An Automated System for Internet Pharmacy Verification

Alberto Cordioli
Prometeia
alberto.cordioli@prometeia.com

Themis Palpanas
Paris Descartes University
themis@mi.parisdescartes.fr

ABSTRACT

In the past years, we have witnessed an explosion of web applications, and in particular of electronic commerce websites. This has led to unquestionable benefits for both producers and consumers of goods. On the other hand, however, untrusted companies have the opportunity to bypass checks and regulations imposed by relevant bodies. This problem is prevalent in the context of online commerce of pharmaceutical products, where it is essential that such products are safe, of good quality, and only used with a proper prescription. In this work, we study the problem of internet pharmacy verification. To this effect, we build a classifier, able to find patterns and predict the class of unseen data. Moreover, we devise algorithms that give a trust score to each pharmacy, in order to have a legitimacy indicator usable by human reviewers. We experimentally evaluate the proposed approach with real data coming from two different time periods. The results demonstrate the effectiveness of our approach, as well as the potential of using similar techniques for automatically checking regulation compliance in electronic commerce.

1 INTRODUCTION

The growth of web-related technologies, and in particular e-commerce, has offered companies the opportunity to increase their own business, selling directly their products and goods to customers within and across borders. Even though this has led to unquestionable benefits for customers, untrusted companies can also access the market and sell products, for which it is not always possible to assess the quality.

The above problem is even more prominent when we are dealing with pharmaceutical products. Online sale of counterfeit drugs has become an important problem, with studies by the World Health Organization (WHO) showing that in more than 50% of the cases, drugs sold by websites that conceal their physical address are counterfeit[1]. The WHO argues that counterfeiting occurs both with branded and with generic products, and counterfeit drugs may include the correct ingredients but fake packaging, the wrong ingredients, insufficient active ingredients, or no active ingredients whatsoever[2]. Evidently, counterfeit drugs represent an enormous public health challenge [26], and also a major illicit economic activity, with an estimated $75 billion market for 2010[3].

Moreover, the mere task of distinguishing a *legitimate* from an *illegitimate* online pharmacy is rather challenging. This is true for domain experts, and is often times impossible to do for simple users, especially since *illegitimate* pharmacies and drugs are designed to look like *legitimate* ones (including the packaging of drugs, and the drugs themselves, which are usually identical to the original ones).

Figure 1 shows the front webpage of two online pharmacies, only one of which is *legitimate*. Evidently, a simple observation of the two webpages is not enough to reveal which one[4]. Hence, there is a pressing need for assessing the quality of pharmaceutical products sold online, and a major step in this direction is to assess the trustiness of online pharmacies, which is the focus of this study.

There are different factors that make a pharmacy *illegitimate*. In U.S.A. (as well as in many other countries), an online pharmacy must satisfy regulations and meet strict requirements. The requirements that are most frequently violated in the U.S. are, for example, the selling of products without prescriptions and the selling of drugs that are not "FDA-Approved"[5] [21]. Evidently, checking these factors is not an easy task, especially for people that do not have any kind of competence and knowledge in this field, such as the normal consumers.

It is for this reason that specialized companies have made the verification of health-related websites their own business. LegitScript[6], for example, offers an internet pharmacy verification service and collaborates with the major search engines (e.g., Google, Bing) in order to enforce policies against *illegitimate* online pharmacies, which can be as much as 90% of the total number of online pharmacies [21].

The process of classifying health-related websites into *legitimate* and *illegitimate* pharmacies is currently mostly *manual*, and requires a great investment of time and human resources. The increasingly large number of online pharmacies, and correspondingly large number of *illegitimate* online pharmacies, leads to the necessity of streamlining the review process with a system capable of automatically giving a trust score to online pharmacies. In this manner the system can assist the human reviewers, relieving them of some tedious and time-consuming tasks.

In this paper, we propose the first systematic approach to the aforementioned problem, using techniques that are based on both text and network features, and we describe a system capable of verifying internet pharmacies. We have made all our code publicly available [7]. Even though previous studies have discussed this problem (e.g., [3, 23, 28]), they did not provide algorithmic solutions for it.

The contributions we make in this paper are as follows.

- We provide the first systematic study for the problem of internet pharmacy verification and formalize two sub-problems: (a) classification of online pharmacies into *legitimate* and *illegitimate*; and (b) ranking online pharmacies according to a *legitimacy* score.
- We study and evaluate indicators that can distinguish between *legitimate* and *illegitimate* pharmacies. We propose

[1] http://who.int/bulletin/volumes/88/4/10-020410/en/

[2] http://apps.who.int/iris/bitstream/10665/65892/1/WHO_EDM_QSM_99.1.pdf

[3] The complete article can be found at: http://www.usatoday.com/money/industries/health/drugs/story/2011-10-09/cnbc-drugs/50690880/1

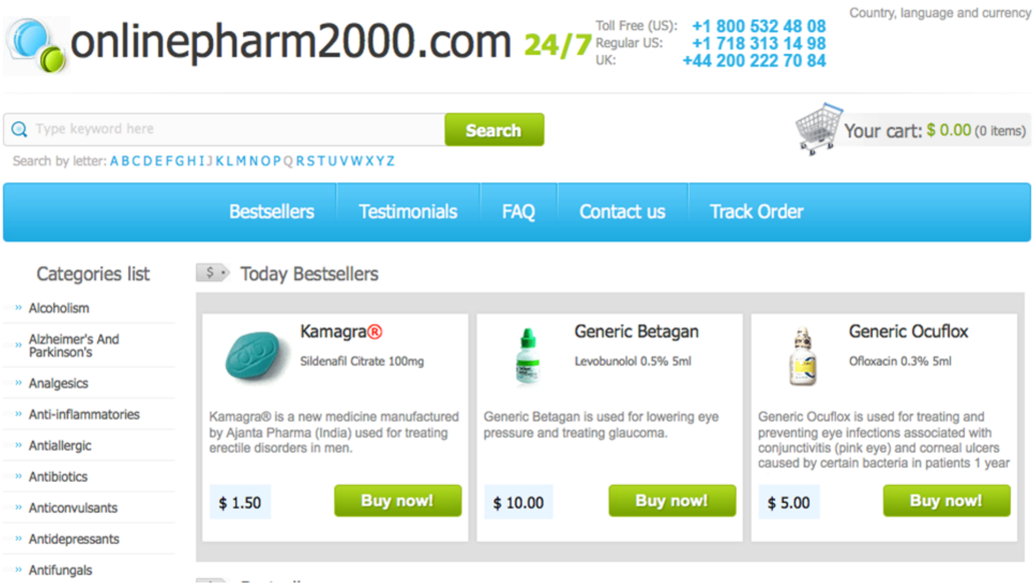
[4] The pharmacy depicted in Figure 1a is *illegitimate*, while the one depicted in Figure 1b is *legitimate*.

[5] If a drug is FDA-Approved it means that specific tests have been conducted to prove the quality of the product.

[6] http://www.legitscript.com

[7] https://sites.google.com/view/acolfplg/home

© 2018 Copyright held by the owner/author(s). Published in Proceedings of the 21st International Conference on Extending Database Technology (EDBT), March 26-29, 2018, ISBN 978-3-89318-078-3 on OpenProceedings.org.

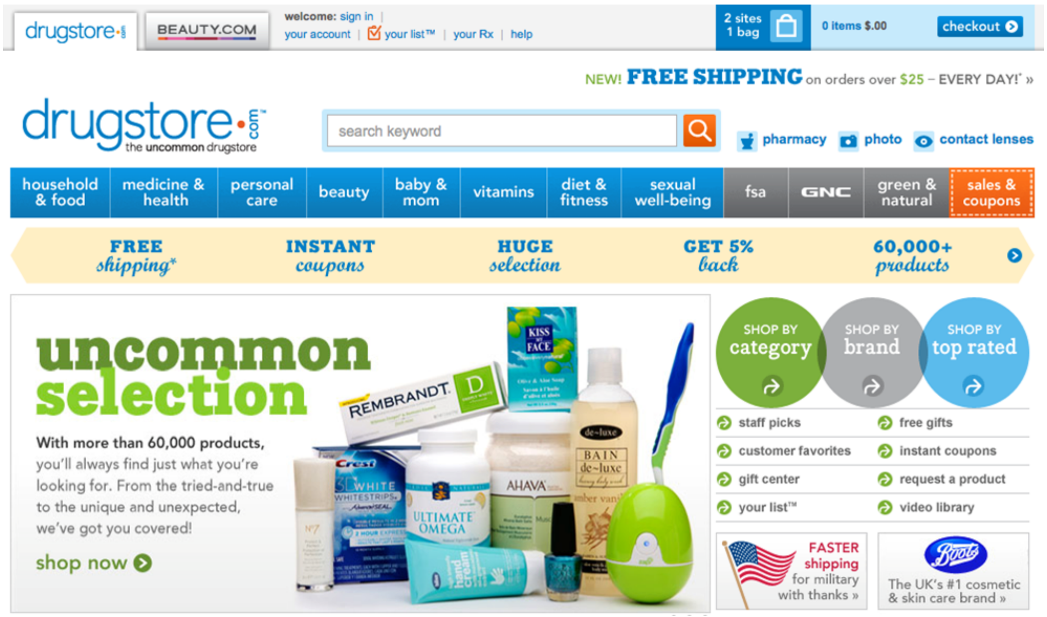


Kamagra® is a new medicine manufactured by Ajanta Pharma (India) used for treating erectile disorders in men.

Generic Betagan is used for lowering eye pressure and treating glaucoma.

Generic Ocuflox is used for treating and preventing eye infections associated with conjunctivitis (pink eye) and corneal ulcers caused by certain bacteria in patients 1 year

(a) front webpage of online pharmacy 1



With more than 60,000 products, you'll always find just what you're looking for. From the tried-and-true to the unique and unexpected, we've got you covered!

(b) front webpage of online pharmacy 2

Figure 1: Examples of two online pharmacies.

novel features that are based on both the text of the online pharmacy website, and its network structure, and integrate these features in our models.

- We describe effective and efficient solutions for the classification and ranking problems. The proposed solutions can automate to a large extent the process of online pharmacy verification.
- We experimentally validate our methods with the two largest real datasets used in the literature, comprising of almost 2500 *illegitimate* pharmacies and 200 *legitimate* pharmacies, crawled in two different time periods. The results demonstrate the accuracy of our approach and its practical value, and showcase the potential of similar techniques in relevant problems in e-commerce.

The rest of this paper is organized as follows. In Section 2, we elaborate on existing work. In Section 3, we provide some background material that is necessary for the discussion that follows, and we formally define the two problems we solve. We describe our proposed solution in Sections 4 and 5. We present the experimental results in Section 6, and finally, we conclude in Section 7 with a brief summary and some thoughts on future work.

2 RELATED WORK

2.1 Pharmacy Verification

Previous works have discussed the problem of online pharmacy verification [22], and advocated the need for studying this problem from the regulation and public-health points of view [27, 28]. A study in the area of medicinal drug commerce has shown that consumers should be able to reduce their risk by relying on trusted lists compiled by credited agencies [3], and by using common sense when examining packaging and pills, while other studies have explored the problem of identifying controversial drugs, by monitoring consumer opinion [33, 34].

Nevertheless, the major problem is that it is not possible to assess the quality of a drug sold online, at least until it has been purchased. And even at that time, determining if such medicine is safe or not, requires several analyses and competences, often not held by the normal consumer. The most promising solution is the one of using official lists of trusted online pharmacies. Nevertheless, this is a daunting task to complete manually, because of the sheer number of online pharmacies and the rate with which they appear (or disappear).

Some studies have focused on the problem of identifying features and signals that distinguish *legitimate* and *illegitimate* online pharmacies [23, 27, 28]. In [23] the authors performed a comparative analysis of website trust features applied to the case of online pharmacies, which showed that proven *legitimate* pharmacies use more extensively verification seals and have more instances of health content than *illegitimate* pharmacies. On the other hand, *illegitimate* pharmacies have fewer store presence features than *legitimate* pharmacies. In [22] the authors try to understand the reasons for the success of unlicensed online pharmacies. They discover that instead of directly competing with licensed pharmacies, unlicensed pharmacies often sell drugs that licensed pharmacies do not, or cannot sell.

Additional studies [27, 28] demonstrate that the problem of online pharmacy verification should be studied at two different levels: from the regulation point of view and from the public health point of view. The very same conclusion has been outlined another study as well [27], where the authors perform a review of the scientific literature and a study of several scientific and institutional databases. They showed that the phenomenon is continuing to spread, and in order to enhance the benefits and minimize the risks, a 2-level approach could be adopted: first, from a policy point of view implementing international level laws, and second on an individual consumer point of view. Note that none of studies mentioned above propose algorithmic solutions for tackling the problem at hand.

CADRE [36] is a cloud-assisted drug recommendation system, which can recommend users with related drugs, according to their symptoms. The system clusters drugs into several groups according to the functional description information, and designs a basic personalized drug recommendation based on user collaborative filtering. While this system can help consumers find alternative drugs for their symptoms, it can not be used for identifying *illegitimate* drugs, or *illegitimate* online pharmacies.

In this study, we show that another direction is possible and effective. We claim that relevant bodies (e.g., LEAs and private agencies in health-care system) could use state-of-the-art data mining techniques in order to find, isolate, and eventually shoutdown *illegitimate* online pharmacies. Our approach outlines a text classification process and a network trust algorithm in order to assess the *legitimacy* of internet pharmacies.

2.2 Text Classification and Network Trust Algorithms

Our problem is reminiscent to spam detection [5]. Even though the problems of internet pharmacy verification and web spam detection exhibit some similarities, they are very different in the definition of what is considered *legitimate*. In our case, we might require very specific knowledge in order to differentiate between *legitimate* and *illegitimate* examples. Moreover, the final scope of the two systems are different. In [5] the authors aim to improve search engines algorithms, while in this work the final users are domain-specific analysts working in the field of online pharmacy verification.

Text Classification (**TC**) is defined as the process, in which a document is automatically classified in one or more categories (classes) [1, 31]. In this process, a set of labeled data is used to train a classifier, which recognize patterns among instances of the same class. These patterns are then used to build a model able to classify unlabeled instances. TC have been used in many contexts, including language identification [6], message filtering (i.e., spam filtering) [2, 19, 25], hierarchical categorization of web pages [7, 11], and others. Recent studies have surveyed text classification algorithms [1], and also studied the behavior of classifiers in the presence of label noise [14, 24].

In the specific context of web content, the classification of web content in one or more classes has been studied by many researchers. In [17], the author analyzes the nature of web content and metadata in relation to requirements for text features, and presents a system for automatically classifying websites into industrial categories. The work presented in [11] explores the use of a hierarchical structure for classifying a large, heterogeneous collection of web content.

In [13], the authors compare three different text representation techniques, based on (character) N-Gram Graphs, the Term Vector model, the Character N-Grams model, and the N-Gram Graphs model, with respect to three different categories of documents: curated, semi-curated and raw documents. They show that each category calls for different classification settings with respect to the representation model; moreover they show that N-Gram Graphs model achieves higher performances on each of the three different categories analyzed. This is a versatile technique that we use in our work, and that we further discuss in the following sections.

Assessing the trustiness in a network of hosts or websites has become very important in the web context, where the number of web spam pages increases by the minute. In order to address this problem, some search engines have adopted trust algorithms to reduce the rank of such pages in query results. TrustRank [15] is a link analysis technique for semi-automatically separating useful webpages from spam. Starting from a seed of reputable web pages, TrustRank uses the underlying network structure to discover other pages that are likely to be legitimate.

In [20] the authors provide a variation of TrustRank algorithm, called Anti-TrustRank, where non-reputable web pages are selected as initial seed. A different algorithm, which is able to decrease the number of downloads of inauthentic files in a peer-to-peer file-sharing network, is presented in [18].

An important characteristic of our domain is that the two classes - *legitimate* and *illegitimate*- are strongly imbalanced: the number of *legitimate* examples represents only a small percentage of the total. The effect of skewed distribution has been studied in many aspects. In [35] the authors analyzed the effect of class distribution on classifier learning and showed that the naturally occurring class distribution is often not the best choice for learning. In [10, 29, 32] the effects of dataset distribution have been studied for two very well known classifiers: C4.5 and SVM. In our work, we compare the results obtained by training classifiers with the natural distribution and with resampling (both undersampling and oversampling).

3 PRELIMINARIES AND PROBLEM DEFINITION

In this section, we recap some concepts and techniques needed for the rest of the paper, and formally define the problems we solve.

3.1 Preliminaries

We first define the concept of an online pharmacy, and we point out the differences between *legitimate* and *illegitimate* online pharmacies. This is important, since the term *illegitimacy* may have different interpretations depending on the contexts.

An online pharmacy is a website that offers medical products for sale. In this context there are different levels of *illegitimacy* for a pharmacy, and such levels depend on certain features that, if present, contribute to make that pharmacy *illegitimate.* We can group *illegitimate* pharmacies in three big categories:

- online pharmacies that do not adhere to accepted standards of medicine and/or pharmacy practice, including standards of safety;
- online pharmacies that violate, appear to violate, encourage violation of, or are not in compliance with applicable national or regional laws or regulations;
- online pharmacies that engage in fraudulent or deceptive business practices.

The first two categories are self explanatory. They include pharmacies that represent a threat for the people's health, as they sell drugs that are not approved, or they are not in compliance with national or regional regulations. The third category is a more general class that includes those websites that scam people, stealing their data, or money (obviously, these websites are not only health-related). These three categories are not mutually exclusive, and a *illegitimate* pharmacy may belong to more than one, which is actually true for the majority of them.

Signals that make a pharmacy more likely to be *illegitimate* include concealing its physical address, being isolated from major trusted websites, as well having fewer store presence features, and fewer health-related text content than *legitimate* pharmacies [3, 28].

3.2 Problem Statement

We now formalize the problems that we solve in this study: first, classification of online pharmacies in *illegitimate* and *legitimate*; and second, ranking of online pharmacies according to their *legitimacy.*

We denote with $\mathcal{P}$ the set of all the pharmacy websites. Let's call $\mathcal{P}^+$ and $\mathcal{P}^-$ the *legitimate* and the *illegitimate* pharmacies, respectively, in $\mathcal{P}$. We also suppose to know the class of a subset of pharmacies $\mathcal{P}_0 \subseteq \mathcal{P}$, i.e., there exists an *oracle function* O that for all the $p \in \mathcal{P}_0$:

$$O(p) = \begin{cases} 1 & \text{if } p \in \mathcal{P}^+ \\ 0 & \text{if } p \in \mathcal{P}^- \end{cases} \tag{1}$$

Assuming that a human reviewer can evaluate if a website is *legitimate* or not, we can think of the oracle function O as an evaluation performed by a human reviewer. However, oracle invocations are expensive and we cannot call O for each pharmacy in $\mathcal{P}$, so we aim to find another function, T, such that for some model Π:

$$T(p) = \begin{cases} 1 & \text{if } \Pi \models p \in \mathcal{P}^+ \\ 0 & \text{if } \Pi \models p \in \mathcal{P}^- \end{cases} \tag{2}$$

We recall that the formal notion $\Pi \models p \in \mathcal{P}^+$ means that Π entails $p \in \mathcal{P}^+$, where Π is a model derived and built from the training set $\mathcal{P}_0$ with some learning algorithm.

Note that there exist infinite T derived by infinite Π, but we are interested in the one that maximizes some evaluation measure (e.g., number of correctly classified instances, recall, precision). We can now formalize the classification problem:

Problem 1 (Online Pharmacy Classification (OPC)). *Given a set of pharmacies $\mathcal{P}$ divided in two classes $\mathcal{P}^+$ and $\mathcal{P}^-$, a set of "known" pharmacies $\mathcal{P}_0 \subseteq \mathcal{P}$, and an evaluation measure φ, we seek a model Π_i such that $\forall p \in \mathcal{P}, T_i(p)$ maximizes φ.*

The second problem we are trying to solve is a ranking problem. We want to give a trust score to each pharmacy, which we could then use to produce an ordered list. More formally, we want to define a *totally ordered set*, where for each pair of pharmacies $p_1, p_2 \in \mathcal{P}$ it holds that $score(p_1) \leq score(p_2)$ or $score(p_2) \leq score(p_1)$. The score for each pharmacy is computed by combining different models.

Problem 2 (Online Pharmacy Ranking (OPR)). *Given a set of pharmacies $\mathcal{P}$ divided in two classes $\mathcal{P}^+$ and $\mathcal{P}^-$, a set of "known" pharmacies $\mathcal{P}_0 \subseteq \mathcal{P}$, and a list of models $\Pi_1, \ldots, \Pi_k$, we seek a totally ordered set $\langle \mathcal{P}, \leq \rangle$ such that, for each pair of elements $p_1, p_2 \in \mathcal{P}$, if $score(p_1) \leq score(p_2)$, then p_1 is "less legitimate" than p_2.*

We expect that the ordered list naturally divides the pharmacies $\mathcal{P}$ in two subsets (i.e., *legitimate* and *illegitimate* pharmacies), with all the elements of one subset at the top of this list, and all the elements of the other subset at the bottom. Without loss of generality, in the following we will focus on a *legitimacy* relation, which builds a list with *legitimate* examples at the top and *illegitimate* ones at the bottom.

In the following sections, we discuss how we solve the two problems formalized above, namely the classification and the ranking problem.

4 ONLINE PHARMACY CLASSIFICATION

Our classification algorithm is based on features that are relevant to both the text contained in the website of the online pharmacy and the web network structure around it.

4.1 Text Analysis

In order to reduce the dimensionality of the problem, it is common practice in Text Classification (TC) to use *preprocessing* and *summarization.*

Preprocessing: In the preprocessing step we remove the stop words found in the documents. In this way the most common words, which could adversely affect classification accuracy, are removed. To do so we rely on Apache Lucene[8] version 3.4.0. We also decided to not use stemming, as the text contains a lot

[8] http://lucene.apache.org/

of technical words and trademarks, and this technique causes undesirable side-effects.

Summarization: The process of transforming a set of web pages into a unique summary is called *summarization*. For each pharmacy, we merge the text content of all the pages crawled into a single document. This step produces documents that include a large number of terms; documents comprising of 160, 000 terms are not unusual. In our experiments, we evaluate the performance of the classifier models when we use the entire content of the document (all terms), as well as subsets of it. During this phase, we generate subsamples of the summary document considering a limited number of terms by randomly selecting 100, 250, 1000 and 2000 terms.

We are now ready to describe the core steps of our classification approach. Considering pharmacy websites as documents, and the two classes, *legitimate* and *illegitimate*, as mutually exclusive, we map our problem into the TC problem. We first convert each document (i.e., the text contained in an online pharmacy website) in a format suitable for the classification phase. We study two different such models, the Term Vector model [30], and the N-Gram Graphs model [12], both outlined below. Then, we use TC to classify unseen data relying on models built using the subset of labeled data $\mathcal{P}_0$. We employ different learning algorithms to build a two-class classifier, including Naïve Bayesian Multinomial (NBM), Support Vector Machine (SVM), and the C4.5 decision tree learning algorithm.

4.1.1 *Term Vector Model.* :

The Term Vector model is the most widely used text representation model in information retrieval, due to its high level of performance and scalability [30]. The model works as follows: given a set of documents $\mathcal{D}$, each document $d \in \mathcal{D}$ is represented as a vector of words $v_d = (v_1, v_2, \dots v_{|\mathcal{W}|})$, where $\mathcal{W}$ is the set of all the distinct words in $\mathcal{D}$. Each position v_i with $1 < i < |W|$ in the vector represents the presence, or the absence, of word w_i in document d.

There are many variants to represent d_i values, but the most popular is the TF-IDF approach, which takes into account the number of occurrences of a term in a document (term frequency), and its overall frequencies in the whole set of documents $\mathcal{D}$ (inverse document frequency).

4.1.2 *N-Gram Graphs.* :

The N-Gram Graph is a graph $\mathcal{G}(V, E)$ which has character n-grams as its vertices, while the edges connecting the n-grams indicate proximity of the corresponding vertex N-Grams [12]. The weights on the edges represent how much the two N-Gram are related (one way of setting these weights is by counting how many times two N-Grams co-occur within a sliding window in the text). The advantage of N-Gram Graphs is that they conserve the order of the characters' appearance in the original text, and hence are more stable than the standard Character N-Gram Model. The three measures characterizing an N-Gram Graph are: (i) the minimum n-gram rank L_{min}, (ii) the maximum n-gram rank L_{max}, and (iii) the minimum neighborhood distance D_{win}. In our experiments, we use $L_{min} = L_{max} = D_{win} = 4$ [13].

In order to use N-Gram Graphs, we first transform each document d in a N-Gram Graph (refer to Figure 2). For each class c, we build an N-Gram Graph derived from merging the individual graphs in c, and we compute the similarities between each document d and the class graph. We then use the following similarity measures, namely, Containment Similarity (CS), Size Similarity (SS), and Value Similarity (VS) in order to build a classifier able to predict the class of unseen data.

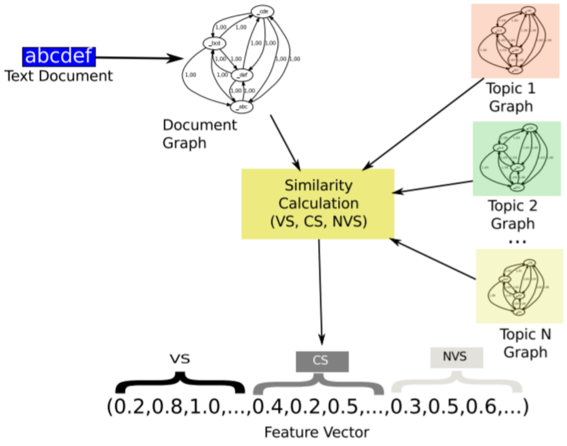


Figure 2: An overview of the classification process using N-Gram Graphs [13].

Containment Similarity (CS) expresses the proportion of edges of a graph G_i that are shared with a graph G_j.

$$CS(G_i, G_j) = \frac{\sum_{e \in G_i} \mu(e, G_j)}{min(|G_i|, |G_j|)}$$

where e is an edge, and $\mu(e, G_i) = 1$ if, and only if $e \in G_i$. The cardinality $|G_i|$ here is intended as the number of edges of the graph G_i.

The ratio of sizes between two graphs is measured by the Size Similarity (SS):

$$SS(G_i, G_j) = \frac{min(|G_i|, |G_j|)}{max(|G_i|, |G_j|)}$$

We recall that with $|G_i|$ we indicate the number of edges in graph G_i

Value Similarity (VS) represents how many of the edges contained in G_i are contained in G_j, considering also the weight of such edges.

$$VS(G_i, G_j) = \frac{\sum_{e \in G_i} \frac{min(w_e^i, w_e^j)}{max(w_e^i, w_e^j)}}{max(|G_i|, |G_j|)}$$

where w_e^i is the weight of edge e in the graph G_i.

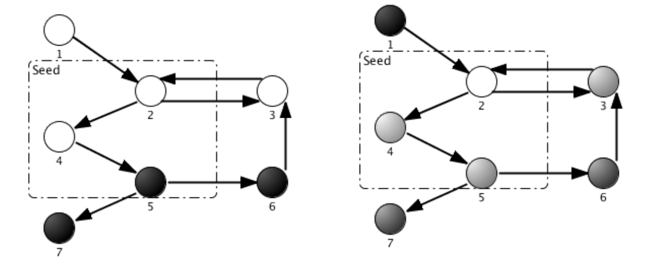
A combination of VS and SS gives another useful measure, called Normalized Value Similarity (NVS):

$$NVS(G_i, G_j) = \frac{VS(G_i, G_j)}{SS(G_i, G_j)}$$

4.2 Network Analysis

Apart from the text features described above that we use for the classification of online pharmacies, we additionally use features derived from the web network in which they are embedded. More specifically, we are interested in the links an online pharmacy web page has with other web pages: the web pages that this pharmacy points to, and the web pages that point to this pharmacy.

To this effect, we use features extracted from the TrustRank algorithm [15]. In TrustRank, the network is represented as a graph $\mathcal{G}(V, E)$, where the set of nodes V are websites (or more generally web pages) and the links between pages, represented by the set E, are drawn as directed edges. The algorithm computes a



(a) Initial state of a network made of "good" (white) and "bad" (black) nodes.

(b) The network after the execution of the TrustRank algorithm, with different levels of "trustiness".

Figure 3: Illustration of the TrustRank algorithm.

Algorithm 1 Creates the network graph $\mathcal{G}(V, E)$

```
GRAPH-CREATION($\mathcal{P}$ : set of pharmacies)
 1: $V \leftarrow \emptyset$
 2: $E \leftarrow \emptyset$
 3: for all Pharmacy $p \in \mathcal{P}$ do
 4:     $V \leftarrow V \cup p$
 5:     $\mathcal{L} \leftarrow outboundLinks(p)$
 6:     for all Link $u \in \mathcal{L}$ do
 7:         $V \leftarrow V \cup endpoint(u)$
 8:         $E \leftarrow E \cup \{u\}$
 9:     end for
10: end for
```

trust score for each node in the graph, based on the premise that "good" pages rarely point to "bad" ones (this property is also called *approximate isolation of good pages*). More specifically, TrustRank starts by selecting a seed of "known" pages, and gives a trust score of 1 to good pages and 0 to all others. After normalizing the values, the trust is propagated at each step until convergence. This process is illustrated in Figures 3a and 3b.

In order to run our version of TrustRank, we first need to construct the graph (refer to Algorithm 1), on which the algorithm will be applied.

Recall that the set $\mathcal{P}$ contains both labeled and unlabeled examples. The *outboundLinks*() function (line 5) accepts the URL of an online pharmacy as input, and returns all the outbound links for this website, namely, all links that point to external domains[9]. The *endpoint*() function (line 7), returns the final destination of a link, extracting the second level domain.

For example, assume that *outboundLinks*(p) for some website p returns the following set of links: "http://www.medicalnewstoday.com/articles/238663.php", and "http://www.fda.gov/forconsumers/consumerupdates/ucm149202.htm". Then, the function *endpoint*() applied on each one of these URLs will return: "medicalnewstoday.com", and "fda.gov", respectively.

This step is important because it allows us to significantly prune the feature space, which in this case is represented by URLs. Please also note that this step will not affect the quality as we can assume all the pages belonging to the same domain having the same trustiness.

In a second step we have to assign an initial trust score to each node in the graph. In our graph we have 4 types of nodes:

(1) *legitimate* nodes in $\mathcal{P}_0$; we denote this set with $\mathcal{P}_0^+$
(2) *illegitimate* nodes in $\mathcal{P}_0$; we denote this set with $\mathcal{P}_0^-$
(3) unknown pharmacy nodes in $\mathcal{P} \setminus \mathcal{P}_0$
(4) non-pharmacy nodes pointed to by nodes in $\mathcal{P}$

Note that the last category includes all the websites extracted from pharmacy links with functions *outboundLinks*() and *endpoint*(). Following our previous example, these are "medicalnewstoday.com" and "fda.gov".

During the initialization phase of TrustRank, we assign to all nodes in $\mathcal{P}_0$ the value returned by the oracle function, O, when invoked on these nodes. Hence, after the initialization, the known *legitimate* nodes (the first category of the list) have a trust score of 1, while all the other nodes have a value of 0. Finally, we can run TrustRank and train a classifier using the output values of corresponding nodes in $\mathcal{P}_0$.

5 ONLINE PHARMACY RANKING

Our goal here is, given a pharmacy, to calculate a value that indicates the degree by which this pharmacy is *legitimate* or *illegitimate*. Having these values for all pharmacies will allow us to compute the *totally ordered set* sought in Problem 2. We assume that the list is ranked in decreasing order of *legitimacy* (if $p_1 \leq p_2$ then p_1 is "less *legitimate*" than p_2). We propose a cumulative model that combines the models built with text and network:

$$rank(p) = textRank(p) + networkRank(p).$$

When we use the Term Vector model with TF-IDF to represent documents, the *textRank* of a pharmacy p is computed as the membership probability of this instance p to the *legitimate* class, as estimated by a classifier solving Problem 1. For example, in the case of the Naïve Bayesian Multinomial classifier, the probability of a document d being in class c is computed as:

$$P(c \mid d) \propto P(c) \prod_{1 \leq k \leq n_d} P(t_k \mid c),$$

where $P(c)$ is the prior probability of class c, and $P(t_k \mid c)$ is the conditional probability of term t_k occurring in a document of class c. If the classifier is non-probabilistic, like for example SVM, we give to *textRank* a value of 1 if the instance is classified as *legitimate* and 0 if it is classified as *illegitimate*, which is the same as the output of function T.

On the other hand, when we use the N-Gram Graphs representation model, we compute *textRank*() using a different formula: rather than considering the output of the classifier, we sum up the graph similarity measures according to this formula:

$$\begin{aligned} textRank(p) = {} & CS_{legitimate}(p) + (1 - CS_{illegitimate}(p)) \\ & + SS_{legitimate}(p) + (1 - SS_{illegitimate}(p)) \\ & + VS_{legitimate}(p) + (1 - VS_{illegitimate}(p)) \\ & + NVS_{legitimate}(p) + (1 - NVS_{illegitimate}(p)) \end{aligned} \tag{3}$$

The abbreviations CS, SS, VS and NVS denote containment, size, value and normalized value similarities for the class in subscript, described in Section 4.1.

The function *networkRank*() simply returns the TrustRank value computed with the algorithms presented in Section 4.2.

6 EXPERIMENTAL EVALUATION

We now evaluate the proposed approach using real data provided by an American company, who is a leader in internet pharmacy verification. We will call this company *PharmaVerComp*. The pharmacies used in our study have been manually labeled as

[9] In other contexts outbound links are sometimes called *external links*, or *outcoming links*.

	Dataset 1	**Dataset 2**
	Date 1	Date 2
		(6 months later)
# Examples	1459 (100%)	1442 (100%)
# Legitimate Examples	167 (12%)	167 (12%)
# Illegitimate Examples	1292 (88%)	1275 (88%)

Table 1: Datasets

legitimate or *illegitimate* by personnel of this company. Therefore, the dataset is consistent and error free.

All our code is publicly available at: https://sites.google.com/view/acolfplg/home.

6.1 Experimental Setup

At the time of this study, PharmaVerComp monitored almost 200, 000 health-related websites, out of which about 42, 000 are active internet pharmacies. Only 0.5% of them are legitimate, while 96.7% are not legitimate. The remaining 2.8% are pharmacies defined as potentially legitimate. This class is represented by pharmacies that do not fully adhere to the PharmaVerComp policies, but are probably not *illegitimate.* In this database, the examples are manually labeled by experts in the sector (i.e., human reviewers), and constitutes our ground truth.

We worked on two different instances of this database, generated with a six months difference, which were provided by PharmaVerComp. The summary statistics of these two instances are provided in Table 1. The intersection between the *illegitimate* instances of Dataset 1 and Dataset 2 is empty, i.e., in Dataset 2 we have 1275 different *illegitimate* domains. The two datasets contain the same *legitimate* instances, but crawled in different periods of time. We used crawler4j[10] in order to crawl each one of the domains in the two datasets, without depth limit, but for a maximum of 200 pages. In our analysis we use Dataset 1 as *base* dataset to test our algorithms, while we use Dataset 2 to inspect how our models evolve over time, i.e., how models built on "old" data behave in dealing with "new" data.

We observe that the two classes in both datasets are strongly imbalanced. In order to cope with this situation, we can use *undersampling*, which modifies the frequencies of the two classes by randomly removing examples belonging to the majority class in the training set, until the minority class reaches the same percentage as the majority class. The other technique we used is *SMOTE* [9]. SMOTE is an oversampling technique in which the minority class is oversampled by creating "synthetic" examples. Examples are created operating in "feature space" rather than in "data space", the opposite of what happens in oversampling with replacement. In our experiments we trained our classifiers using the natural class distribution as well as the ones generated using these two sampling techniques. Then, for each classifier, we took the one which offered the best results.

6.2 Evaluation Measures

In the following, we call "positive" the *legitimate* class, and "negative" the *illegitimate* class. With the notions TN, TP, FN, FP we denote respectively the number of true negatives, true positives, false negatives and false positive. Based on those, the evaluation measures we use are the following.

Overall Accuracy. Overall accuracy is the general correctness of the classifier and it is calculated as the sum of correct classified instances divided by the total number of instances:

$$Acc = \frac{(TP + TN)}{(TP + TN + FP + FN)}.$$

Note that in the case of imbalanced classes this indicator does not provide a very good evaluation measure. In fact, given the actual distribution of our dataset (12% *legitimate*, 88% *illegitimate*), a simple strategy of guessing the majority class would have an accuracy of 88%, but, of course, such a kind of classifier does not help us to distinguish between *illegitimate* and *legitimate* pharmacies.

Precision. Precision is a measure of the accuracy provided for a specific class. It is defined as the number of correct classified instances for a specific class divided by the total number of instances for that class. For example, the precision for the *legitimate* class is computed as:

$$Precision_{legitimate} = \frac{TP}{TP + FP}.$$

Recall. Recall measures how many examples of one class are classified correctly. The recall of the *legitimate* class is computed as:

$$Recall_{legitimate} = \frac{TP}{TP + FN}.$$

Due to the imbalance of the two classes, we expect a good classifier to have a high *illegitimate* precision and an high *legitimate* recall. This would mean that the classifier is able to correctly identify the *legitimate* examples.

Area Under ROC Curve. The ROC curve is drawn by plotting the *False Positive Rate* ($FPR = \frac{FP}{TN+FP}$) against the *True Positive Rate* ($TPR = \frac{TP}{TP+FN}$) of the classifier, at various threshold settings. The ideal point on this curve would be on the top left corner, meaning that all the positive examples are classified correctly, and no negative examples are classified as positive. The area under ROC curve is a useful measure, especially in the case of imbalanced datasets.

Pairwise Orderedness. For what concern the second problem, we rely on a measure generally adopted in ranking problems, *pairwise orderedness*, which is an indicator of the number of "violations" of the ordered property in a list. First of all we define a function I as follows:

$$I(p, q) = \begin{cases} 1 & \text{if } rank(p) \geq rank(q) \text{ and } O(p) < O(q) \\ 1 & \text{if } rank(p) \leq rank(q) \text{ and } O(p) > O(q) \\ 0 & \text{otherwise} \end{cases}$$

This function give 1 if and only if a *illegitimate* pharmacy receives an equal or higher score than a *legitimate* pharmacy. Then, we evaluate our ranking computing the fraction of the pairs for which there is not such a violation:

$$pairord(\mathcal{X}) = \frac{|\mathcal{X}| - \sum_{(p,q) \in \mathcal{X}} I(p, q)}{|\mathcal{X}|},$$

where $\mathcal{X}$ is the set of all the pairs of pharmacies $(p, q), p \neq q$, in the set $\mathcal{P} \setminus \mathcal{P}_0$. If *pairord* is equal to 1 there are no violations in the pairs and we have all the *legitimate* pharmacies at the top of the list and all the *illegitimate* ones at the bottom.

6.3 Classification Results

We ran experiments on Dataset 1 in order to test the methods described in Sections 4.1 and 4.2. In particular, we trained different classifiers with several combinations of text and network features. Below we present the results of the text and network

[10] https://github.com/yasserg/crawler4j

classification, and ensemble classification, where we combine both.

For all results, we computed the corresponding confidence interval, which indicates the reliability of the results. We used a confidence level of $100(1-\alpha)\%$ with $\alpha = 0.05$ (i.e., confidence level 95%). In all cases, the confidence intervals for our classifiers are very small (less than 1%): this means that the classifiers are stable, with the results of each fold being very close to the mean.

6.3.1 *Text Classification.* :

Text classifiers were trained with different text representation techniques on different subsets of the data. We employed 3-fold cross validation, where two folds were used for training and the third for testing. In the case of N-Gram Graphs, we randomly selected half of the training instances to build the class graph [13]. Hence, we compared these class graphs with all our instances (training and test). The N-Gram Graphs library we used to build the graphs and compute similarities is JInsect[11], while for the implementations of the classifiers we relied on Weka [16].

Table 2 lists the abbreviations concerning the classifiers and the sampling techniques employed in our experiments. We performed various tests with all combinations among classifiers and sampling techniques. However, due to space requirements, for each classifier we present only the sampling technique that performed best.

The overall accuracy with the TF-IDF representation are reported in Table 3. In all cases, accuracy is above 88%, with the best performers reaching 99%. However, as we can observe in Table 4, the J48 classifier has low *legitimate* recall for small subsamples of data. (As we already explained in Section 6.2, overall accuracy is not enough to properly evaluate a classifier in an imbalanced classes context.)

Increasing the number of words considered in the subsample generally results in better performance. SVM is the classifier that performs the best in terms of overall accuracy. Also for what concerns *legitimate* precision and *illegitimate* recall, the best performer is SVM (refer to Tables 4 and 5). It is interesting to note the inverse trend of NBM, whose efficacy decreases, especially in *legitimate* precision, as we consider larger term subsets. As expected, *illegitimate* precision is generally high, with all values above 93%. This derives directly by the imbalance of the two classes. In fact, since we have much less *legitimate* than *illegitimate* examples, if the classifier put some *legitimate* instances in the wrong class, this does not heavily affect the recall of the *illegitimate* class.

The AUC ROC curve, which is more robust to the case of imbalanced classes, in shown in Table 6. NBM is the winner in all cases considered. We note that SVM, which offers good results in terms of *legitimate* precision and *illegitimate* recall, does not have high AUC ROC values, especially for small subsets of terms. Another observation is that the choice of the sampling technique makes almost no difference for NBM and SVM. Instead, for J48, the sampling technique leads to substantial variations in performance. In particular, SMOTE is the sampling technique that offered the best results. Similar observations have also been documented in previous studies [8, 10].

We performed the same experiments using N-Gram Graphs, in order to compare the two text representation techniques. For N-Gram Graphs we do not use sampling, because of the nature of this representation.

Abbreviation	Description
NBM	Naïve Bayesian Multinomial
NB	Naïve Bayesian
SVM	Support Vector Machines
J48	Java implementation of C4.5 algorithm
MLP	Multilayer perceptron (Artificial Neural Networks)
NO	No sampling technique used
SUB	Subsampling
SMOTE	Oversampling with SMOTE algorithm

Table 2: Abbreviations

		#Terms				
		100	250	1000	2000	All
NBM	NO	0.97	0.97	0.96	0.96	0.95
SVM	NO	**0.97**	**0.99**	**0.99**	**0.99**	**0.99**
J48	SMOTE	0.89	0.92	0.93	0.96	0.95

Table 3: TF-IDF - Overall Accuracy

			#Terms				
			100	250	1000	2000	All
Recall	NBM	NO	**0.90**	**0.98**	**0.98**	**0.97**	**0.96**
	SVM	NO	0.76	0.90	0.93	0.92	0.95
	J48	SMOTE	0.57	0.61	0.71	0.83	0.78
Precision	NBM	NO	0.84	0.81	0.78	0.78	0.72
	SVM	NO	**0.96**	**0.98**	**0.97**	**0.98**	**0.98**
	J48	SMOTE	0.58	0.71	0.76	0.83	0.82

Table 4: TF-iDF - *legitimate* recall and precision

			#Terms				
			100	250	1000	2000	All
Recall	NBM	NO	0.98	0.97	0.96	0.96	0.94
	SVM	NO	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
	J48	SMOTE	0.94	0.96	0.97	0.97	0.98
Precision	NBM	NO	**0.98**	**0.99**	**0.99**	**0.99**	**0.99**
	SVM	NO	0.97	0.98	0.99	0.98	0.99
	J48	SMOTE	0.94	0.94	0.96	0.98	0.97

Table 5: TF-IDF - *illegitimate* recall and precision

		#Terms				
		100	250	1000	2000	All
NBM	NO	**0.99**	**0.99**	**0.99**	**0.99**	**0.98**
SVM	NO	0.88	0.95	0.97	0.96	0.97
J48	SMOTE	0.77	0.79	0.83	0.87	0.88

Table 6: TF-IDF - Area Under ROC Curve

		#Terms				
		#100	#250	#1000	#2000	All
NB	NO	0.89	0.89	0.94	0.95	0.92
SVM	NO	0.92	0.95	0.97	0.96	0.93
J48	NO	0.95	0.96	0.96	0.95	0.96
MLP	NO	**0.97**	**0.98**	**0.98**	**0.99**	**0.99**

Table 7: N-Gram Graphs - Classifiers Accuracy

			#Terms				
			#100	#250	#1000	#2000	All
Recall	NB	NO	0.53	0.48	0.63	0.70	0.60
	SVM	NO	0.43	0.61	0.77	0.73	0.60
	J48	NO	0.76	0.80	0.83	0.79	0.87
	MLP	NO	**0.90**	**0.89**	**0.94**	**0.95**	**0.97**
Precision	NB	NO	0.59	0.58	0.93	0.91	0.82
	SVM	NO	**0.99**	**0.98**	**0.97**	**0.98**	**0.91**
	J48	NO	0.87	0.91	0.88	0.83	0.89
	MLP	NO	0.88	0.93	0.94	0.95	0.94

Table 8: N-Gram Graphs - *legitimate* recall and precision

[11] http://sourceforge.net/projects/jinsect

			#Terms				
			#100	#250	#1000	#2000	All
Recall	NB	NO	0.94	0.95	0.99	0.99	0.98
	SVM	NO	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
	J48	NO	0.98	0.99	0.98	0.98	0.98
	MLP	NO	0.98	0.99	0.99	0.99	0.99
Precision	NB	NO	0.93	0.92	0.95	0.96	0.93
	SVM	NO	0.92	0.94	0.97	0.96	0.93
	J48	NO	0.96	0.97	0.98	0.97	0.98
	MLP	NO	**0.98**	**0.98**	**0.99**	**0.99**	**0.99**

Table 9: N-Gram Graphs - *illegitimate* recall and precision

		#Terms				
		#100	#250	#1000	#2000	All
NB	NO	0.90	0.91	0.94	0.92	0.95
SVM	NO	0.71	0.81	0.88	0.86	0.80
J48	NO	0.93	0.92	0.91	0.88	0.95
MLP	NO	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**

Table 10: N-Gram Graphs - Area Under ROC Curve

The overall accuracies are shown in Table 7. MLP (Artificial Neural Networks) is the classifier that offers the best accuracy results. MLP is also the winner when we consider *legitimate* recall and *illegitimate* precision, s well as the AUC ROC, though similarly to the TF-IDF case, SVM gives better results for *illegitimate* recall and *legitimate* precision (see Tables 8 and 9). Note that J48 is the second best classifier.

When we compare the results obtained by the two text representation techniques, we realize that they perform very close to one another. Nevertheless, TF-IDF has a small edge when compared to N-Gram Graphs, since it leads to slightly better *legitimate* recall and AUC ROC values for "small" documents, which are easier to handle.

The conclusion of these experiments is that the use of text classification in the process of internet pharmacy verification leads to good results, independently of the text representation technique used. The reason for such good performance resides on the fact that online *legitimate* and *illegitimate* pharmacies behave very differently when selling products.

Taking a close look at the most frequent terms used by *illegitimate* websites, we noticed that words like "viagra", "cialis" and "no prescription" appear more frequently compared to *legitimate* pharmacies, which usually target a more broad and educated audience. Therefore, our classifiers built on top of the text representations of the pharmacy websites (using TF-IDF, and N-Gram Graphs) are in general able to recognize these differences and correctly predict the class of new instances.

6.3.2 *Network Classification.* :

In Table 11, we report the ten most linked-to websites by *legitimate* and *illegitimate* pharmacies. We observe that the most linked-to websites we find in the *legitimate* list are the two major social networks, Facebook and Twitter. This is in accordance with a previous study, which claims that *illegitimate* online pharmacies have fewer store presence features than *legitimate* pharmacies [23].

We can also find in the *legitimate* list many government websites that are not present in its *illegitimate* counterpart. For example, fda.gov, which is an agency responsible for protecting and promoting public health, is the third most linked-to website among the *legitimate* examples, while it is not even present in the *illegitimate* list. This is also the case for other government websites, like the National Institute of Health (nih.gov), and the Centers for Disease Control and Prevention (cdc.gov).

pointed by *legitimate* website	pointed by *illegitimate* website
facebook.com	wikipedia.org
twitter.com	wordpress.org
fda.gov	drugs.com
google.com	securebilling-page.com
youtube.com	rxwinners.com
nih.gov	google.com
adobe.com	providesupport.com
cdc.gov	euro-med-store.com
doubleclick.net	statcounter.com
nabp.net	cipla.com

Table 11: Websites pointed to by *legitimate* and *illegitimate* pharmacies (top 10)

Classifier	Overall Accuracy	AUC ROC
NB	0.96	0.95

Table 12: Network - Overall Accuracy and AUC ROC

	legitimate precision	*legitimate* recall	*illegitimate* precision	*illegitimate* recall
NB	0.904	0.732	0.966	0.990

Table 13: Network - precision and recall

On the other hand, we note that the two most linked-to websites in the *illegitimate* list are not directly related to the health sector (wikipedia.org and wordpress.org). Furthermore, in the *illegitimate* list there are some websites that are themselves classified as *illegitimate* pharmacies (e.g., rxwinners.com)[12] This fact is supported by many studies, which report networks of *illegitimate* pharmacies connected together in an affiliated way, where there is a central website and multiple other sites link to it[13].

For the network experiments, we used the same settings as for the text classification. The dataset is divided into 3 folds (2 train, 1 test) and each experiment is repeated three times, changing the folds, according to cross-validation. Note that the two folds used for training represent the initial seed $\mathcal{P}_0$. In the set of graph nodes V, we assigned 1 to those nodes that represent *legitimate* pharmacies in $\mathcal{P}_0^+$, 0 to the others. We use the scores computed by TrustRank algorithm to train and test the classifiers, and the Naïve Bayes as the base classifier.

The overall accuracy and the AUC ROC are summarized in Table 12. The overall accuracy is around 96%, that is fairly close to the case of text classification, but for what concerns the AUC ROC curve the result is significantly worse. This is reflected also to *legitimate* recall, shown in Table 13, which is around 0.73, while for the other measures the method exhibits quite good results.

Given these results, we conclude that network analysis offers good performance in terms of *illegitimate* precision and recall, and could be used to assess the *legitimacy* of a pharmacy, even though it does not reach the level of confidence provided by the text analysis method.

6.3.3 *Ensemble Classification.* :

In order to enhance our results, we also combine the two analyses techniques, building a single model that embodies the characteristics of both the text and the network. In order to implement

[12] *Illegitimate* status verified via the LegitScript public interface (http://www.legitscript.com).

[13] http://legitscript.com/research

		legitimate		*illegitimate*		
	Acc.	Rec.	Prec.	Rec.	Prec.	AUC ROC
Ensem. Sel.	**0.96**	0.92	**0.96**	**0.99**	**0.99**	**0.99**
Neural (Text)	0.98	**0.94**	0.94	0.99	0.99	0.99
NB (Network)	0.95	0.73	0.90	0.99	0.97	0.95

Table 14: Ensemble Classification Results

			pairord
TF-IDF	NBM	NO	0.998
	SVM	NO	**0.999**
	J48	SMOTE	0.994
N-Gram Graph			0.998

Table 15: Ranking using TF-IDF and N-Gram Graphs

		Old-Old		**New-New**		**Old-New**	
		#Terms					
		250	1000	250	1000	250	1000
NBM	NO	0.99	0.99	0.99	0.99	0.99	0.99
SVM	NO	0.95	0.97	0.96	0.96	0.90	0.93
J48	SMOTE	0.79	0.83	0.80	0.84	0.74	0.78

Table 16: TF-IDF - Model over Time - Area Under ROC Curve

		Old-Old		**New-New**		**Old-New**	
		#Terms					
		250	1000	250	1000	250	1000
NBM	NO	0.81	0.78	0.73	0.74	0.61	0.57
SVM	NO	0.98	0.97	0.98	0.96	0.95	0.93
J48	SMOTE	0.71	0.76	0.67	0.79	0.66	0.51

Table 17: TF-IDF - Model over Time - *legitimate* Precision

this approach, we relied on “Ensemble Selection” [4], which is a method for constructing ensembles from libraries of models. We used an implementation of “Ensemble Selection” available in Weka. According to the process explained in [4], the library of models was trained, building a new model with standard parameters.

The results we obtained are presented in Table 14, comparing ensemble selection with the two best single models on text and network. For simplicity we report only the results considering subsamples of 1000 words; the other cases exhibit very similar results.

We note that ensemble selection increases the overall accuracy, *legitimate* precision and *illegitimate* recall when compared to the other classifiers. This results in a higher AUC ROC value, as well, which means that our ensemble classifier is the preferred method for this task.

6.4 Ranking Results

In Table 15, we summarize the ranking results for both the TF-IDF and the N-Gram Graph representations. As we expected, the results reflect the trend observed in the classification results. The best ranking is the one that is computed with the SVM classifier. Also the other classifier results follow the patterns highlighted in Section 6.3, with SVM and NBM performing better than J48.

As part of the ranking analysis, we performed an analysis of the *legitimate* and *illegitimate outliers*, i.e., the *illegitimate* examples that appear high in our ranking, and the *legitimate* examples that obtained poor score and appear at the bottom of the list. Performing a manual analysis on such examples gives us the possibility to check which *illegitimate* pharmacies are able to fool our system. Moreover, these insights could be very useful in helping *legitimate* pharmacies to better market themselves.

We extracted a subset of the *legitimate* and *illegitimate* outliers, and provided those to PharmaVerComp, in order to obtain feedback about their characteristics. The domain experts pointed out that *illegitimate* outliers, are in general not part of any *illegitimate* networks[14]. On the other hand the *legitimate* outliers are the pharmacies that offer new prescriptions, while the majority of them simply give the possibility to refill existing prescriptions.

6.5 Model Evolution over Time

As previously mentioned, we also want to test the behavior and robustness of our models over time. Given the good performance obtained with Dataset 1, we are now interested in evaluating how much our models are affected by time. In particular, we want to answer the following two questions:

(1) *Will models computed on the new dataset (Dataset 2) get the same performance as the models computed on the old dataset (Dataset 1)?*
(2) *Are models trained with the old data still valid on the new data?*

The answer to the first question will give us the opportunity to evaluate the robustness of the model. The answer to the second question, will evaluate the validity of the proposed models over time, and test whether it is necessary to train the models often, or not.

In the following analysis, we only consider the classification part of the proposed approach. Note that ranking models are derived directly from the classifiers. We report here the performance for the two most meaningful classification measures for our problem, that is, AUC ROC and *legitimate* precision.

As we have seen, the first one offers a good overall indication of how well our classification process works, while the second measure is very sensible due to the small number of *legitimate* examples. Moreover, we focus on results for the Term Vector with TF-IDF weights and subsets of 250 and 1000 words.

6.5.1 *New model with new data.* :

In order to answer the first question posed above, we ran the same experiments conducted in Section 6.3, using Dataset 2. The results are used to verify if our models are effective even when applied to a new test dataset, despite the fact that *illegitimate* pharmacies appear in and disappear form the web at a relatively high rate.

We report the results in Tables 16 and 17. To do the comparison we report also the results obtained with the old dataset, namely, Dataset 1. In particular, we indicate with “Old-Old” the results obtained when computing and testing models on Dataset 1, and with “New-New” the results obtained by building and testing models on Dataset 2. We observe that the two models achieve almost the same performance for both measures. The conclusion of this analysis is that our approach is stable in analyzing different datasets that follow the natural distribution of instance classes.

6.5.2 *Old model with new data.* :

The answer to the second question will help us understand whether or not we need to adapt our models to pharmacy behavior changes. In particular, we expect that pharmacies change their text content and their relationship with other websites over

[14] We recall that *illegitimate* examples are very likely to belong to *illegitimate* networks (see Section 6.3.2).

time. This may affect our approach, which is strongly based on these two factors.

We tested models computed on old dataset, i.e., Dataset 1, on the new Dataset 2. Recall that these two datasets were crawled with a difference of six months. In this period of time, online pharmacies may have changed their characteristics, especially the *illegitimate* pharmacies, since they may be closed by the inspection authorities.

The results of this analysis are presented in Tables 16 and 17 (column Old-New). We can observe that there is a small reduction in *legitimate* precision is evident, while the AUC ROC value remains almost the same.

The conclusion of this experiment is that our model is fairly robust over time. However, it has some problems related to the *legitimate* accuracy measure. In turn, this means that re-training and maintenance of the model is necessary to ensure good quality of results, though, it is not necessary that this re-training takes place very often.

7 CONCLUSIONS AND FUTURE WORK

In this paper, we proposed solutions for the problem of automatic internet pharmacy verification, which is becoming an increasingly relevant and important problem, gaining attention from both the public and the private sectors.

We formalized two different problems: first, a binary classification problem, where we define two classes, *legitimate* and *illegitimate*, and we classify the online pharmacies in one of them; and second, a ranking problem, where we seek a *totally ordered set* that defines a ranking among pharmacies. We then described solutions for both these problems, based on features that are relevant to the text and the network structure of online pharmacies. These are the first solutions that have been proposed in the literature in order to address the internet pharmacy verification problem.

We experimentally validated the effectiveness of our approach using two real datasets from two different time periods. The experiments demonstrate that the proposed algorithms can very accurately perform the classification and ranking tasks, and that the models we use are fairly robust over time. Our results confirm that our system could be effectively employed in the process of internet pharmacy verification, as well as in other similar tasks, offering considerable assistance to the human analysts dealing with such real-world problems.

As part of future work, we intend to extend our algorithms across two dimensions: (a) include in our network analysis non pharmacy websites that point to pharmacies, as well as consider websites at distances greater than one to our working set, and (b) study and evaluate classification schemes with combined (network and text), or additional features. In both cases, the aim will be to employ a richer input, and therefore to improve the performance of the algorithms.

Moreover, we plan to apply the proposed techniques to other domains of electronic commerce, where it will be possible to create publicly available datasets that can serve for making further progress in this area.

REFERENCES

[1] Charu C. Aggarwal and ChengXiang Zhai. A survey of text classification algorithms. In Charu C. Aggarwal and ChengXiang Zhai, editors, *Mining Text Data*. Springer, 2012.

[2] Rasim M. Alguliev, Ramiz M. Aliguliyev, and Saadat A. Nazirova. Classification of textual e-mail spam using data mining techniques. *Appl. Comp. Intell. Soft Comput.*, January 2011.

[3] Roger Bate and Kimberly Hess. Assessing website pharmacy drug quality: Safer than you think? *PLoS ONE*, 5(8), 08 2010.

[4] Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew, and Alex Ksikes. Ensemble selection from libraries of models. In *ICML'04*, 2004.

[5] Carlos Castillo, Debora Donato, Aristides Gionis, Vanessa Murdock, and Fabrizio Silvestri. Know your neighbors: Web spam detection using the web topology. In *SIGIR '07*, 2007.

[6] William B. Cavnar and John M. Trenkle. N-gram-based text categorization. In *SDAIR-94*, 1994.

[7] Soumen Chakrabarti, Byron Dom, Rakesh Agrawal, and Prabhakar Raghavan. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *The VLDB Journal*, 7(3), August 1998.

[8] Nitesh V. Chawla. C4.5 and imbalanced data sets: investigating the effect of sampling method, probabilistic estimate, and decision tree structure. In *ICML'03*, 2003.

[9] N.V. Chawla, K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *JAIR*, 16, 2002.

[10] Chris Drummond and Robert C Holte. *C4.5, Class Imbalance, and Cost Sensitivity: Why Under-sampling Beats Over-sampling*. 2003.

[11] Susan Dumais and Hao Chen. Hierarchical classification of web content. In *ACM SIGIR*, SIGIR '00, 2000.

[12] George Giannakopoulos, Vangelis Karkaletsis, George Vouros, and Panagiotis Stamatopoulos. Summarization system evaluation revisited: N-gram graphs. *TASLP*, 5(3), October 2008.

[13] George Giannakopoulos, Petra Mavridi, Georgios Paliouras, George Papadakis, and Konstantinos Tserpes. Representation models for text classification : a comparative analysis over three web document types. In *WIMS 2012*. ACM, 2012.

[14] George Giannakopoulos and Themis Palpanas. Revisiting the effect of history on learning performance: the problem of the demanding lord. *Knowl. Inf. Syst.*, 36(3):653–691, 2013.

[15] Zoltan Gyongyi, Hector Garcia-Molina, and Jan Pedersen. Combating web spam with trustrank. Tr, Stanford InfoLab, March 2004.

[16] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1), November 2009.

[17] J.Pierre. On the automated classification of web sites, 2001.

[18] Sepandar D. Kamvar, Mario T. Schlosser, and Hector Garcia-Molina. The eigentrust algorithm for reputation management in p2p networks. In *WWW'03*, 2003.

[19] Ahmed Khorsi. An overview of content-based spam filtering techniques. *Informatica (Slovenia)*, 31(3), 2007.

[20] Vijay Krishnan and Rashmi Raj. Web spam detection with anti-trustrank. In *AIRWeb*, 2006.

[21] LegitScript. No prescription required: Bing.com prescription drug ads. Technical report, LegitScript, 2009.

[22] Nektarios Leontiadis, Tyler Moore, and Nicolas Christin. Pick your poison: pricing and inventories at unlicensed online pharmacies. In *ACM EC'13*, 2013.

[23] Tamilla Mavlanova and Raquel Benbunan-Fich. What does your online pharmacy signal? a comparative analysis of website trust features. In *HICSS*, 2010.

[24] Katsiaryna Mirylenka, George Giannakopoulos, Le Minh Do, and Themis Palpanas. On classifier behavior in the presence of mislabeling noise. *Data Min. Knowl. Discov.*, 31(3):661–701, 2017.

[25] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly. Detecting spam web pages through content analysis. In *WWW*, 2006.

[26] World Health Organization. Counterfeit medicines. Technical report, World Health Organization, 02 2006.

[27] G. Orizio, A. Merla, J.P. Schulz, and U. Gelatti. Quality of online pharmacies and websites selling prescription drugs: A systematic review. *JMIR*, 13(3), 2011.

[28] G. Orizio, P. Schulz, S. Domenighini, L. Caimi, C. Rosati, S. Rubinelli, and U. Gelatti. Cyberdrugs: a cross-sectional study of online pharmacies characteristics. *EJPH*, 2009.

[29] B. Raskutti and A. Kowalczyk. Extreme re-balancing for svms: A case study. *SIGKDD*, 6(1), June 2004.

[30] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill, Inc., 1986.

[31] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1), March 2002.

[32] Aixin Sun, Ee-Peng Lim, and Ying Liu. On strategies for imbalanced text classification using svm: A comparative study. *Decis. Support Syst.*, 48(1), December 2009.

[33] Mikalai Tsytsarau and Themis Palpanas. Managing diverse sentiments at large scale. *IEEE Trans. Knowl. Data Eng.*, 28(11):3028–3040, 2016.

[34] Mikalai Tsytsarau, Themis Palpanas, and Kerstin Denecke. Scalable detection of sentiment-based contradictions. In *International Workshop on Knowledge Diversity on the Web (DiversiWeb), in conjunction with the World Wide Web Conference (WWW)*, 2011.

[35] G. Weiss and F. Provost. The effect of class distribution on classifier learning: An empirical study. Technical report, 2001.

[36] Y. Zhang, D. Zhang, M.M. Hassan, A. Alamri, and L. Peng. Cadre: Cloud-assisted drug recommendation service for online pharmacies. *MONET*, 20(3), 2015.