framework. The true positive

recognition rate (sensitivity) reported measurement is 0.54 and true negative recognition rate (specificity) measurement is 0.70. The total average accuracy measurement, which is independent of class priors, reported is 0.66 as shown in Table 2.

5. CONCLUSION AND FUTURE EXTENSION

In the research work described in this paper, the authors presented a framework for extracting the required information about nodes, links and the required labels in order to creating the corresponding PPIs graph from textual patient's review transcripts (focusing on drug side effects part of individual patients' reviews) that have been submitted by patients and caregivers published on-line in HTML format (at www.druglib.com). Using sentiment analysis techniques to extract PPIs graphs will enable the visualization of the high-level structure of these disagreements graphically in order to gain a deeper insight into the information embedded in an abundant amount of these textual representation for the corresponding patients' reviews, which is the goal of the research described. The framework's functionality is demonstrated and assessed using 255 drugs data from the www.druglib.com website. The dataset included 3,763 patients' reviews linked to 255 different drugs. According to the promising results achieved, it is possible to: (i) use inter-document similarity to capture the PPIs structure, which represents patients as nodes; (ii) use sentiment analysis techniques, adopting SentiWordNet 3.0, to discover patients' attitudes. However, it may be necessary to develop specialized medical lexicons to increase attitude prediction overall accuracy, or machine learning techniques may be employed for the same purpose. The use of machine learning techniques more especially, classification techniques instead of lexicon-based sentiment analysis will be the primary focus of future research to improve the performance of prediction of the patients' attitudes from patients' reviews. Expanding the instances (examples) of our DPR collection is another goal. The authors plan, in the long run, to concentrate on mining the PPIs graphs that are produced in an effort to predict the effectiveness and side effects of drugs by utilizing the information (interesting patterns) embedded in the structure of these graphs.

REFERENCES




- [1] A. Asmi and T. Ishaya, "A framework for automated corpus generation for semantic sentiment analysis," in *Lecture Notes in Engineering and Computer Science*, pp. 436-444, 2012.
- [2] S. Grijzenhout, V. Jijkoun, and M. Marx, "Opinion mining in dutch hansards," *Proceedings Workshop from Text to Political Positions*, 2010.
- [3] Z. Salah, F. Coenen, and D. Grossi, "Generating domain-specific sentiment lexicons for opinion mining," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pp. 13-24, 2013. doi: 10.1007/978-3-642-53914-5_2.
- [4] A. Esuli and F. Sebastiani, "SENTIWORDNET: a publicly available lexical resource for opinion mining," in *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006*, pp. 417-422, 2006.
- [5] N. A. Ebrahim, M. Fathian, and M. R. Gholamian, "Sentiment classification of online product reviews using product features," *International Journal of Information Processing and Management*, vol. 3, no. 3, pp. 242-245, 2012, doi: 10.4156/ijipm.vol3.issue3.4.
- [6] G. Demiroz, B. Yanikoglu, D. Tapucu, and Y. Saygin, "Learning domain-specific polarity lexicons," in *Proceedings - 12th IEEE International Conference on Data Mining Workshops, ICDMW 2012*, pp. 674-679, 2012. doi: 10.1109/ICDMW.2012.120.
- [7] K. Denecke, "Using SentiWordNet for multilingual sentiment analysis," in *Proceedings - International Conference on Data Engineering*, pp. 507-512, 2008. doi: 10.1109/ICDEW.2008.4498370.
- [8] K. Denecke, "Are SentiWordNet scores suited for multi-domain sentiment classification?," in *4th International Conference on Digital Information Management, ICDIM 2009*, pp. 1-6, 2009. doi: 10.1109/ICDIM.2009.5356764.
- [9] J. Martineau and T. Finin, "Delta TFIDF: an improved feature space for sentiment analysis," in *Proceedings of the 3rd International AAAI Conference on Weblogs and Social Media, ICWSM 2009*, pp. 258-261, 2009. doi: 10.1609/icwsm.v3i1.13979.
- [10] A. Montejo-Raez, E. Martinez-Cámara, M. T. Martín-Valdivia, and L. A. U. na-López, "RandomWalk weighting over SentiWordNet for sentiment polarity detection on Twitter," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 3-10, 2012.
- [11] B. Ohana and B. Tierney, "Sentiment classification of reviews using SentiWordNet," *9th. IT & T Conference*, 2009.
- [12] G. Paltoglou and M. Thelwall, "Twitter, MySpace, Digg: unsupervised sentiment analysis in social media," *ACM Transactions on Intelligent Systems and Technology*, vol. 3, no. 4, pp. 1-19, 2012, doi: 10.1145/2337542.2337551.
- [13] R. Prabowo and M. Thelwall, "Sentiment analysis: a combined approach," *Journal of Informetrics*, vol. 3, no. 2, pp. 143-157, 2009, doi: 10.1016/j.joi.2009.01.003.
- [14] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *HLT/EMNLP 2005 - Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pp. 347-354, 2005. doi: 10.3115/1220575.1220619.
- [15] M. Herland, T. M. Khoshgoftaar, and R. Wald, "A review of data mining using big data in health informatics," *Journal of Big Data*, vol. 1, no. 1, pp.1-35, 2014, doi: 10.1186/2196-1115-1-2.
- [16] B. Pourebrahim and M. Keyvanpour, "Adverse drug reaction detection using data mining techniques: a review article," in *2020 10th International Conference on Computer and Knowledge Engineering, ICCKE 2020*, pp. 118-123, 2020. doi: 10.1109/ICCKE50421.2020.9303709.
- [17] A. Sarker et al., "Utilizing social media data for pharmacovigilance: a review," *Journal of Biomedical Informatics*, vol. 54, pp. 202-212, 2015. doi: 10.1016/j.jbi.2015.02.004.
- [18] J. Lardon et al., "Adverse drug reaction identification and extraction in social media: a scoping review," *Journal of Medical Internet Research*, vol. 17, no. 7, p.e171, 2015. doi: 10.2196/jmir.4304.

- [19] R. Sloane, O. Osanlou, D. Lewis, D. Bollegala, S. Maskell, and M. Pirmohamed, "Social media and pharmacovigilance: a review of the opportunities and challenges," *British Journal of Clinical Pharmacology*, vol. 80, no. 4, pp. 910-920, 2015, doi: 10.1111/bcp.12717.
- [20] R. B. Correia, L. Li, and L. M. Rocha, "Monitoring potential drug interactions and reactions via network analysis of instagram user timelines," in *Pacific Symposium on Biocomputing*, pp. 492-503, 2016. doi: 10.1142/9789814749411_0045.
- [21] R. Kiguba, S. Olsson, and C. Waitt, "Pharmacovigilance in low- and middle-income countries: a review with particular focus on Africa," *British Journal of Clinical Pharmacology*, vol. 89, no. 2, pp. 491-509, 2023. doi: 10.1111/bcp.15193.
- [22] A. Sarker and G. Gonzalez, "Portable automatic text classification for adverse drug reaction detection via multi-corpus training," *Journal of Biomedical Informatics*, vol. 53, pp. 196-207, 2015, doi: 10.1016/j.jbi.2014.11.002.
- [23] K. Jiang and Y. Zheng, "Mining Twitter data for potential drug effects," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pp. 434-443, 2013. doi: 10.1007/978-3-642-53914-5_37.
- [24] M. J. Paul and M. Dredze, "You are what you tweet: analyzing twitter for public health," in *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media, ICWSM 2011*, pp. 265-272, 2011. doi: 10.1609/icwsm.v5i1.14137.
- [25] G. Chowell and H. Nishiura, "Transmission dynamics and control of Ebola virus disease (EVD): A review," *BMC Medicine*, vol. 12, no. 1, pp. 1-17, 2014. doi: 10.1186/s12916-014-0196-0.
- [26] P. Carbonell, M. A. Mayer, and A. Bravo, "Exploring brand-name drug mentions on Twitter for pharmacovigilance," in *Studies in Health Technology and Informatics*, pp. 55-59, 2015. doi: 10.3233/978-1-61499-512-8-55.
- [27] F. Gräßer, H. Malberg, S. Kallumadi, and S. Zaunseder, "Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning," in *ACM International Conference Proceeding Series*, pp. 121-125, 2018. doi: 10.1145/3194658.3194677.
- [28] B. Xu et al., "Incorporating user generated content for drug drug interaction extraction based on full attention mechanism," *IEEE Transactions on Nanobioscience*, vol. 18, no. 3, pp. 360-367, 2019, doi: 10.1109/TNB.2019.2919188.
- [29] J. Hadzi-Puric and J. Grmusa, "Automatic drug adverse reaction discovery from parenting websites using disproportionality methods," in *Proceedings of the 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2012*, pp. 792-797, 2012. doi: 10.1109/ASONAM.2012.144.
- [30] I. Segura-Bedmar, M. Crespo, C. de Pablo-Sánchez, and P. Martínez, "Resolving anaphoras for the extraction of drug-drug interactions in pharmacological documents," *BMC Bioinformatics*, vol. 11, no. 2, pp. 1-9, 2010, doi: 10.1186/1471-2105-11-S2-S1.
- [31] H. Y. Wu, C. W. Chiang, and L. Li, "Text mining for drug-drug interaction," *Methods in Molecular Biology*, vol. 1159, pp. 47-75, 2014, doi: 10.1007/978-1-4939-0709-0_4.
- [32] S. Sohn, J. P. A. Kocher, C. G. Chute, and G. K. Savova, "Drug side effect extraction from clinical narratives of psychiatry and psychology patients," *Journal of the American Medical Informatics Association*, vol. 18, no. SUPPL. 1, pp. i144-i149, 2011, doi: 10.1136/amiajnl-2011-000351.
- [33] R. Xu and Q. Q. Wang, "Comparing a knowledge-driven approach to a supervised machine learning approach in large-scale extraction of drug-side effect relationships from free-text biomedical literature," *BMC Bioinformatics*, vol. 16, no. 5, pp. 1-8, 2015, doi: 10.1186/1471-2105-16-S5-S6.
- [34] S. Stieglitz, M. Mirbabaie, B. Ross, and C. Neuberger, "Social media analytics – challenges in topic discovery, data collection, and data preparation," *International Journal of Information Management*, vol. 39, pp. 156-168, 2018, doi: 10.1016/j.ijinfomgt.2017.12.002.
- [35] Y. Wang and N. Hajli, "Exploring the path to big data analytics success in healthcare," *Journal of Business Research*, vol. 70, pp. 287-299, 2017, doi: 10.1016/j.jbusres.2016.08.002.
- [36] J. D. Duke, X. Li, and S. J. Grannis, "Data visualization speeds review of potential adverse drug events in patients on multiple medications," *Journal of Biomedical Informatics*, vol. 43, no. 2, pp. 326-331, 2010, doi: 10.1016/j.jbi.2009.12.001.
- [37] S. Li, C. H. Yu, Y. Wang, and Y. Babu, "Exploring adverse drug reactions of diabetes medicine using social media analytics and interactive visualizations," *International Journal of Information Management*, vol. 48, pp. 228-237, 2019, doi: 10.1016/j.ijinfomgt.2018.12.007.
- [38] T. I. Oprea et al., "Associating drugs, targets and clinical outcomes into an integrated network affords a new platform for computer-aided drug repurposing," in *Molecular Informatics*, vol. 30, no. 2-3, pp.100-111, 2011. doi: 10.1002/minf.201100023.
- [39] L. Yang and P. Agarwal, "Systematic drug repositioning based on clinical side-effects," *PLoS ONE*, vol. 6, no. 12, 2011, doi: 10.1371/journal.pone.0028025.
- [40] M. Thelwall and K. Buckley, "Topic-based sentiment analysis for the social web: the role of mood and issue-related words," *Journal of the American Society for Information Science and Technology*, vol. 64, no. 8, pp. 1608-1617, 2013, doi: 10.1002/asi.22872.
- [41] S. Bratus, A. Rumshisky, A. Khrabrov, R. Magar, and P. Thompson, "Domain-specific entity extraction from noisy, unstructured data using ontology-guided search," *International Journal on Document Analysis and Recognition*, vol. 14, no. 2, pp. 201-211, 2011, doi: 10.1007/s10032-011-0149-5.
- [42] Y. Choi and C. Cardie, "Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification," in *EMNLP 2009 - Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: A Meeting of SIGDAT, a Special Interest Group of ACL, Held in Conjunction with ACL-IJCNLP 2009*, pp. 590-598, 2009. doi: 10.3115/1699571.1699590.
- [43] A. Weichselbraun, S. Gindl, and A. Scharl, "Using games with a purpose and bootstrapping to create domain-specific sentiment lexicons," in *International Conference on Information and Knowledge Management, Proceedings*, pp. 1053-1060, 2011. doi: 10.1145/2063576.2063729.
- [44] S. S. Tan and J. C. Na, "Expanding sentiment lexicon with multi-word terms for domain-specific sentiment analysis," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pp. 285-296, 2016. doi: 10.1007/978-3-319-49304-6_34.
- [45] J. Zhang and Q. Peng, "Constructing Chinese domain lexicon with improved entropy formula for sentiment analysis," in *2012 IEEE International Conference on Information and Automation, ICIA 2012*, pp. 850-855, 2012. doi: 10.1109/ICInfA.2012.6246900.
- [46] Y. Wilks and M. Stevenson, "The grammar of sense: using part-of-speech tags as a first step in semantic disambiguation," *Natural Language Engineering*, vol. 4, no. 2, pp. 135-143, 1998, doi: 10.1017/S1351324998001946.
- [47] H. Chim and X. Deng, "Efficient phrase-based document similarity for clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 9, pp. 1217-1229, 2008, doi: 10.1109/TKDE.2008.50.




- [48] S. Hariharan and R. Srinivasan, "A comparison of similarity measures for text documents," *Journal of Information and Knowledge Management*, vol. 7, no. 1, pp.1-8, 2008, doi: 10.1142/S0219649208001889.
- [49] S. Poomagal and T. Hamsapriya, "K-means for search results clustering using URL and Tag contents," in *Proceedings of 2011 International Conference on Process Automation, Control and Computing, PACC 2011*, pp. 1-7, 2011, doi: 10.1109/PACC.2011.5978906.
- [50] A. Amine, Z. Elberrichi, and M. Simonet, "Evaluation of text clustering methods using WordNet," *International Arab Journal of Information Technology*, vol. 7, no. 4, pp. 349-357, 2010.
- [51] A. Kuhn, S. Ducasse, and T. Gırba, "Semantic clustering: identifying topics in source code," *Information and Software Technology*, vol. 49, no. 3, pp. 230-243, 2007, doi: 10.1016/j.infsof.2006.10.017.
- [52] H. M. Li, C. X. Sun, and K. J. Wang, "Clustering web search results using conceptual grouping," in *Proceedings of the 2009 International Conference on Machine Learning and Cybernetics*, pp. 1499-1503, 2009, doi: 10.1109/ICMLC.2009.5212322.
- [53] A. Madylova and Ş. G. Öğüdücü, "Comparison of similarity measures for clustering Turkish documents," *Intelligent Data Analysis*, vol. 13, no. 5, pp. 815-832, 2009, doi: 10.3233/IDA-2009-0394.
- [54] J. Fan, X. Xu, and L. Zhao, "A bibliometric analysis of the theme trends and knowledge structures of pulmonary embolism from 2017 to 2021," *Frontiers in Medicine*, vol. 10, p.1052928, 2023, doi: 10.3389/fmed.2023.1052928.
- [55] O. Iparraguirre-Villanueva et al., "Search and classify topics in a corpus of text using the latent dirichlet allocation model," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 30, no. 1, pp. 246-256, 2023, doi: 10.11591/ijeecs.v30.i1.pp246-256.
- [56] A. C. Alhadi, A. Deraman, M. M. A. Jalil, W. N. J. W. Yussof, and R. Mohemad, "A computational analysis of short sentences based on ensemble similarity model," *International Journal of Electrical and Computer Engineering*, vol. 9, no. 6, pp. 5386-5394, 2019, doi: 10.11591/ijece.v9i6.pp5386-5394.
- [57] M. Hammad, M. Al-Smadi, Q. B. Baker, and S. A. Al-Zboon, "Using deep learning models for learning semantic text similarity of Arabic questions," *International Journal of Electrical and Computer Engineering*, vol. 11, no. 4, pp. 3519-3528, 2021, doi: 10.11591/ijece.v11i4.pp3519-3528.
- [58] A. Srivastav and S. Prajapat, "Text similarity algorithms to determine Indian penal code sections for offence report," *IAES International Journal of Artificial Intelligence*, vol. 11, no. 1, pp. 34-40, 2022, doi: 10.11591/ijai.v11.i1.pp34-40.

BIOGRAPHIES OF AUTHORS






Zaher Salah    received his Ph.D. degree in Computer Science from the University of Liverpool, UK, in 2014, his MSc degree in Computer Science from Yarmouk University, Jordan, in 2004, and his B.Sc. degree in Computer Science from University of Jordan, Jordan, in 2001. He is currently an Associate Professor in the Information Technology Department of The Hashemite University, Zarqa, Jordan. His research interests include machine learning, cyber security, information retrieval, opinion mining, sentiment analysis, biometrics, digital image and analysis, and pattern recognition. He can be contacted at email: zaher@hu.edu.jo.






Esraa Elsoud    received the B.Sc. degree in Electrical Engineering from The Hashemite University, Jordan, in 2013 and M.Sc. in Cyber Security from The Hashemite University in 2023. Eng. Esraa's current research interests include cyber security, machine learning, big data and mobile network. She is currently a Lecturer in Zarqa University, Zarqa, Jordan. She can be contacted at email: eabuelsoud@zu.edu.jo.






Kamal Salah    received the B.Sc. degree in Physics from Yarmouk University, Jordan in 2003, his M.Sc. degree in Applied Physics (Experimental Atomic and Molecular Physics) from The Hashemite University Jordan in 2007. He is currently a Lecturer in the Deanship of Preparatory Year and Supporting Studies of the Imam Abdulrahman Bin Faisal University, P.O Box 1982, Dammam, Saudi Arabia. He can be contacted at email: kisalah@iau.edu.sa.






Waleed T. Al-Sitt    received his Ph.D. degree in Electrical Engineering and Electronics from the University of Liverpool, UK, in 2015, his M.Sc. degree in Electrical and Computer Engineering from New York Institute of Technology (NYIT), Jordan, in 2008, and his BSc degree in Computer Engineering from Mu'tah University, Jordan, in 2006. He is currently an Associate Professor in the Department of Computer Engineering, Mutah University, Al-Karak, Jordan; Higher Colleges of Technology, Dubai, UAE. His research interests include machine learning, cyber security, sentiment analysis, networking, and pattern recognition. He can be contacted at email: w_sitt@mutah.edu.jo.



Manal Maaya'a    received the B.Sc. degree in Computer Science and Applications from The Hashemite University, Jordan, in 2009 and M.Sc. in Computer Information Systems from The University of Jordan in 2015. She current research interests include artificial intelligence, data analysis, machine learning, natural language processing and computer linguistics. She can be contacted at email: ManalS@hu.edu.jo.



Ahmad Al Khawaldeh    received his Ph.D. degree in Chemistry from The University of Jordan, Jordan, in 2018, his M.Sc. degree in Chemistry from The University of Jordan, Jordan, in 2014, and his B.Sc. degree in Chemistry from The Hashemite University, Jordan, in 2009. He is currently an Assistant Professor in Al-Balqa Applied University, Zarqa University College, Zarqa, Jordan. He can be contacted at email: ahmad.khawaldeh@bau.edu.jo.