Learning Interpretable and Statistically Significant Knowledge from Unlabeled Corpora of Social Text Messages: A Novel Methodology of Descriptive Text Mining

Giacomo Frisoni, Gianluca Moro* and Antonella Carbonaro

Department of Computer Science and Engineering – DISI, University of Bologna, Via dell'Università 50, I-47522, Cesena, Italy

Keywords: Text Mining, Descriptive Analytics, Explainability, Latent Semantic Analysis, Unsupervised Learning, Rare Diseases.

Abstract: Though the strong evolution of knowledge learning models has characterized the last few years, the explanation of a phenomenon from text documents, called *descriptive text mining*, is still a difficult and poorly addressed problem. The need to work with unlabeled data, explainable approaches, unsupervised and domain independent solutions further increases the complexity of this task. Currently, existing techniques only partially solve the problem and have several limitations. In this paper, we propose a novel methodology of descriptive text mining, capable of offering accurate explanations in unsupervised settings and of quantifying the results based on their statistical significance. Considering the strong growth of patient communities on social platforms such as Facebook, we demonstrate the effectiveness of the contribution by taking the short social posts related to Esophageal Achalasia as a typical case study. Specifically, the methodology produces useful explanations about the experiences of patients and caregivers. Starting directly from the unlabeled patient's posts, we derive correct scientific correlations among symptoms, drugs, treatments, foods and so on.

1 INTRODUCTION

More and more large online communities of patients aggregate to share experiences and to look for answers to questions in order to safely improve their health conditions, such as "which are the most effective and safe medical treatments from patients' viewpoint ?", "what contributes to the failure of a certain medical treatment ?", "which foods cause or lighten a certain symptom ?" or "for what purposes an expert centre is more suitable than another ?".

This kind of answers require to discover, from large unstructured corpora of unlabeled short text messages, relationships among concepts of various nature (e.g., symptoms, treatments, drugs, foods). Furthermore, only the most significant ones, according to objective measures, should be selected.

In particular, we want to understand the underlying reasons that explain some phenomena of interest by bringing out quantifiable correlations according to usual statistical significance (i.e., the degree of certainty about the fact that the relationship between two or more variables is not caused by chance), thus enabling sorting, selection and filtering. In the medical field, these could be "citrus fruit" $\leftrightarrow$ "acid reflux": 87%, or "GERD" $\leftrightarrow$ "pantoprazole": 82%. The importance of the problem is evidenced also by the recent Kaggle competition on Covid-19[1].

Discovering facts based on significant relationships applies to a large number of completely different domains. For instance, understanding the main causes of destructive plane crashes or the reasons behind negative hotel reviews, directly from aviation text reports or from customer reviews respectively.

We refer to this task of discovering explanations of phenomena from unstructured texts as *descriptive text mining*, which is completely different from the predictive one, where instead the goal is to estimate the likelihood of a future outcome based on labeled data, like for instance text classification or sentiment analysis (Weiss et al., 2015).

Aspect-based sentiment analysis (Liu and Zhang, 2012) touches just partially the goal of descriptive text

*Contact author: gianluca.moro@unibo.it

[1]https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

Frisoni, G., Moro, G. and Carbonaro, A.
Learning Interpretable and Statistically Significant Knowledge from Unlabeled Corpora of Social Text Messages: A Novel Methodology of Descriptive Text Mining.
DOI: 10.5220/0009892001210132
In *Proceedings of the 9th International Conference on Data Science, Technology and Applications (DATA 2020)*, pages 121-132
ISBN: 978-989-758-440-4

mining. In fact, this kind of analysis infers an overall rating of each review from the evaluations assigned to each known feature that characterizes an entity, for example cleanliness, food, quietness and kindness for a hotel. It is applied to data which are labeled and restricted to an expected comparison schema between equivalent products or services (i.e., with the same known features), and it does not assess the correlations between the expressed concepts and the outcome rating of features. However, this approach cannot be applied in a general context like the one of patients, where the messages are unlabeled, without a predefined comparison schema, the features are unknown and the goal is to discover significant relationships among unbounded combinations of concepts underlying the explanation of phenomena.

Descriptive text mining is not the goal of deep learning, where the tasks are highly supervised, require large datasets (partly alleviated with transfer learning(Pagliarani et al., 2017)) to achieve satisfactory results and are not designed to explain the learned knowledge (Montavon et al., 2018). This lack of neural network transparency has recently opened a new research thread, called Explainable Artificial Intelligence (XAI) (Gunning, 2017; Liu et al., 2018). Even if the boundary between explaining a deep learning model and explaining a phenomenon may seem blurred, the concepts of explainability and descriptive text mining should not be confused.

Decision trees (Safavian and Landgrebe, 1991) represent a halfway solution between descriptive text mining and explainability, but they do not allow the identification of fine-grained correlations.

In this paper we propose a novel methodology of descriptive text mining for the explanation of phenomena from the unsupervised learning of underlying relationships with their statistical significance. The methodology is modular in various parts, which include documents preprocessing and classification, term weighting, and language model application for the representation of documents and terms inside a single latent semantic space. Subsequently, the adoption of information retrieval methods within the space thus constructed makes it possible to derive correlations between the represented vectors, and to progressively generate textual explanations. Chi-square hypothesis test is used to determine the statistical significance of the extracted knowledge. We introduce an implementation of the methodology based on an original use of LSA, as practical example of language model. However, the main contribute of the paper is not focused on LSA (which we consider easily replaceable by other solutions), but on the potential that derives from the combined use of the various techniques mentioned above, even in tasks deemed complex to manage (such as descriptive text mining) and currently of increasing importance. To evaluate its effectiveness, the methodology has been applied to the scenario initially described, with the aim of offering high-value answers to the many questions asked by a community of people living with a rare disease. To this end, the results have been validated with domain experts.

The paper is organized as follows. Section 2 briefly analyzes the existing works in the literature. In Section 3 we introduce our descriptive text mining methodology with a LSA implementation. Section 4 shows the application of the contribution on the medical case study and the results obtained. Finally, Section 5 sums up the work with conclusive remarks.

2 RELATED WORK

According to (Fisher and Marshall, 2009), *descriptive statistics* have the purpose of analyzing and summarizing the data collected in an experiment, expressing information mainly by means of charts and statistical indicators. Similarly, in text mining and data mining, *descriptive analytics* is often treated as a set of sub-tasks, each aimed at highlighting a certain type of useful information and exploring a particular aspect (e.g., word frequencies, strongest correlations among words and topic extraction). The search for causality factors behind a phenomenon is called *causal analysis* and, in the literature, the application of text mining for this purpose is scarcely addressed.

In (Ahonen et al., 1998), the authors employ data mining techniques for descriptive phrase extraction, based on episodes and episode rules, also accompanied by indications about their statistical significance. However, episode rules are not sufficiently expressive and require a different construction form depending on the specific problem that needs to be addressed.

A comprehensive survey on *aspect-based sentiment analysis* is reported in (Liu and Zhang, 2012). In many applications the selection of the aspects to be evaluated is carried out manually by an expert user, who specifies them in a supervised manner. Four main lines of work can be identified to perform the extraction of latent aspects from text: (i) frequent nouns and noun phrases identification, (ii) nouns in conjunction with opinion words, (iii) topic modeling, (iv) associations with opinion ratings related to documents. The primary focus of the contribution (even if independent of the sentiment analysis task) can be linked to the approaches supported by the fourth group.

Deep learning networks have achieved considerable performance among many tasks, like speech recognition (Nassif et al., 2019), image classification (Brinker et al., 2019), and question answering (Lan et al., 2019). These methods are typically applied to classification and regression problems, which differ significantly from descriptive text mining. Recent works demonstrate how simple it is to deceive neural networks with adversary inputs, and all this further increases the need both to investigate their reliability and alternative solutions (Nguyen et al., 2015; Moosavi-Dezfooli et al., 2017; Papernot et al., 2016; Jia and Liang, 2017). Explainability seeks to give answers on how a black-box model achieves the results it produces (e.g., the features considered by a classification model that distinguishes benign from malignant tumor cells), but not on *why* a phenomenon occurs. In medicine, above all, where wrong decisions by a system can be harmful, the ability to explain models and phenomena is considered essential (Mathews, 2019). XAI is a young and rapidly growing research area, but current solutions do not yet allow the exclusive adoption of deep learning models for the resolution of descriptive text mining. In fact, almost all of them refer to local explanations, while descriptive text mining aims to provide global explanations of phenomena. For example, if a typical XAI solution is able to detect the terms most responsible for the class prediction of a single textual document, a descriptive text mining task has instead the objective of finding the most representative terms of the distribution related to a certain class on the whole corpus.

Decision trees (Safavian and Landgrebe, 1991) can be used to explain knowledge with a symbolic and interpretable model, where the highest nodes are also the most important ones. With them we can discover the most relevant terms for a certain class (e.g., "pain" for negative opinion), but not correlations belonging to many classes that go beyond this aspect, such as those between drugs, symptoms, lifestyles and so on.

3 METHODOLOGY

Here we discuss a novel methodology of descriptive text mining capable of offering accurate probabilistic explanations in unsupervised settings. The proposed methodology is independent of the domain and lends itself to operate in two ways: *interactive* and *automatic*. Section 3.1 illustrates the interactive mode (where the user is part of the learning process and can explore the data during the analysis). Section 3.2 shows the transition to a fully automatic mode. Finally, Section 3.3 presents some observations on the contribution.

3.1 Interactive Knowledge Extraction

The overall approach comprises several stages, following the general knowledge discovery process. Below, all the phases are discussed in detail, paying attention to the importance of their combination in the context of the presented work.

3.1.1 Early Steps

Textual content typically has numerous imperfections, such as grammatical and spelling errors, and various types of noise. As an initial stage, a transformation pipeline can be applied with the aim of increasing the quality of documents (e.g., encoding uniformization, symbols normalization, URL removal, word lengthening fixing). Thanks to this phase, a document like *"i suffrer from achalasiaaa"* can be converted into *"I suffer from achalasia"*. Cleaning the text promotes the identification of concepts and correlations between them, improving the results of all subsequent phases and reducing the dimensionality.

At this point of the analysis, a pre-trained Named Entity Recognition (NER) system can be used for unsupervised categorization of terms contained in the corpus (e.g., places, foods, symptoms, drugs). This gives the opportunity for more in-depth analyzes of entity types, a better understanding of the description for the phenomenon of interest, and the possibility of being able to connect these concepts to those of already existing knowledge bases, such as Wikidata. Its placement among the first steps within the methodology is justified by the typical dependence of NER systems on preprocessing operations, like lowercasing, which could make them ineffective. Information regarding recognized entities must be reported directly in the textual content of the documents (entity tagging), such as *"I suffer from <achalasia;/medicine/disease;Q661015>"*.

The documents to be analyzed can be selected applying a filter on their content, for example by using regex patterns. This operation allows to distinguish global analyzes (carried out on all documents; e.g., "*") from local analyzes (focused on documents related to a particular concept, like a medical treatment; e.g., "poem|endoscopic myotomy").

Lemmatization allows to increase the similarity between the terms, and therefore their frequencies.

3.1.2 Documents Classification

After these preliminary steps, it is necessary to define the phenomenon to be investigated, which can be represented by the way in which a certain class is distributed over documents. The attribute to be considered as a class could already be available within the dataset, or it could be calculated for each document at this stage. The classification could coincide with an opinion mining task on patients' social posts, and the description could have the objective of understanding why the opinion of a certain medical treatment is overall negative. As another example, the classification could refer to a category of air accidents, and the description could be aimed at highlighting the factors that lead to destructive ones.

3.1.3 Analysis Preprocessing

This stage has the objective of preparing data for analysis. It typically includes transformations such as case-folding, replacement of punctuation and numbers with spaces (except for entity tags), extra white-spaces removal and stopwords removal. *Tokenization* is another central aspect in this step. The descriptive text mining methodology illustrated in this document can be used in general with any N-Gram tokenization (e.g., Unigram, Bigram, Trigram), both at the word and character level. Models based on the latter can be very powerful (Bojanowski et al., 2017), significantly increasing the correlations between tokens and therefore the ability to bring out latent associations. Since character-models would require additional steps (in order to reconstruct the words and return a meaningful explanation), the next stages of the methodology refer to a unigram word-level approach for simplicity.

3.1.4 Term-document Matrix Construction

A term-document matrix is extracted from the corpus, where each row stands for a unique term t, each column stands for a unique document d, and each cell contains the frequency with which t appears in d.

3.1.5 Feature Selection

Irrelevant terms (in addition to the stopwords) can be further removed in this stage. One way to accomplish this is to keep only the terms with a percentage frequency above a certain threshold, such as 1%. From a co-occurrence point of view, each term should appear at least twice. However, the percentage threshold for standard terms should be distinguished from that for entity terms. Depending on the domain and the specific data source, in fact, a concept of an interesting type (e.g., a drug) could be of fundamental importance for the purposes of the analysis even if it is scarcely mentioned in the documents. Term selection simplifies the model and makes it easier to interpret, reducing also computational times. In any case, this phase must be performed with caution because the more terms are eliminated, the more the latent correlations become weak.

3.1.6 Term Weighting

Raw counts do not consider the significance a term has in the document in which it appears. To better represent the importance of each term in each document, term weighting methods are applied to the term-document matrix. A good comparison of the available schemes is proposed in (Domeniconi et al., 2015). We suggest the use of a variation of the classic tf-idf, making use of a factor inverse to the *entropy* of the term (as defined by Shannon) in the non-local part of the formula. See Equation 1.

$$w_{t,d} = log(1 + tf_{t,d}) \times (1 - sEntropy(tdm)) \quad (1)$$

It strongly affects the production of the description for the phenomenon investigated. More specifically, it determines the norms of the vectors in the new space and therefore their impact on the result.

3.1.7 Language Model

This stage involves the application of a language model (LM), which forms the basis for the whole analysis. For space reasons, we have verified and used only *Latent Semantic Analysis (LSA)* (Landauer and Dumais, 1997), reinterpreting it in light of recent developments in the NLP field. Nonetheless, we believe it is replaceable with other approaches, such as those based on neural networks. BERT (Devlin et al., 2018) and SBERT (Reimers and Gurevych, 2019) are some examples. There are four reasons why we focused on LSA (extending it) in this first research.

1. It is an *algebraic* method, and therefore solid and explainable in the semantic correlations it returns.
2. It allows mapping *both terms and documents within the same latent semantic space*, in a coherent way. Though there are some advancements, word embeddings and document embeddings are instead meant to work only on words or documents in a mutually exclusive manner.
3. It gives the possibility to perform analyzes with a *reduced number of dimensions* (even just two). On the other hand, word embeddings require a large number of features to function properly. The BERT model released by Google, for example, uses 768 hidden units (Devlin et al., 2018).
4. It *does not require labeled data* and training.

LSA induces global knowledge indirectly from local *co-occurrence* data (without using syntax, linguistic, pragmatics or perceptual information about the physical world). It performs a mapping of the weighted term-document matrix in a reduced vector space (called *"latent semantic space"*) which approximates the original one, focusing on the essence of data. The mapping is based on *Singular Value Decomposition (SVD)*, a technique in linear algebra that factorizes any matrix C into the product of three separate matrices (Equation 2).

$$\underset{M\times N}{C} = \underset{M\times M}{U}\ \underset{M\times N}{\Sigma}\ \underset{N\times N}{V^T} \tag{2}$$

U and V are two orthogonal matrices, and Σ is a diagonal matrix containing the singular values of C. Formally, $\Sigma = diag(\sigma_1, \ldots, \sigma_p)$ where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$ and $p = min(M,N)$.

The singular values in Σ are the components of the new dimensions, indicating also their importance. Being in descending order, the first of them capture the greatest variation of the data (i.e., contain more information). SVD reduces dimensionality by selecting only the k largest singular values, and only keeping the first k columns of U and the first k rows of V^T. So, given a matrix C $M \times N$ and a positive integer k, SVD finds the matrix $C_k = U_k \Sigma_k V_k^t$ of rank at most k (between all matrices with k linearly independent vectors) that minimizes the difference with the original matrix $X = C - C_k$, according to the Frobenius norm (Equation 3). Figure 1 resumes this process.

$$\|X\|_F = \sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N} X_{ij}^2} \tag{3}$$

The positions of all terms and documents in the latent semantic space are obtained respectively from the products of matrices $U_k \times \Sigma_k$ and $V_k \times \Sigma_k$.

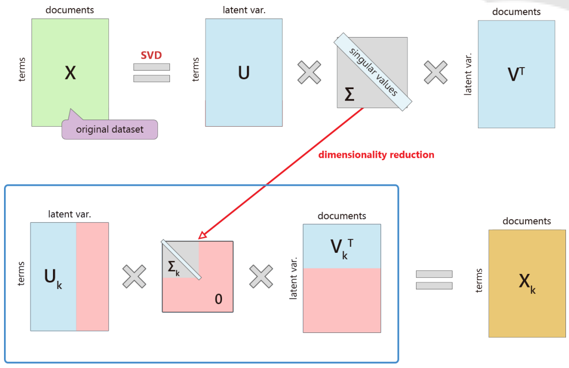

Figure 1: LSA dimensionality reduction through SVD.

Similarity is measured by the cosine between vectors (Equation 4).

$$sim(A,B) = cos(\theta) = \frac{A \cdot B}{\|A\|\,\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{4}$$

k is a *hyperparameter* we can select and adjust to delete noise and unnecessary data, as well as better capture the mutual implications of terms and documents. The number of dimensions retained in LSA is an empirical issue, as well as being highly dependent on the goal of the single application. Some heuristics are available, but to date there is still no way to establish it optimally.

Though its representation of reality is basic, relatively simple, and surely imperfect, LSA performs a powerful induction of knowledge that closely matches human meaning similarities and has the potential to address Plato's Problem (Landauer et al., 1998).

The computational cost of LSA is the same as SVD: $O(min\{MN^2, M^2N\})$. No-exact solutions for SVD significantly reduce costs by orders of magnitude compared to the exact version typically implemented (Halko et al., 2011). In some circumstances, this allows to obtain a decrease in execution times from a few hours to a few minutes.

3.1.8 2D Space Representation

The visualization of terms and documents in the latent semantic space (built in the previous step) is useful for several reasons. Starting from it, it is possible to:

- identify the correlations between terms and terms, documents and documents, and terms and documents;
- have a graphical feedback on the distribution of documents based on their class;
- recognize the presence of any clusters;
- understand the effectiveness of the model and whether it is necessary to intervene again on the previous phases to make adjustments.

Even if the new space is made up of k dimensions, a 2D representation must be adopted to make the graph suitable for human observation. Under this point of view, *t-SNE* (Maaten and Hinton, 2008) can be useful for mapping high-dimensional data to two dimensions, compressing all the original ones so as to minimize divergences. However, this task can be achieved independently of it. For example, with LSA as LM, the choice of the two dimensions to be adopted for visualization purposes can be made directly from the singular values in the matrix Σ. Since the latter are in descending order and indicate the importance of their dimensions in the transformed space, a good choice concerns the dimensions associated with two high singular values without too much difference between them (to have a good approximation and avoid a strong crushing of data on one axis with respect to the other). In any case, generally it does not go beyond the fourth dimension, because the information captured is much lower than the previous ones. The *power law curve* formed by the singular values in Σ is a valid tool to make this decision (Figure 2). To prevent terms and documents from being displayed at different scales, a good practice is to *normalize* vectors.

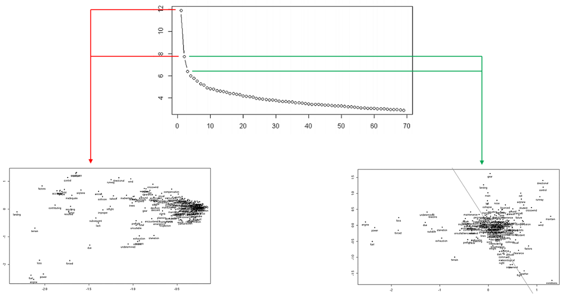

Figure 2: 2D visualization of the latent semantic space, starting from the power law curve obtained after the SVD decomposition (LSA). In the first case (dimensions 1 and 2), the resulting terms distribution is not satisfactory and takes the form of an ellipsoid. In the second case (dimensions 2 and 3), the distribution is less concentrated and consequently it is better for visual observations.

The cosine similarity allows the visual recognition of semantically related terms and documents. The more a pair of terms forms a small angle with the origin, the greater the similarity between them (i.e., they often appear together in documents and so are frequently associated). However, having compressed an originally high-dimensional space into only two dimensions, close terms in the graph may not necessarily be such (approximation).

A complete graph of the latent space can be obtained by considering the overlapping representation of normalized terms and documents (Figure 3). In order to better understand the distribution of vectors in space and the quality of the correlations, terms and documents can be colored according to their class.

3.1.9 Selection of the Number of Dimensions

If the use of only two dimensions is suitable for visualization purposes, in the rest of the analysis it could involve a significant loss of information. After applying the decomposition by choosing a number of dimensions k for the new space, a further reduction in dimensionality can be made to lighten the computational load required by the subsequent analytical phases. To decide an optimal value of the number of dimensions to use, it is possible to analyze the descending sequence of singular values to search for a *knee point* in the progression. This can be done both

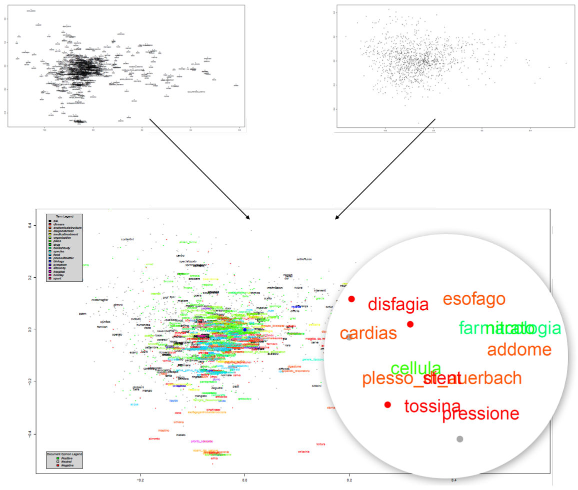

Figure 3: Graph of the latent semantic space with normalized and overlapping terms and documents.

visually and formally. In fact, a knee point is a point where the radius of the *curvature* of the function that interpolates the hyperbola corresponds to a local minimum. Considering that the curvature of a function $y = f(x)$ is $c = y''/(1+(y')^2)^{3/2}$, a valid number of dimensions can therefore coincide with one of its local minima (Figure 4). The idea is that the informative contribution given by the dimensions associated with the eigenvalues that follow a knee point is lower, making an approximation possible.

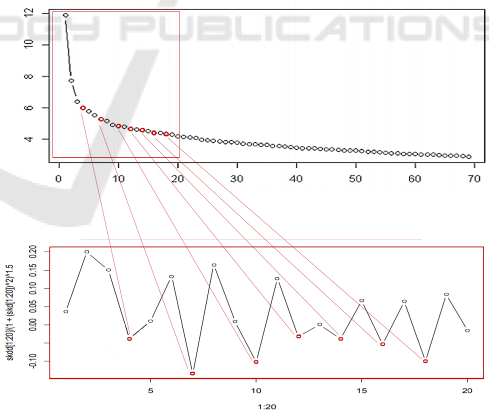

Figure 4: Selection of the number of dimensions in LSA, through the search for knee points in the curvature function given by the sequence of singular values.

In general it is advisable to perform tests with multiple minimums to choose the potentially optimal one. A way to conduct these tests is to verify the consistency with respect to a particular query of the first N documents semantically most similar to it, and so the semantic precision of the query itself (as better explained in Section 3.1.11). Since SVD (and so LSA) is based on co-occurrences, it is important to note that during these checks, real positives documents (i.e., actually related to the query) may be listed though these do not directly contain the query term(s). This could not happen with a Boolean research model based on lexical (and not semantic) matches.

3.1.10 Qualitative Analysis of the Graph

In the case of uniform distribution, imagining to divide the space into quadrants, a quantity of documents proportional to the original one should be found for each class. The areas of the space outlined by the LM in which there are unexpected concentrations (different from those foreseen in the case of random distribution), indicate the presence of elements of interest for the analysis (Figure 5). By researching which terms are found in these areas, it is possible to interpret them to identify the causes that contribute to the phenomenon. This phase of the methodology therefore has the aim of recognizing the possible presence of areas to be investigated.

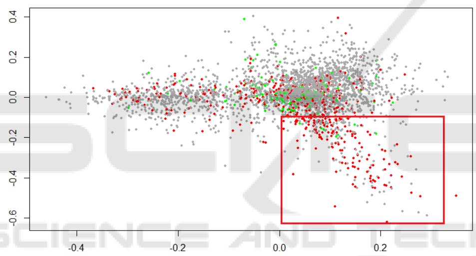

Figure 5: Example of unusual concentration of documents belonging to a certain class, in the 2D representation of the latent semantic space.

3.1.11 Description Construction

In this last phase, the methodology foresees to calculate the correlations between terms and terms, documents and documents, and terms and documents. The original paper with which LSA was introduced (Landauer and Dumais, 1997) focuses on the application of SVD and on the general use of cosine similarities, but neither describes nor offers solutions to the various types of correlations. We therefore introduce an *expansion of LSA operations* with those necessary for the objective of the research.

Within the reconstructed term-documents matrix C_k (Equation 2), the semantic similarities between the pairs of terms or documents are measured with the cosine of the respective scalar products ($C_k C_k^T = U_k \Sigma_k^2 U_k^T = (U_k \Sigma_k)(U_k \Sigma_k)^T$ or $C_k^T C_k = V_k \Sigma_k^2 V_k^T = (V_k \Sigma_k)(V_k \Sigma_k)^T$) and the specific instances of U_k or V_k. We argue that one of the most interesting features of LSA is the ability to *fold-in*[2] new documents, realizing the transposition of queries in the latent semantic space. A query q is equivalent to a set of terms in C (pseudo-document vector). It must undergo the same preliminary transformations that the cell entries of C received before the SVD application. Transforming a query vector q in a new document q_k means transforming it into a row of the matrix V_k. Since $V = C^T U \Sigma^{-1}$, it follows that $q_k = q^T U_k \Sigma_k^{-1}$. The position of the query in the latent semantic space is given by $q^T U \Sigma^{-1} \Sigma = q^T U$. Table 1 summarizes the equations for calculating the similarities between the vectors in the transformed space.

Table 1: Similarities between terms (u), documents (v) and queries (q) in the latent semantic space.

v_i and v_j	$cos(v_i \Sigma_k, v_j \Sigma_k)$
u_i and u_j	$cos(u_i \Sigma_k, u_j \Sigma_k)$
q and v_j	$cos(q^T U_k, v_j \Sigma_k)$
u_i and v_j	$cos(u_i \Sigma_k^{1/2}, \Sigma_k^{1/2} v_j)$
u_i and q	$cos(u_i \Sigma_k^{1/2}, \Sigma_k^{-1/2} U_k^T q)$

By following the original operations explained below, it is possible to construct a *probabilistic description* (step by step) for the phenomenon we have chosen to analyze. In particular, the resulting description will consist of a query (set of terms) that best characterize the distribution of the class representing the phenomenon in the latent semantic space.

First, the most representative term must be visually identified in the area highlighted in the previous stage. In doing this, it is necessary to focus on the terms placed in a central position within the area itself and at a greater distance from the origin (with a high norm and so a high relevance). The selected term can be seen as the first one of the descriptive query.

In order to mathematically demonstrate the correlation between the query q and the class c (representing the phenomenon), the *chi-squared (χ^2) test* can be used in conjunction with *R-precision* (Equation 5). To give the query the possibility of retrieving all the documents to which it refers, R is set equal to the number of instances of class c.

$$\chi^2(\mathbb{D}, q, c) = \sum_{e_q \in \{0,1\}} \sum_{e_c \in \{0,1\}} \frac{(N_{e_q e_c} - E_{e_q e_c})^2}{E_{e_q e_c}} \quad (5)$$

where:

$\mathbb{D}$ = corpus (repository of documents)
q = query
c = class

[2]Folding is the process of adding new vectors to a space after its construction, without rebuilding it.

e_q = p-a in the top-R identified by q
e_c = p-a of the class
$N_{e_q e_c}$ = documents number observed with e_t and e_c
$E_{e_q e_c}$ = documents number expected with e_t and e_c

While the number of documents observed is directly reflected in the data, the expected frequencies are calculated as $E_{e_q e_c} = |\mathbb{D}| \cdot P(t) \cdot P(c)$. For example, $E_{11} = |\mathbb{D}| \cdot ((N_{11} + N_{10})/|\mathbb{D}|) \cdot ((N_{11} + N_{01})/|\mathbb{D}|)$.

The higher χ^2, the lower the probability that the hypothesis of independence between q and c holds. Therefore, to obtain the level of *statistical significance* associated with the description of the phenomenon (for which only LSA is not enough), we consider the *p-value* obtainable from the χ^2 distribution table between q and c with a degree of freedom.

To establish whether the null hypothesis is rejected or not, a *p-value threshold* must be set. In the case of statistical dependence, let n_c be the number of documents belonging to class c in top-R and $|c|$ the number of instances of class c (used as R-precision), it is possible to say that the query characterizes a number of instances related to the phenomenon being described equal to $n_c/|c| \cdot 100$, with a probability corresponding to the p-value for the calculated χ^2.

After having verified with a formal approach the relevance of the term chosen for the class associated with the unusual concentration, it is possible to proceed with the extension of the description. The analysis therefore continues with the search for terms closest to the query. Among the terms with higher similarity, we choose the most significant one that also has a high norm and we insert it as a second element in q. Like before, the chi-squared test is rerun to verify the correlation between the new query and the class of interest. The process is repeated iteratively, as long as the confidence indicated by the p-value does not fall below a certain threshold. By alternating searches for query-terms and query-documents correlations, the proposed approach makes the description of the phenomenon progressively more specific. Once the analysis is complete, the methodology returns a set of terms that has not a sentence structure but is often easily interpretable. Figure 6 shows an example.

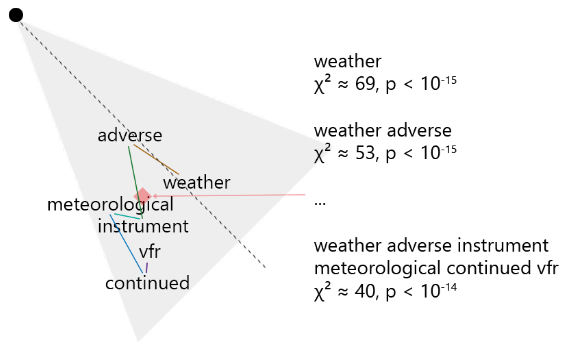

Figure 6: Example of description construction. At each step the query is enriched with a semantically close term, and the correlation with the class tends to decrease.

3.1.12 Evaluation

There are several ways to evaluate the correctness of the results produced by the methodology in its various parts. The description of the phenomenon can be interpreted and compared to existing scientific reports or data. As regards the correlations between concepts as a whole, a manual judgment can be applied to clusters of similar terms in the transformed space. Under this point of view, we suggest a more formal approach based on the definition of a file containing a set of *gold standards* (i.e., positive and negative known correlations expressed by an expert user). It can be made up of elements having a structure of this type: $x_1, \ldots, x_n \leftrightarrow y_1, \ldots, y_m$. An example of positive correlation (known in the literature for its truth) is *alcohol* $\leftrightarrow$ *GERD*. Vice versa, an example of negative correlation (known in the literature for its falsity) is *lung* $\leftrightarrow$ *gastroenterology*. Consequently, each correlation generally involves two fictitious documents (sets of terms). The effectiveness of the model can be proven with the same theoretical framework illustrated in Section 3.1.11 (i.e., by applying the chi-squared test between the two queries involved in each known correlation and verifying that the resulting p-value is below a certain threshold for positive examples and above for negative ones). This allows the creation of a *confusion matrix* and the calculation of several metrics.

3.2 Automatic Knowledge Extraction

This section describes a possible solution to fully automate the execution of the presented methodology.

3.2.1 Dimensionality Calibration

The choice of the number of dimensions to work with concerns both the transformed space generated by the language model (Section 3.1.7) and the reduction of dimensionality aimed at lightening the computational load (Section 3.1.9). Making this decision according to the domain is not always easy, and for automation purposes it cannot be standardized for all application cases. We propose a solution focused on the reuse of gold standards as training set. The idea is to iterate over the two parameters, make multiple attempts and choose the pair of values that most bring out the known correlations (e.g., best ratio between True Positive Rate and False Positive Rate). This mechanism emulates that of back-propagation. Some known facts can be used for *calibration*, while the rest as *tests* (also with *k-fold cross validation*).

3.2.2 Starting Query

Within a latent semantic space, there may be multiple areas of concentration (with non-random distributions) linked to the documents of the class of interest, and not necessarily just one. A way to solve the problem in an automated context is to ask the user to indicate a starting query, thus identifying the point of the space from which to proceed with the analysis (enriching the provided description with other statistically significant terms, if possible). In a medical domain, for example, this query can express personal information that you want to take into account (such as symptoms and performed medical treatments).

3.2.3 First Term Selection

If a user-specified starting query is not available, the automated version of the methodology must still be able to manage the choice of the initial term. This task can be solved by considering all the terms with norm higher than a certain threshold, and applying the chi-squared test on them in order to calculate their correlation with the class. The term chosen is the one with minimum p-value and highest norm. The risk linked to the automation of this phase is the choice of a term that is not particularly interesting or significant. In this regard, a valid preprocessing is fundamental.

3.2.4 Choice of the Next Term

Another aspect that needs to be automated concerns the choice of the term with which to enrich the description at each step. This can be done by searching for the N terms semantically closest to the current query, but in any case respecting a minimum cosine similarity threshold. Among them, the term used to extend the query is the one with the highest norm, capable of reaching a description with a sufficiently low p-value.

3.3 Considerations

A relevant strength of the proposed methodology lies in its flexibility. Some of the phases of which it is composed can be seen as *modules*, independent of the particular implementation (which therefore can be adapted to the application scenario or compared to alternative solutions). Furthermore, the methodology produces a *global explanation* of the phenomenon and extracts correlations between various concepts without using domain-dependent dictionaries or resources. This last feature also allows the *inference of new knowledge* (i.e., previously unknown correlations with high statistical significance). Gold standards are required only for evaluation purposes, or as possible semantic-based solution to the automatic choice of the optimal number of dimensions to work with. However, the calibration phase is not essential and could for example be replaced with the use of heuristics and with the adoption of the first minimum in the curvature function. In any case, the amount of data required is minimal and inexpensive to produce compared to that frequently needed by other solutions.

4 CASE STUDY

This section shows the application of the contributions to a case study of conversational messages (Domeniconi et al., 2016b) in the medical field, specifically focused on the domain of rare diseases. Experiments are conducted on unlabeled conversational posts about a rare disorder, called "Esophageal Achalasia". A descriptive text mining analysis is performed with the aim of collecting useful information to improve patients' living conditions.

4.1 Esophageal Achalasia Overview

Idiopathic Achalasia (*ORPHA:930*) is a rare disorder of the esophagus, with a prevalence rate estimated to be 1/10,000 (Patel and Vaezi, 2014). It is characterized by an impaired ability to push food down toward the stomach (peristalsis), due to the failure of the lower esophageal sphincter (LES) to relax.

4.2 Dataset

The dataset was built with the collaboration of *Associazione Malati Acalasia Esofagea (AMAE) Onlus*[34], the main Italian patient organization for the disease under consideration. In particular, we used the Facebook Group directly managed by AMAE, named *Acalasia esofagea... I malati "rari" non sono soli...!*[5]. It has been collecting data since 2008 and currently has around 2000 users (mainly patients, caregivers and doctors). Using the *Facebook Graph API*[6], data were downloaded for **6,917 posts**

[3] http://www.amae.it/

[4] https://www.orpha.net/consor/cgi-bin/SupportGroup_Search.php?lng=EN&data_id=106412

[5] https://www.facebook.com/groups/36705181245/

[6] https://developers.facebook.com/docs/graph-api/

and **61,692 first-level comments**, published between 21/02/2009 and 05/08/2019. The private nature of the group further improved the quality of the dataset, strongly limiting the presence of fake news and harmful content for other patients.

4.3 Methodology Implementation

The implementation of the methodology for the case study perfectly follows the steps described in Section 3, and is realized with a combined use of R and Python (through `reticulate` package).

To cope with typical text distortions in social contexts, the quality preprocessing pipeline includes operations such as emotes normalization, word lengthening fixing and Internet slang translation. In this first work NER and lemmatization were not used.

A common need among patients is to know whether other users' thoughts on a certain topic are positive or negative, and why. In this case, therefore, the classification corresponds to an opinion mining task. To estimate the opinion score associated with each document, we made use of a very simple algorithm based on *opinion words count*: words known for their semantic expression of polarity (Liu and Zhang, 2012). In particular, we used the opinion lexicon published by Hu and Liu[7] (containing 6800 positive and negative words), appropriately translated into Italian to manage the language mismatch. The specificity of the case study also made it appropriate to insert additional opinion words, both positive (e.g., "reborn") and negative (e.g., "reflux", "regurgitation"). The score is calculated according to Equation 6.

$$score(d) = nMatches(d, pos\_words) - nMatches(d, neg\_words) \tag{6}$$

4.4 Experiments

To survey the usefulness of the methodology (in its fully-automatic version) and the discovered knowledge in the context of the case study, we made two types of experiments. Firstly, through local analyzes, we investigated the reasons behind the positive and negative opinions on the two main surgical treatments for Achalasia: Heller-Dor and POEM. Secondly, we assessed the quality of the correlations obtainable with a global analysis on all documents. To consolidate the effectiveness of the contribution in quantifying the truth value of the identifiable correlations, gold standards were used (as indicated in Section 3.1.12). The structure of the latter is shown below, together with the results of the two experiments.

[7]https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon

4.4.1 Achalasia Gold Standards

The set of gold standards created consists of 120 positive correlations (i.e., true positive) and 104 negative ones (i.e., true negative). The ability to recognize correlations between various types of concepts was tested by dividing known facts into several categories: food $\leftrightarrow$ symptom, symtpom $\leftrightarrow$ drug, symtpom $\leftrightarrow$ anatomical_structure, drug $\leftrightarrow$ drug_class, place $\leftrightarrow$ doctor, etc. Depending on the topic, their definition was made by domain experts, who accompanied their facts with scientific sources.

4.4.2 Results

Interpreting the explanations returned by the local analyzes (Table 2), it can be appreciated how the names of the main doctors and expert centres (positive polarity), as well as the problems known in the literature for the treatments considered (negative polarity) have been identified with p-value < 0.01 (χ^2 test).

Table 2: Translated explanations returned by the methodology for positive and negative opinions about Achalasia treatments.

	Pos Explanation	Neg Explanation
Heller-Dor	equipe, dr, mario, costantini, salvador, padua, antireflux, plastic	problems, drink, eat
POEM	rome, prof, gemelli, costamagna, equipe, familiari, lombardy	reflux, problems, liquid, pain, inflammation, antacids

Table 3, on the other hand, shows the quality of the correlations extracted in the global analysis through the application of several statistical indices.

The underlying confusion matrix is constructed as follows.

- *TP*, number of known positive correlations identified (e.g., dysphagia $\leftrightarrow$ swallowing).
- *TN*, number of known negative correlations identified (e.g., stomach $\leftrightarrow$ tachycardia).
- *FP*, true negative correlations that the method incorrectly considers above the acceptance threshold, and therefore as positive.
- *FN*, false negative correlations that the method incorrectly considers below the acceptance threshold, and therefore as negative.

From the results it is possible to observe how the choice of dimensionality through a calibration phase leads to better results (albeit slight). In this case, the methodology allows to obtain $\approx 78\%$ accuracy.

Table 3: Comparison of confusion matrices and relative statistical indices obtained with and without the calibration phase based on gold standards, using three different acceptance thresholds. ACC = Accuracy, PRE = Precision, MR = MisclassificationRate, TPR = TruePositiveRate, TNR = TrueNegativeRate, FPR = FalsePositiveRate, FNR = FalseNegativeRate.

		Confusion Matrix				Statistical Indices						
Solution	P-value Threshold	TP	TN	FP	FN	ACC	PRE	MR	TPR	TNR	FPR	FNR
With Calibration k = 100, min = 6	0.7	101	68	36	19	75.45	73.72	24.55	84.17	65.38	34.62	15.83
	0.8	96	78	26	24	**77.68**	**78.69**	**22.32**	**80.00**	**75.00**	**21.31**	**20.00**
	0.9	72	98	6	48	75.89	92.31	24.11	60.00	94.23	5.77	40.00
Without Calibration k = 969, min = 4	0.7	99	67	37	21	74.11	72.79	25.89	82.5	64.42	35.58	17.50
	0.8	94	78	26	26	**76.79**	**78.33**	**23.21**	**78.33**	**75.00**	**25.00**	**21.67**
	0.9	71	97	7	49	75.00	91.03	25.00	59.17	93.27	6.73	40.83

5 CONCLUSIONS

We proposed a general and unsupervised methodology of descriptive text mining, capable of working with unlabeled data and accompanying the results with accurate probabilistic information. By modeling terms and documents together in a latent semantic space, we used a personal expansion of LSA to identify global textual explanations of phenomena, extracting also correlations between concepts of various kinds. We conducted experiments as part of a case study focused on Esophageal Achalasia. Through the discovery of statistically significant evidences, the methodology allowed the identification of scientific medical correlations directly from the patients' posts.

The work can be extended in several directions: the introduction of a NER system on the application level increases the expressiveness of the results, moreover a conditional GAN and/or the modeling of dependencies between terms in space, allows the construction of meaningful sentences. Pre-computing all the correlations between pairs of terms, with p-values below a certain threshold, helps patients to easily navigate among the most significant knowledge. Modern hierarchical clustering techniques, popular in many other domains (Cerroni et al., 2015), can also be applied over the results, e.g. to automatically extract semantically related terms. Transfer learning techniques increasingly play a key role with unlabeled data, from medical field (Domeniconi et al., 2014a; Domeniconi et al., 2014b; Domeniconi et al., 2016a) to NLP tasks such as opinion mining (Domeniconi et al., 2017; Moro et al., 2018), which we intend to deepen. The methodology can be applied on other diseases or completely different domains and languages, including the scientific medical literature. We also plan to represent the extracted knowledge by means of logic (Riccucci et al., 2007), knowledge graphs and semantic web techniques (Carbonaro et al., 2018), enabling reasoning.

ACKNOWLEDGMENTS

The research was developed starting from a university lesson by Professor Gianluca Moro, who is the author of the core contribution. We would like to thank Cristina Lanni[8], University of Pavia, for participating in the realization of the drug-related gold standards due to her expertise in the pharmacological field. We also thank Celeste Napolitano (President of AMAE and National Secretary of the Italian Society of Narrative Medicine), for her precious help in known correlations about doctors and expert centres.

REFERENCES

Ahonen, H., Heinonen, O., Klemettinen, M., and Verkamo, A. I. (1998). Applying data mining techniques for descriptive phrase extraction in digital document collections. In *IEEE ADL'98*, pages 2–11.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Brinker, T. J., Hekler, A., Enk, A. H., Klode, J., Hauschild, A., et al. (2019). Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. *European Journal of Cancer*, 113:47–54.

Carbonaro, A., Piccinini, F., and Reda, R. (2018). Integrating heterogeneous data of healthcare devices to enable domain data management. *Journal of e-Learning and Knowledge Society*, 14(1).

Cerroni, W., Moro, G., Pasolini, R., and Ramilli, M. (2015). Decentralized Detection of Network Attacks Through P2P Data Clustering of SNMP Data. *Computers & Security*, 52:1–16.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Domeniconi, G., Masseroli, M., Moro, G., and Pinoli, P. (2014a). Discovering new gene functionalities from

[8]https://www.researchgate.net/profile/Cristina_Lanni/research

random perturbations of known gene ontological annotations. In *KDIR 2014 - Proceedings of the International Conference on Knowledge Discovery and Information Retrieval, Rome, Italy*, pages 107–116. SciTePress.

Domeniconi, G., Masseroli, M., Moro, G., and Pinoli, P. (2016a). Cross-organism learning method to discover new gene functionalities. *Computer methods and programs in biomedicine*, 126:20–34.

Domeniconi, G., Moro, G., Pagliarani, A., and Pasolini, R. (2017). On deep learning in cross-domain sentiment classification. In *Proceedings of the 9th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - (Volume 1), Funchal, Madeira, Portugal, 2017*, pages 50–60. SciTePress.

Domeniconi, G., Moro, G., Pasolini, R., and Sartori, C. (2014b). Iterative refining of category profiles for nearest centroid cross-domain text classification. In *International Joint Conference on Knowledge Discovery, Knowledge Engineering, and Knowledge Management*, pages 50–67. Springer.

Domeniconi, G., Moro, G., Pasolini, R., and Sartori, C. (2015). A Study on Term Weighting for Text Categorization: A Novel Supervised Variant of tf. idf. In *DATA*, pages 26–37.

Domeniconi, G., Semertzidis, K., Lopez, V., Daly, E. M., Kotoulas, S., et al. (2016b). A novel method for unsupervised and supervised conversational message thread detection. In *DATA*, pages 43–54.

Fisher, M. J. and Marshall, A. P. (2009). Understanding descriptive statistics. *Australian Critical Care*, 22(2):93–97.

Gunning, D. (2017). Explainable Artificial Intelligence (XAI). *Defense Advanced Research Projects Agency (DARPA), nd Web*, 2.

Halko, N., Martinsson, P.-G., and Tropp, J. A. (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288.

Jia, R. and Liang, P. (2017). Adversarial examples for evaluating reading comprehension systems. *arXiv:1707.07328*.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., et al. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv:1909.11942*.

Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284.

Liu, B. and Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In *Mining text data*, pages 415–463. Springer.

Liu, H., Yin, Q., and Wang, W. Y. (2018). Towards explainable nlp: A generative explanation framework for text classification. *arXiv:1811.00196*.

Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605.

Mathews, S. M. (2019). Explainable Artificial Intelligence Applications in NLP, Biomedical, and Malware Classification: A Literature Review. In *Intelligent Computing-Proc. of the Computing Conference*.

Montavon, G., Samek, W., and Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15.

Moosavi-Dezfooli, S.-M., Fawzi, A., Fawzi, O., and Frossard, P. (2017). Universal adversarial perturbations. In *CVPR*, pages 1765–1773.

Moro, G., Pagliarani, A., Pasolini, R., and Sartori, C. (2018). Cross-domain & in-domain sentiment analysis with memory-based deep neural networks. In *Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, IC3K 2018, Volume 1: KDIR, Seville, Spain, 2018*, pages 125–136.

Nassif, A. B., Shahin, I., Attili, I., Azzeh, M., and Shaalan, K. (2019). Speech recognition using deep neural networks: A systematic review. *IEEE Access*, 7:19143–19165.

Nguyen, A., Yosinski, J., and Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *CVPR*, pages 427–436.

Pagliarani, A., Moro, G., Pasolini, R., and Domeniconi, G. (2017). Transfer learning in sentiment classification with deep neural networks. In *Knowledge Discovery, Knowledge Engineering and Knowledge Management - 9th International Joint Conference, IC3K 2017, Funchal, Madeira, Portugal, 2017, Revised Selected Papers*, pages 3–25.

Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., et al. (2016). The limitations of deep learning in adversarial settings. In *EuroS&P*, pages 372–387.

Patel, D. and Vaezi, M. (2014). Idiopathic achalasia. https://www.orpha.net/consor/cgi-bin/OC_Exp.php?lng=EN&Expert=930. Accessed 6 Apr 2020.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv:1908.10084*.

Riccucci, S., Carbonaro, A., and Casadei, G. (2007). Knowledge acquisition in intelligent tutoring system: A data mining approach. In *Mexican International Conference on Artificial Intelligence*, pages 1195–1205. Springer.

Safavian, S. R. and Landgrebe, D. A. (1991). A survey of decision tree classifier methodology. *IEEE Trans. Systems, Man, and Cybernetics*, 21:660–674.

Weiss, S. M., Indurkhya, N., and Zhang, T. (2015). *Fundamentals of predictive text mining*. Springer.