# What Drives Research Efforts? Find Scientific Claims That Count!

José María González Pinto
pinto@ifis.cs.tu-bs.de
Technische Universität Braunschweig
Institute for Information Systems
Brunswick, Lower Saxony, Germany

Janus Wawrzinek
wawrzinek@ifis.cs.tu-bs.de
Technische Universität Braunschweig
Institute for Information Systems
Brunswick, Lower Saxony, Germany

Wolf-Tilo Balke
balke@ifis.cs.tu-bs.de
Technische Universität Braunschweig
Institute for Information Systems
Brunswick, Lower Saxony, Germany

## ABSTRACT

Researchers often struggle to solve a common problem: how does one know whether a research hypothesis *is worth investigating?* Given the increasing number of research publications, it is complicated to guide such decisions. Previous work has shown how predicting generally emerging research topics can provide some help. Yet, in specialized scientific domains, only little is known about how to provide a service that allows users to ease the identification of *scientific claims* worth investigating. Scientific claims here means a natural language sentence that expresses a relationship between two entities. In particular, how one of them affects, manipulates, or causes the other entity. In this paper, we propose a data-driven approach aiming at filling this gap and empowering users at query level: given the results of a query, we deliver a *characterization* of clusters of the query results to discover the *contextualization* of scientific claims and the *identification* of those claims that may be worth more research efforts. To do so, we cluster documents with scientific claims that share the same context by leveraging co-clustering. After that, we characterize the clusters to annotate them. Our annotation focuses on two core aspects: *controversy* and *diversity* of claims in a given cluster. Controversy arises when two or more claims semantically contradict each other; diversity means the presence of different semantics of the claims that do not contradict each other but provide different insights expressed by some paper. To evaluate the benefits of our approach, we performed an extensive retrospective analysis on PubMed.

## CCS CONCEPTS

• **Information systems → Digital libraries and archives**; *Clustering*; *Content analysis and feature selection.*

## KEYWORDS

scientific claims, metadata generation, cluster characterization

## 1 INTRODUCTION

Due to the increasing number of available publications [22] with an estimate of one paper published every 30 seconds ‘It is practically impossible for researchers to keep up’ [6]. Thus, satisfying complex information needs is becoming more difficult. Consider for example the task of understanding the landscape of current research trends to design new hypothesis and experiments. In general, this task has implications for research funding, peer-review assessment, and new grad students. Consider for instance Anna, a new grad student in the medical school of some university. She would really like to understand better whether ‘smoking causes lung cancer’ and to discover novel aspects of such a claim that may need further investigation. Thus, Anna will need at least three steps:

(1) Find relevant documents, i.e. research papers, where any association between ‘smoking’ and ‘lung cancer’ has been studied within particular problem settings (the document space),
(2) Find out what the individual context of each document is, e.g., what other entities are involved (the contextual space),
(3) Given all these documents, organize them to decide which areas of research may represent opportunities to design new hypotheses and experiments (a grouping method to ease new hypothesis generation).

In general, this is a time-consuming task and modern discovery systems in high-quality Digital Libraries may support Anna. Thus, she might consider submitting a query to a curated Digital Library with high-quality sources of information such as PubMed. Anna’s query in PubMed will deliver more than 13 thousand results. In Figure 1 we show a frequency plot regarding the number of publications retrieved per year. The topic indeed seems still engaging; at least we can conclude that from the increasing number of publications. However, figuring out what specific aspects need further research is still cumbersome. Thus, to ease Anna’s task, we focus on how to re-organize the result set of Anna’s query. In our quest, we use clustering as one of the core steps but with a focus on how to *characterize and annotate* the clusters to ease Anna’s task. Our characterization focusses on revealing indicators within a cluster such as the presence of ‘controversy’, ‘diversity’ and how ‘homogenous’ the cluster is. Why do we focus on these aspects? Finding controversy manually has helped researchers to put into context the challenges behind the study of certain diseases such as in [38]. Thus, to alleviate the burden of manual work, we provide an automatic method to annotate such cases. Diversity can help Anna to get a consensus of what is known about the different associations that exist between entities [9]. Clustering plays a critical role in our proposed approach to ease the semantic interpretation of the result set of a query. In particular, clusters should be homogenous, e.g., semantically coherent (cf. [41]).

To achieve our goal, we build on the following observations: Firstly, we concentrate on a specific aspect of each scientific paper that allows for a meaningful organization of the results: *claims.* Claims are sentences that express some association between entities, see [8, 24]. This crucial step is part of our ‘Filtering and representation’ (see Figure 2); secondly, ‘controversy’ regarding the semantics of claims within clusters could also unveil what drives research efforts. The controversy arises when in a given cluster two or more claims contradict each other. For instance, when we

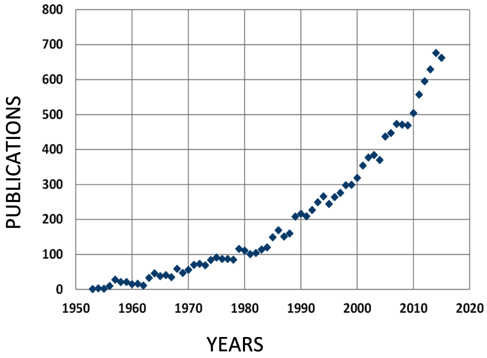

**Figure 1: Numbers of PubMed results for the query "smoking causes lung cancer" per year.**

have two documents with the following situation: document 1 contains the following claim: 'drug X helps to prevent disease Y', while document 2 states the following claim: 'drug X induces disease Y'. Thirdly, we consider the degree of 'diversity', e.g., the existence of claims that connect the same entities but with different associations that not necessarily contradict each other. Finally, we also consider how 'homogenous' a given group of papers is regarding some domain entities. With these factors, we propose to 1) represent documents by mapping them to a semantic space 2) construct a semantic matrix representation of them and cluster the semantic matrix and 3) characterize clustering with our proposed indicators. A natural way to re-organize our semantic matrix is to perform clustering. In particular, we propose to re-organize the matrix in such a way that allows the *grouping* of a subset of documents with claims that are more related with each other because they share the same context, e.g., share the same set of words and entities. Thus, we build on the idea of 'co-clustering': simultaneously clustering of rows and columns to re-organize the matrix, to facilitate Anna's task [5, 7, 21]. In Figure 2 we show our proposed methodology.

## 2 RELATED WORK

Our work strongly builds on a recent body of knowledge in the field of argumentation mining. Argumentation mining has a focus on modeling and extracting argument structures. Particularly relevant to our current efforts [8, 10, 11] is the work of [24] where the idea of context-dependent claims was introduced, which we in turn extended by our work. Instead of a claim being a general statement that supports or contests a given topic, a claim in our work is a sentence in a scientific document that relates two entities given in a query. For a broader perspective on the impact of the argumentation mining field see e.g., [26].

Recent efforts towards extracting and organizing the information of documents relevant to a query such as [39] are also related to our work. In [39] a case study using text mining techniques of scientific literature to build a network around the tumor suppressor p53 to predict new protein interactions with p53. In our work, we focus on a more general setting using scientific claims as our basic unit of representations towards a characterization of clusters that can be used to boost prototypes such as the one introduced by the article mentioned above. The work in [29] presents a generalization framework of [39] using network analysis to predict protein interactions using scientific literature. The system that the authors developed, called Knowledge Integration Toolkit, includes a reasoning component to predict new interactions between proteins that users can rely on to formulate new hypothesis [29].

In a similar line of thought, the work of [37] offers a new approach to detect the emergence of new research topics. What makes the work of [37] stand from other previous approaches focusing on topic detection is the fact that they can detect research topics at an earlier stage instead of topics that are already associated with a certain number of publications [37].

The work of [33] shows a model that focuses on document metadata such as 'objectives' or 'conclusions' contained in research papers to predict the rise and fall of popularity of scientific topics represented with keywords. Our approach can complement the previously mentioned body of work that aims at detecting the emergence of new topics, or new associations between scientific entities, by introducing a semantic characterization of clusters of documents based on scientific claims.

In our experiments, we prove the benefits of our characterization by looking at a retrospective analysis task that shows the potential of our characterization. Studying the dynamics of topics in Digital Libraries [31] is another example of the need for approaches that can ease user's understanding over a set of documents. In our work, we focus on providing a characterization of clusters based on scientific claims that can push further some of these previous efforts.

Our work is also related to the text mining efforts in biomedical literature to help researchers in the difficult task of finding where new researcher efforts are needed. In particular, our work adds value to the ongoing effort referred to as extra-propositional meaning that focuses on the detection of uncertainty, negations, hedging, opinions, and beliefs see [17] for a more in-depth overview. For instance, consider the work of Light et al. [25] that emphasized the relevance of detecting speculations instead of well-established facts. Light et al. found that the existence of speculative language in MEDLINE is not rare; it accounts for an estimate of 11% of sentences in MEDLINE abstracts. Moreover, even if definitive statements are of primary interest, knowing that a statement is not definite, i.e., speculative, is relevant [25]. In a nutshell, Light et al. observations can help a user to find where more research efforts are needed. As we will show, our work can add value to these efforts by introducing a claim-based characterization of semantically coherent clusters where controversy might exist.

## 3 PROBLEM DEFINITION AND APPROACH

We now introduce the idea of claims as a basic unit to filter the result set of some given query. The query that we study in our work follows the pattern: $(e_i, e_j)$, where $e_i$ and $e_j$ are arbitrary entities such as 'cigarettes' and 'cancer'. Moreover, a scientific claim (or merely claim) is a natural language sentence in a scientific paper that expresses a certain relationship between two entities. In particular, how one of them affects, manipulates, or causes the other entity.

An example of a claim is the following "Smoking cigarettes has the potential to increase the risk of lung cancer." In this example,

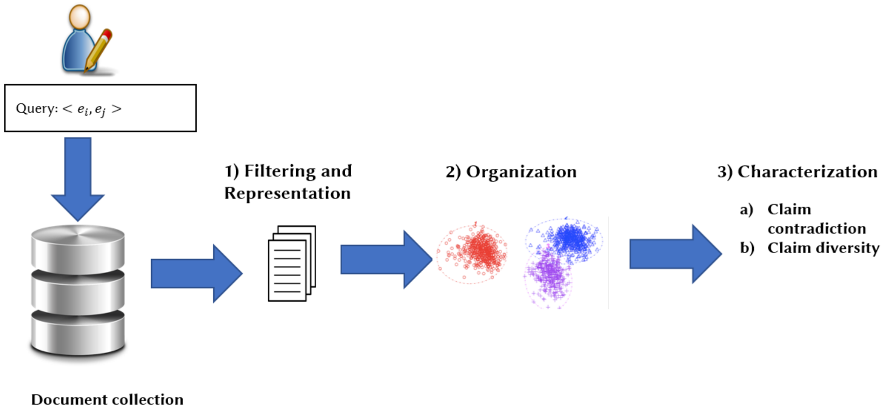

**Figure 2: Approach Overview.**

'cigarettes' and 'lung cancer' are the entities, and the relationship between them is 'increase the risk'. The relationships and entities of interest are domain dependent, and our approach can be applied once they are explicitly defined and identified.

We can now formally define the problem as follows:

Problem Definition. *Given the results $d = \{d_1, ..., d_n\}$ of some query $(e_i, e_j)$, discover a characterization of the grouping of the results based on their individual relationships, such that claims can be detected, which are worth investigating (i.e. which with high probability will lead to the publication of more papers in future).*

To instantiate this characterization, we first annotate the semantic orientation of claims in the documents automatically. Afterward, we assess and annotate three cluster properties: 'controversy', 'diversity' and 'homogenous'. They together have the potential to help users to formulate, discover or put into context scientific claims that may need more research efforts. For instance, in the case of controversy, we mean that a cluster contains scientific claims, whose semantic orientation contradict each other. For instance, if we have in some given cluster $Cl_1$ a document $d_1$ with a claim such as 'drug X treats disease Y' and a document $d_2$ with a claim such as 'drug X causes disease Y'. Here 'treats' and 'causes' clearly contradict each other.

To solve our problem, we propose three steps: 1) Filtering and Representation; 2) Organization; 3) Characterization, see Fig. 2.

Firstly, Filtering and Representation, we focus on two main tasks: firstly, we filter each document that does not contain a claim connecting the entities of the query. Secondly, once we have documents with claims that connect the entities of the query, we proceed to label the semantic orientation of each claim. By semantic orientation we mean what specific association exists between the entities in a given claim, e.g., 'causal', 'affects'. In the organization step, we focus on assessing the value of our proposed clustering approach by comparing its semantic quality with K-Means. Moreover, thirdly, we proceed to characterize the clusters with our proposed claim-based indicators.

To evaluate the value of our proposed characterization of the clusters, we perform a study of the proposed properties to predict clusters of scientific claims that will lead to an increase in the number of papers beyond what is expected using retrospective analysis.

## 3.1 Filtering and representation

Our first step assumes that a query has been submitted to a retrieval system that relies on high-quality content such as a Digital Library and that results have been retrieved. One specific example relevant to our domain would be the submission of a query to PubMed. Thus, the results of such a query are the input to our first step. In our first step, we perform two tasks: firstly, we filter some results of the query beyond what a retrieval system will typically do. In particular, we focus on assessing for each document whether a claim linking the two entities of the query exist. Whenever there is an identification of some claim in a document, we consider that document as part of our next steps. In other words, we focus on assessing that a document is relevant if it contains a scientific claim related to the query.

To distinguish sentences that correspond to our definition of a claim we rely on domain knowledge from the medical domain because our corpus of study comes from PubMed. Therefore, we restrict our current efforts to a subset of the relations from the Unified Medical Language System (UMLS)[1] . We used in our work SemmedDB [18] a database that contains semantic predications in the form of subject-predicate-object triples extracted from PubMed research papers. As discussed previously in [35], the tool used to extract the predicates of SemmedDB detects all predicates concerning pharmacogenomics (e.g., 'affects', 'augments', 'disrupts'), substance interactions (e.g., 'interacts-with', 'inhibits', 'stimulates'), genetic etiology of disease (e.g., 'associated-with', 'causes', 'predisposes') and clinical medicine (e.g. 'treats', 'diagnoses', 'process-of').

We manually inspected the predicates covered in the SemmedDB, and we found that some of them such as 'is-a' or 'location-of' do not fit our definition of claims. After carefully analyzing the database, we decided to use the following: "affects", "associated-with", "causes", "inhibits", "prevents", "process-of", and "treats" Moreover,

[1] For more information about UMLS see https://www.nlm.nih.gov/research/umls/

for each relation we included its negation. In other words, for "causes" there is a relation named "neg-causes".

We used the content of the database to develop an automatic tool that given a claim in natural language form, delivers the relation that exists in the claim. We called this the 'semantic orientation' of claims. We can formalize the Semantic Orientation of Claims as a classification problem as follows:

Semantic Orientation of Claims Problem. *Formally, we want to learn a function $f : R^n \rightarrow \{1, ...k\}$. When $y = f(x)$, the model assigns an input described by vector $x$ to a category identified by a numeric code $y$. The vector $x$ in our case is a vector representation of some claim.*

Different alternatives exist to learn such a function in a data-driven fashion, and we will explore some of them in this work. In the following section, we introduce each model used in this task and provide details of their performance in the experimental section of the paper (Section 4.1).

*Models.* We trained different machine learning models in our quest to have an automatic tool that given a claim in natural language text can output the semantic orientation of the corresponding entities. We decided to use only Deep Learning approaches for our classification task. Moreover, as a baseline, we used FastText [16] an approach used for text classification tasks that have shown good performance on different datasets see [16]. Next, we briefly describe the deep learning models used.

*LSTM Stack.* The Long-Shot Term Memory Network (LSTM) is one of the most widely used recurrent neural networks, and it was introduced by Hochreiter and Schmidhuber [15]. It has been applied to solve different problems mostly for several time-series or sequence data. Recently, researchers have used LSTMs to build sentence embeddings for information retrieval [30], or the translation of sentences into different languages [1]. Training and optimizing such networks is a complex task and requires much computational power. Fortunately, a recent extensive empirical study [13] has shed the light of different variations of LSTM's performance and tuning of its parameters.

The architecture that we used in this work consists of two consecutive LSTM layers followed by a stack of fully connected layers for prediction. We applied a single dropout of 0.50 to force the model into learning more general abstractions from the data. The LSTM layers consume most of the computational power as they account for over 90% of the total weights of the model.

Additionally, since the LSTMs are recurrent layers, they take much longer to train than a simple fully connected layer with the same amount of weights. The reason is that any recurrent network needs a particular variant of the optimization algorithm called Backpropagation through Time. This algorithm needs to compute the gradient not only once, but multiple times to allow the model to learn abstractions through the time domain. For a more in-depth introduction to this fascinating topic, please see [12].

*Convolutional Neural Networks.* Convolutional Neural Networks (CNNs) have excelled in computer vision in different tasks [4, 20, 40] and are responsible for the renaissance of interest in neural networks [23]. Recently, they have also shown to achieve state of the art results on sentence classification. For instance, the work of Kim [19] proposed not only to look at one specific amount of word vectors but also to use multiple context sizes in parallel. The idea here is as Zhang, and Wallace [42] demonstrated: to capitalize on the distributed representation of word embeddings. Zhang and Wallace provided practical guidelines of what can be achieved using a CNN for text classification tasks. Goodfellow et al. [12] have emphasized that three essential ideas motivate the use of CNNs in different machine learning tasks, including text classification: sparse interactions, parameter sharing, and equivariant representations. For text classifications task, sparse interactions allow for automatically learning linguistic n-grams patterns, i.e. without manual feature engineering; parameter sharing influences computation storage requirements, and equivariant representation allows for robustness in the patterns learned regarding of the position in the sentence.

To better understand the CNN model, in this section we provide the necessary background. A CNN is a specialized kind of neural network for processing data that has a grid-like topology. Examples include time-series data, which can be thought of as a 1-D grid taking samples at regular time intervals, and image data, which can be thought of as a 2-D grid of pixels [12]. Recently, they have also been applied to text data. Here we follow the discussion from [42] to explain how sentences can have a matrix-like representation so they can be of any use for CNN's. We begin with a tokenized sentence, which we then convert to a sentence matrix. In this matrix, each row is a word vector representation of each token. These word vector representations can be obtained from models such as word2vec [28] or Glove [32]. We denote the dimensionality of the word vectors by $d$. If the length of a given sentence is $s$, then the dimensionality of the sentence matrix is $s \times d$. Suppose that there is a filter matrix $w$ with region size $h$; $w$ will contain $h \times d$ parameters to be estimated. We denote the sentence matrix by $A \in R^{s \times d}$, and use $A[i : j]$ to represent the sub-matrix of $A$ from row $i$ to row $j$. The output sequence $o \in R^{s-h+1}$ of the convolution operator is obtained by repeatedly applying the filter on sub-matrices of $A$:

$$o_i = w \cdot A[i : i + h - 1]$$

where $\cdot$ is the dot product between the sub-matrix and the filter (a sum over element-wise multiplications). We add a bias term $b \in R$ and an activation function $f$ to each $o_i$ to induce the feature map $c \in R^{s-h+1}$ for this filter:

$$c_i = f(o_i + b)$$

Next, a pooling function is applied to each feature map to get a fixed-length vector. Researchers use Max-pooling [2], which extracts the maximum value for each feature map. Then the outputs generated from each filter map can be concatenated into a fixed-length feature vector, which is then fed through a softmax function to generate the final classifications. Usually, dropout [14] is applied as regularization. Optimization is performed using Stochastic Gradient Descend and back-propagation [36]. We present the details of these experiments in Section 4.1.

## 3.2 Organization

The primary goal of our second step is to cluster the result set of the query. To accomplish our first task, we build on the idea of modeling the co-occurrence of claims and entities. For this task, we represent each document as a bag of words [27]. Our vector space model consists of each document represented as a vector of weighted frequencies of its tokens. Tokens in our setting are not only words but also entities relevant to our domain such as drugs and diseases.

In particular, we model this interaction as a bipartite Graph model and apply Co-clustering of documents and tokens. For self-containment in what follows we use and adapt the notation from the original work on Co-clustering [5]. Thus, formally, we have $Docs = \{doc_i, i = 1, ..., m\}$: a set of $m$ documents and $Tokens = \{token_j, j = 1, ..., n\}$: a set of $n$ tokens

Let us start with the following definitions:

Definition 1. *A graph* $G = (V, E)$ *is a set of vertices* $V = \{1, .., |V|\}$ *and a set of edges* $\{i, j\}$ *each with edge weight* $E_{ij}$. *The adjacency matrix M of a graph is defined by*

$$M_{ij} = \left\{ \begin{array}{ll} E_{ij}, & \textit{if there is an edge \{i,j\}} \\ 0, & \textit{otherwise} \end{array} \right\}$$

Definition 2. *Cut of a graph. Given a partition of the vertex set* $V$ *into multiple subsets* $V_1, ..., V_k$, *the cut of the graph is:*

$$cut(V_1, ...V_k) = \sum_{i<j} cut(V_i, V_j)$$

*where*

$$cut(V_1, V_2) = \sum_{i \in V_1, j \in V_2} M_{ij}$$

We now consider the bipartite graph model for representing our claims collection.

Definition 3. *Bipartite Graph. An undirected bipartite graph is a triple* $G = \{Docs, Tokens, E\}$ *where* $Docs$ *and* $Tokens$ *are two sets of vertices and E is the set of edges each with weight* $a_{ij}$. *The weights indicate an association between claims and entities. One possibility is to use simple entity frequencies.*

Why a bipartite graph model? The intuition: we would like to organize the result set of a query such that documents with claims of the same group are more related to one subset of tokens compare to the other subsets of tokens. We built on work [5] and based our proposed solution considering the following observation:

***The duality of tokens and documents clustering***. Token clustering induces document clustering while document clustering induces token clustering.

Given disjoint document clusters $Dl_1, ..., Dl_k$, the corresponding token clusters $Tn_1, ..., Tn_k$ may be determined as follows: a given token $t$ belongs to the token cluster $Tn_m$ if its association with the document cluster $Dl_m$ is greater than its association with any other claim cluster. Using the proposed graph model, a natural measure of the association of a token with a document cluster is the sum of the edge-weights to all documents in the cluster. Thus, each of the token clusters is determined by the document clustering. Similarly, given token clusters $Tn_1, ...Tn_k$, we can find the induced document clustering in a similar fashion. Thus, we can observe the recursive nature of this characterization: document clusters determine token clusters, which in turn determine document clusters.

Thus, the "best" token and document clustering would correspond to a partitioning of the graph such that crossing edges between partitions have a minimum weight. Using Definition 2, we can further formulate the dual clustering of document and tokens as a solution of the minimization of graph cut:

$$cut(Tn_1 \cup Dl_1, ...Tn_k \cup Dl_k) = min\, cut(V_1, ...V_k)$$

where $V_1, ...V_k$ is a $k-$partioning of the bipartite graph.

***Spectral Co-clustering***. The idea of modeling documents with claims and tokens with a bipartite graph motivates us to use spectral graph theory to induce the co-clusters [5]. Spectral graph clustering uses the eigenvalues of the adjacency matrix to map the original relationships of co-occurrence onto a new space to project each claim and entity. After the projection, the documents and tokens are simultaneously partitioned into disjoint clusters with minimum cut optimization. The solution to the partitioning of the bipartite graph studied in [5], uses the $k$ left and right singular vectors to find the new partition space.

### 3.3 Characterization

In our last step, we use the organization of the clusters from Co-clustering and proceed to characterize each cluster. The main characterization involves finding and annotating whether 'controversy' or 'diversity' exist in a given cluster. For instance, if we have in the cluster $Cl_1$ a document $d_1$ with a claim such as 'drug X alleviates disease Y' and a document $d_2$ with a claim such as 'drug X induces disease Y', then we annotate the cluster as having 'controversy'. To accomplish our goal of identifying 'controversy', we focus on the identification of direct negative semantic orientation of claims. For instance, 'affects' versus 'neg_affects'. Thus, we annotate a cluster as exhibiting 'controversy' if we can find at least two claims with the same pairs of entities that have opposite semantic orientation.

Diversity, on the other hand, involves another core aspect. It has to do with the existence of different semantics of claims that are pointing at the same entities. The idea here is to capture the presence of scientific claims that may deserve further investigation because they have different semantics although they share almost the same context. We hypothesize that these two claim-based aspects can provide a 'semantic' indicator of the significant growth of future papers.

## 4 EXPERIMENTAL SETUP

In this section, we provide details of the experimental setting of our proposed approach to evaluate its scope and limitations.

### 4.1 Semantic Orientation of claims

We now look at our first step towards the characterization of the clusters. In other words, we deal with the problem of automatically detecting the semantic orientation of a claim. By semantic orientation, we mean the type of association that exists between the entities in a given claim. For instance, the claim: "Diabetes was induced by alloxan injection"[2] corresponds to some claim, where a causal association exists between Diabetes and alloxan injection. According to our definition of claims we focus on the following associations that

[2] Contained in paper with pmid 21629542

are part of the semantic relations in the Unified Medical Language System (UMLS): "affects", "associated-with", "causes", "inhibits", "prevents", "process-of", "treats" as well as its corresponding negative counterparts, e.g., "neg-affects", "neg-associated-with". Thus, we limit our work to fourteen different semantic types that fit in our definition of claims.

*Data description.* We randomly sampled from SemmedDB database 10K sentences per each of the semantic relations that we previously described. However, for each of the negative classes, e.g., "neg-affects", we could only sample 2000 per class because there were not too many of them in the database. Out of these samples, we split our data into two sets as follows: training (80%) and testing (20%). We used weighted average F1-score as a metric to measure the performance of each model.

We considered the weighted average because it considers the differences in the class frequencies between the positive semantic relations and their opposites. As a result, we can more objectively assess the overall performance of the models.

*Preprocessing.* We tokenized each sentence and used in our work word embeddings to represent each sentence as a sequence of its embedding words. Using word embeddings allowed the models to account for multiple synonyms and expressions with the same meaning. In our work, we used the word vector representations learned with the Word2Vec algorithm by Mikolov et al. [28]. In a nutshell, word embeddings pack more information into far fewer dimensions. Researchers have shown [42] two approaches to take advantage of word embeddings for classification tasks: 1) learn word embeddings jointly with the problem at hand and 2) use embedding vectors from a pre-computed embedding space that might exhibit useful properties (captures general aspects of language structure). We will show in our tailored models the effect of both approaches.

*Parameter details.* One of the challenges of the deep learning models that we used in our work is to find the right combination of parameters that can solve the problem optimally. In our case, we mainly iterate: we began with a simple small model, gradually increased its capacity and we kept doing that until the validation score no longer improved after three consecutive epochs. In addition to that, we also used dropout [14] of 50 as regularization to avoid overfitting. Furthermore, 80% of the data was used for model training and 20% for testing.

*Discussion of the results.* In Table 1 we show the results of our experiments. 'FastText' refers to the model learned using the FastText algorithm [16]. 'CNN-Static' refers to the model using Convolutional Neural Networks using word embeddings pre-trained on PubMed as described in [34]. 'CNN-Dynamic' is the model that learns the embeddings as part of the classification task. 'LSTM-Stack-Static' refers to the model trained using a two LSTM stack using the pre-trained word embeddings mentioned before. Finally, the 'LSTM-Stack-Dynamic' refers to the model that learns the word embedding as part of the classification task.

We can observe that all models that had to learn the representation of the word vectors as part of the classification task performed only slightly better than the models that used the pre-trained word embeddings. This means that for our problem setting the computationally expensive process is not worth it. Our finding supports the idea that using transfer learning for this task results in models that can provide comparable results regarding more time-consuming models that do not use pre-trained word embeddings.

**Table 1: Results Semantic Orientation Task**

| Model | Weighted F1 |
|---|---:|
| FastText | 0.67 |
| CNN-Static | 0.70 |
| CNN-Dynamic | 0.72 |
| LSTM-Stack-Static | 0.69 |
| LSTM-Stack-Dynamic | 0.71 |

**Table 2: Results of the CNN-Dynamic Model**

| Class | Precision | Recall | F1 | samples |
|---|---:|---:|---:|---:|
| AFFECTS | 0.62 | 0.65 | 0.63 | 2000 |
| ASSOC_WITH | 0.62 | 0.68 | 0.65 | 2000 |
| CAUSES | 0.76 | 0.83 | 0.79 | 2000 |
| INHIBITS | 0.81 | 0.77 | 0.79 | 2000 |
| NEG_AFFECTS | 0.69 | 0.74 | 0.71 | 400 |
| NEG_ASSOC_W | 0.65 | 0.75 | 0.70 | 400 |
| NEG_CAUSES | 0.89 | 0.85 | 0.87 | 400 |
| NEG_INHIBITS | 0.85 | 0.81 | 0.82 | 400 |
| NEG_PREVENTS | 0.84 | 0.89 | 0.86 | 400 |
| NEG_PROC_OF | 0.69 | 0.82 | 0.75 | 400 |
| NEG_TREATS | 0.74 | 0.70 | 0.72 | 400 |
| PREVENTS | 0.84 | 0.87 | 0.85 | 2000 |
| PROCESS_OF | 0.61 | 0.64 | 0.62 | 2000 |
| TREATS | 0.76 | 0.51 | 0.61 | 2000 |
| Weighted AVG | 0.73 | 0.72 | 0.72 | 16800 |

To our surprise, the CNN based-models outperform the LSTM Stack models regardless of the existence of transfer learning. Given that LSTMs are very often used in sequence data and usually provide more accurate results, we were surprised that in our case they did not outperform their CNN counterparts. It seems that for our task CNNs can find patterns in the data that have a higher impact that the sequence modeling approach of LSTMs.

In Table 2 we show the performance of the best model per class. We can see that some classes were challenging to recognize by the model. For instance class 'TREATS' has the lowest F1 score due to its low recall value. An analysis of the data revealed that 'TREATS' requires domain knowledge, e.g., other medical entities to assess the presence of this type of association more accurately. This could probability boost our best model, and it will be part of our future work. The performance on the negative classes such as 'NEG-CAUSES' with 0.87 of F1 score reflects the impact of our decision of keeping all types of negation and hedges such as may, might, instead of treating them as stop words.

## 4.2 Semantic cluster coherence

In this section, we examine the qualitative properties of the co-clustering approach compared to the results provided by K-Means clustering. First, we explain what we mean by cluster coherence, and we define our hypothesis. Afterward, we will empirically prove our hypothesis and, in this context, we describe our ground-truth evaluation corpus, followed by experimental set-up as well as our implementation decisions.

*4.2.1 Cluster coherence.* As we have already mentioned, the algorithm co-clustering is based on the grouping of rows and columns. The columns represent our pharmaceutical entities that are either a drug or a disease. In this context, our coherence-hypothesis is: If pharmaceutical entities that often co-occur together in documents are grouped with co-clustering, then we should have more meaningful clusters than those obtained by a more conventional approach such as K-Means. Here, meaningful means that the entities of a cluster are semantically similar to each other. This is particularly important since increasing coherence makes it easier to identify an intrinsic semantic relationship within a cluster. How can semantic similarity be determined for pharmaceutical entities?

In order to compare the semantic similarity of pharmaceutical entities and to evaluate the semantic quality of an entity cluster, each entity (active substance/disease) needs a unique class label. The most common pharmaceutical classification systems, such as the Anatomical Therapeutic Chemical (ATC) Classification System[3], Medical Subject Headings (MeSH) Trees[4] or the American Hospital Formulary Service (AHFS) Pharmacologic-Therapeutic classification[5], are suitable sources for these labels. In the following investigations, we use these classification systems to determine entity labels.

*4.2.2 Experimental setup.*

*Corpus.* We crawled documents from PubMed. PubMed[6] currently with more than 28 million document citations is the largest and most comprehensive digital library in the biomedical field. Since full-text access is not available for the most publications, we used only abstracts for our evaluation corpus. More abstracts per entity (active ingredient) allow us to perform a retrospective analysis that spans a more extended period (20 years). Thus, we decided to use a minimum of the 1000 most relevant abstracts for each entity (active substance).

Moreover, we relied on the relevance weighting of PubMed’s search engine. Diseases, as well as drugs, often consist of several words (e.g., diabetes mellitus). To account for such cases, we 1) recognize the entities in documents and 2) place a unique identifier at the entity’s position in the text. For the recognition of the entities, we used PubTator[7], a tool which can recognize pharmaceutical entities and returns a MeSH-Id for each of them.

*Query entities.* As query entities for the evaluation, we randomly selected 350 drugs from the DrugBank[8] collection, which is a 10% sample of all approved drugs. Thus, our final document set for evaluation contains $\sim$ 2.5 million abstracts for 350 drugs and a period between 1900-01-01 and 2018-01-01.

*Entity labels.* Drugs can be grouped according to different properties and thus semantics. For example, drugs can be grouped according to anatomical and therapeutic properties (e.g., as with ATC) and in other cases according to pharmacological properties (e.g., as with AHFS). Therefore, we use the different classification systems ATC, AHFS, and MESH to determine labels for drugs in the best possible way. These classification systems have a hierarchical structure, and so for example, the drug aspirin (Table 3) contains the ATC label “N02BA01”. Where, in ATC, “N” stands for the nervous system and “N02” for analgesics. Since the labels can become too fine-granular, we select only the highest level as a possible class label. Besides, a drug can have several classes and in different classification systems. Since we have to determine a unique label for each active ingredient, we use the majority label approach described in [41] to determine a unique label for each drug. For diseases, we use the same approach, but here we rely on MeSH-Trees only to determine the labels.

**Table 3: Example classes in different classification systems for the drug ‘Aspirin’**

| Classification System | Assigned Classes |
|---|---|
| ATC | N02BA01, B01AC06, C10BX05 |
| AHFS | 28:08:04:24 |
| MeSH Trees | D02.455.426.559.389.657.410.595.176 |

*4.2.3 Experiment implementation and parameter settings.* In the following, we describe the steps we perform for our evaluation:

(1) Historical Corpus Generation: One of our core steps in this paper is to create semantically coherent-entity clusters using co-clustering, and to annotate them using our proposed claim-based characterization. To investigate whether such a forecast makes sense at all, we use a retrospective analysis (Section 4.3). For this purpose, we generate a historical corpus from our corpus (period 1900-01-01 to 2018-01-01) for the period 1900-01-01 to 1998-01-01. This corpus is the basis for the following retrospective investigations.
(2) Query Candidate Generation: For each drug-disease query, we determine in the historical corpus the number of documents (abstracts) in which the pair occurs together. If the number of documents is too small, clustering makes less sense. Therefore, we determine all pairs which occur in at least 200 publications together. We use this procedure to determine a total of 214 query candidates.
(3) Clustering: We cluster the query candidates documents determined in the last step using the co-clustering approach. For comparison, we also cluster the documents with K-Means. We determine the optimal number of clusters per query candidate using the cluster silhouette, which is calculated by a distance-based coherence.
(4) Entity Pair Generation: Next, we extract all active ingredients and diseases from each cluster and generate all possible combinations between an active ingredient and a disease.

[3] https://www.whocc.no/atc_ddd_index/
[4] https://www.nlm.nih.gov/mesh/intro_trees.html
[5] http://www.ahfsdruginformation.com/ahfs-pharmacologic-therapeutic-classification/
[6] https://www.ncbi.nlm.nih.gov/pubmed/
[7] https://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/PubTator/
[8] https://www.drugbank.ca/

**Table 4: Comparison between Co-clustering and K-Means**

| Clustering Approach | Precisin | Recal | F1-Score |
|---|---|---|---|
| Co-Clustering | 0.37 | 0.62 | 0.46 |
| K-Means | 0.54 | 0.14 | 0.23 |

(5) Entity Label Identification: Finally, we determine a unique label for each active ingredient and disease. To determine a meaningful label, we use the approach described in [41].

To continue with our experimental evaluation, we first have to determine what quality criteria an entity-centric clustering approach should meet for dynamically creating coherent semantic clusters. Thus, the following criterion should be fulfilled:

*Semantic Cluster Coherence.* Elements of a cluster such as drugs and diseases that belong respectively to the same class should also be grouped in the same cluster. From a user perspective, this facilitates the thematic interpretation of the individual clusters as well as the semantic differentiation from other clusters. Moreover, this semantic differentiation is simplified if there is a moderate and thus manageable number of clusters. The semantic cluster coherence can be evaluated using the F1 score.

*4.2.4 Experimental Evaluation.* In our evaluation, we compare the results of a co-clustering with the results achieved with K-Means. For comparison, we calculate the clusters Precision, Recall and F1-Score with the approach described in [27]. Here, we calculated the F1 score based on the AVG precision and the AVG recall. The results are presented in Table 4.

We can observe that co-clustering leads to generally better results regarding F1 score. Therefore, co-clustering leads to more semantically coherent clusters compared to K-Means.

K-Means has a low recall, suggesting that entities with the same class are more likely to be distributed across many clusters. This means that the semantic differences between clusters will be more difficult to distinguish. On the other hand, K-Means has a higher AVG precision. This may be due to that K-Means leads to many clusters with smaller sizes. Therefore, the next step was to analyze the number of clusters. Figure 3 shows the comparison between the numbers of clusters for the two clustering approaches. On average, co-clustering leads to 3.5 clusters, while K-Means leads to 10 clusters.

Are the two sets of cluster sizes significantly different? We tested this with a Welch two sample t-test at a confidence interval of 99%. With a p-value < 2.2e-16 we are below the threshold of 0.05 and therefore the two sets are significantly different.

In conclusion, we can say that co-clustering not only leads to a semantically more coherent grouping of pharmaceutical entities but also that the average number of clusters is moderate. Therefore, from a user perspective, a co-clustering is more accessible to interpretation in comparison to K-Means.

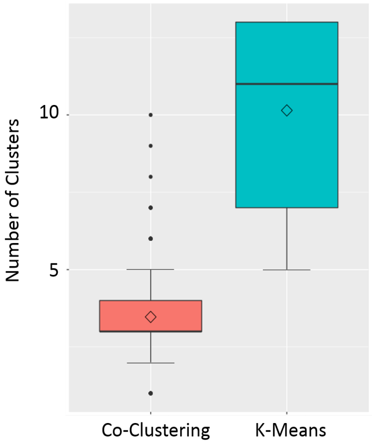

**Figure 3: Cluster size comparison between Co-Clustering and K-Means where the rhombus represents the mean**

## 4.3 Retrospective Analysis

In this section, we evaluate the merits of our proposed characterization of the clusters. In particular, the clusters computed using the Co-clustering algorithm because of the findings that we reported in the previous section. Thus, for each cluster of each query, we annotate the semantic orientation of each claim using our deep learning tailored model that we have already discussed in Section 4.1. Afterward, we compute the following attributes per each cluster: 1) contradiction: the number of contradictions that we can find. We count as a contradiction the presence of a semantic orientation of claims such as ‘treats’ and ‘neg-treats’ in the same cluster; 2) diversity: here we compute the different number of semantic orientations of the claims that we can find per cluster.

Because our proposed characterization aims at providing an overview that helps users to formulate, discover or put into context scientific claims that may need more research efforts, we decided to evaluate our characterization using retrospective analysis.

The idea behind is to apply our overall approach and characterize the clusters as we described before using information of 1998 and then, predict the number of papers in 2008 and 2018. To do so, we query PubMed using the entity pairs of each claim found per cluster and count the number of papers in 2008 and 2018. Thus, we introduce here a classification task aiming at predicting if a given cluster in 1998 with its corresponding contradiction and diversity attributes will lead to some papers beyond what might be expected, e.g., higher than the overall mean. In Table 5 we show the statistics of the data we use to evaluate our proposed characterization. In the table, ‘Significant’ means the number of clusters whose characterization corresponds to some papers greater than the mean for the corresponding year. ‘Nonsignificant’ are those clusters less than or equal to the mean. ‘Data_2008’ refers to the dataset that considers up to 2008 as the ground truth of the number of papers. Similarly, ‘Data_2018’ corresponds to the counting of papers up to 2018.

Given the statistics of the dataset, we will report weighted F1 to measure the performance of the prediction task. We used a Support Vector Machine (SVM) [3] to evaluate the merits of our proposed characterization of clusters.

*Preprocessing.* As a preprocessing step, we standardized our three attributes by removing the mean and scaling to unit variance. This preprocessing step is particularly crucial for SVMs because if an

**Table 5: Summary of the Datasets Used to Evaluate the Characterization of the Clusters**

| Dataset | Significant | Nonsignificant |
|---|---|---|
| Data_2008 | 135 | 416 |
| Data_2018 | 132 | 419 |

**Table 6: Summary of models performance using weighted F1 score to evaluate the Characterization of the Clusters**

| Model | Data_2008 | Data_2018 |
|---|---|---|
| SVM+Contradictions | 0.83 | 0.81 |
| SVM+DiffSemantics | 0.81 | 0.79 |

attribute has a variance several orders of magnitude larger than others, the algorithm will not learn from the other attributes correctly because the attribute with larger variance will dominate the objective function.

*Models.* We trained four SVM models using the Sckit-Learn machine learning library in Python. We refer to the models as follows: 'SVM+DiffSemantics' is a model using an SVM that used the different semantic orientations of the claims. The other model named 'SVM+Contradictions' is an SVM trained considering the number of contradictions found by our algorithm. Notice also that both models used as a control variable the 'size' of the clusters in 1998.

All the models were trained using a random stratified sample of the data (60%), leaving the rest for evaluation. We show in Table 6 the results of each model for the two datasets.

*Discussion of the results.* To put our findings into the right context, please notice that a model that in addition to 'size' of a cluster in 1998 uses our proposed characterization gets a 5% gain improvement in weighted F1 score in 2008 and a 3% gain in 2018. We can observe that there are small differences in performance between the models in both datasets. Considering that we only used documents in 1998 to automatically generate our 'controversy' and 'diversity' features, the results indeed already look very promising. In particular, this is because our claim-based characterization can provide the semantics behind the significant increase of papers for some claims. We can also see for instance that the best model only uses the 'controversy' attribute. It seems to fit the idea that the more controversy exhibited in the clusters, the more likely it is that more papers will further investigate these issues. Our finding complements the idea introduced [25]: researchers argued that the existence of speculative claims could form the basis of new hypotheses. We can push this idea further and say that in addition to speculation, controversy is a factor that can help to characterize clusters and ease hypothesis generation to the interested user. Finally, as expected, we can also observe that both models obtained slightly less F1 score in 2018. This means that the more we look into the future, the less prediction power we can obtain.

## 5 CONCLUSIONS AND FUTURE WORK

We introduced a novel characterization of the results sets of a query using a key aspect of scientific papers: claims. Claims are natural language sentences in a scientific paper that expresses a relationship between two entities. In particular, how one of them affects, manipulates, or causes the other entity. When mining clusters with respect to the relationships expressed in the respective claims, our approach leveraged co-clustering technology to characterize the clusters by two fundamental properties: contradiction and diversity of scientific claims.

We tested our claim-based characterization in an extensive retrospective analysis using more than 200 queries on a corpus build from biomedical literature in PubMed. To quantitatively evaluate the benefits of our approach, we validated it in the challenging, yet meaningful task of predicting cases where a significant increase of publications will appear in two different moments: 10 and 20 years after our characterization. Our findings support the potential of our approach to providing an innovative service that can help users in the difficult task of finding claims in need of further investigation. We achieved 80% of F1 score with our proposed attributes showing the potential of our claim-based characterization of the clusters.

As possible directions of future work, we intend to improve our modeling of 'controversy'; we only scratched the surface of 'controversy' using direct negations of the semantic orientation. However, certain contextual conditions could also imply controversy, for instance 'causes' and 'alleviates' when linking drug and diseases could also be considered controversial in specific contexts, and our experts can help us to make such assertions.

Finally, our work provides an example for information providers in need of new intelligent services aiming at empowering users to take advantage of the richness of knowledge within our collections. Indeed, as our collections grow, one can argue that there is more gold than ever as well as precious gems within the manuscripts, but we owed to our users to provide them with intelligent services to solve complex information needs such as hypothesis generation.

## ACKNOWLEDGMENTS

Part of this research has been funded within the project: PubPharm: Specialized Information Service for Pharmacy (Gepris 267140244) supported by the German Research Foundation (DFG).

## REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *3rd International Conference on Learning Representations*. arXiv, San Diego, California, 1–15. https://doi.org/10.1146/annurev.neuro.26.041002.131047

[2] Y-Lan Boureau, Jean Ponce, and Yann LeCun. 2010. A Theoretical Analysis of Feature Pooling in Visual Recognition. In *Proceedings of the 27th Interrnational Conference on Machine Learning (ICML-2010)*. Omnipress, Haifa, Israel, 111–118. http://www.ece.duke.edu/~lcarin/icml2010b.pdf

[3] C J C Burges. 1998. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery* 2, 2 (1998), 121–167. /papers/Burges98.ps.gz

[4] Dan Claudiu Cireşan, Ueli Meier, Luca Maria Gambardella, and JÃijrgen Schmidhuber. 2011. Convolutional neural network committees for handwritten character classification. In *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*. IEEE, Beijing, China, 1135–1139. https://doi.org/10.1109/ICDAR.2011.229

[5] Inderjit S. Dhillon. 2001. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '01*. ACM, San

Francisco, California, 269–274. https://doi.org/10.1145/502512.502550

[6] Andy Extance. 2018. How AI technology can tame the scientific literature. *Nature* 561, 7722 (9 2018), 273–274. https://doi.org/10.1038/d41586-018-06617-5

[7] Thomas George and Srujana Merugu. 2005. A scalable collaborative filtering framework based on co-clustering. In *Proceedings - IEEE International Conference on Data Mining, ICDM*. IEEE Computer Society, Washington, DC, USA, 625–628. https://doi.org/10.1109/ICDM.2005.14

[8] José María González Pinto and Wolf-Tilo Balke. 2017. Can Plausibility Help to Support High Quality Content in Digital Libraries?. In *TPDL 2017 21st International Conference on Theory and Practice of Digital Libraries*. Springer International Publishing, Thessaloniki, Greece., 169–180.

[9] José María González Pinto and Wolf-Tilo Balke. 2017. Result Set Diversification in Digital Libraries Through the Use of Paper's Claims. In *International Conference on Asian Digital Libraries*. Springer, Springer, Bangkok, Thailand, 225–236.

[10] José María González Pinto and Wolf-Tilo Balke. 2018. Assessing plausibility of scientific claims to support high-quality content in digital collections. *International Journal on Digital Libraries* 19, 59 (10 2018), 1–14. https://doi.org/10.1007/s00799-018-0256-8

[11] José María González Pinto and Wolf-Tilo Balke. 2018. Scientific Claims Characterization for Claim-Based Analysis in Digital Libraries. In *TPDL 2018 - 22nd International Conference on Theory and Practice of Digital Libraries*, Eva Méndez, Fabio Crestani, Cristina Ribeiro, Gabriel David, and JoÃčo Correia Lopes (Eds.). Springer International Publishing, Porto, Portugal, 257–269.

[12] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press, Cambridge, MA. 775 pages. https://doi.org/10.1038/nmeth.3707

[13] Klaus Greff, Rupesh K. Srivastava, Jan Koutnik, Bas R. Steunebrink, and Jurgen Schmidhuber. 2017. LSTM: A Search Space Odyssey. *IEEE Transactions on Neural Networks and Learning Systems* 28, 10 (2017), 2222–2232. https://doi.org/10.1109/TNNLS.2016.2582924

[14] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. , 18 pages. http://arxiv.org/abs/1207.0580

[15] Sepp Hochreiter and J Urgen Schmidhuber. 1997. Long Short-Term Memory. *Journal of Neural Computation* 9, 8 (1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[16] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, Vol. 2. Association for Computational Linguistics, Valencia, Spain, 427–431. http://arxiv.org/abs/1607.01759

[17] Halil Kilicoglu, Graciela Rosemblat, and Thomas C. Rindflesch. 2017. Assigning factuality values to semantic relations extracted from biomedical research literature. *PLOS ONE* 12, 7 (07 2017), 1–20. https://doi.org/10.1371/journal.pone.0179926

[18] Halil Kilicoglu, Dongwook Shin, Marcelo Fiszman, Graciela Rosemblat, and Thomas C. Rindflesch. 2012. SemMedDB: A PubMed-scale repository of biomedical semantic predications. *Journal of Bioinformatics* 28, 23 (2012), 3158–3160. https://doi.org/10.1093/bioinformatics/bts591

[19] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1746–1751. https://doi.org/10.3115/v1/D14-1181

[20] Alex Krizhevsky, Ilya Sutskever, and Hinton Geoffrey E. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc., Harrah's Lake Tahoe, USA, 1âĂŞ9. https://doi.org/10.1109/5.726791

[21] Bongjune Kwon and Hyuk Cho. 2010. Scalable co-clustering algorithms. In *Algorithms and Architectures for Parallel Processing. ICA3PP 2010. Lecture Notes in Computer Science.*, Vol. 6081. Springer Berlin Heidelberg, Busan,Korea, 32–43. https://doi.org/10.1007/978-3-642-13119-6{_}3

[22] Peder Olesen Larsen and Markus von Ins. 2010. The rate of growth in scientific publication and the decline in coverage provided by science citation index. *Scientometrics* 84, 3 (2010), 575–603. https://doi.org/10.1007/s11192-010-0202-z

[23] Yann Lecun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521, 7553 (2015), 436–444. https://doi.org/10.1038/nature14539

[24] Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context Dependent Claim Detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics*. Dublin City University and Association for Computational Linguistics, Dublin, Ireland, 1489–1500. http://www.aclweb.org/anthology/C14-1141.pdf

[25] Marc Light, Xin Ying Qiu, and Padmini Srinivasan. 2004. The Language of Bioscience: Facts, Speculations, and Statements In Between. In *HLT-NAACL 2004 Workshop: BioLINK 2004, Linking Biological Literature, Ontologies and Databases*. Association for Computational Linguistics, Boston, Massachusetts, USA, 17–24. https://www.aclweb.org/anthology/W04-3103

[26] Marco Lippi and Paolo Torroni. 2016. Argumentation Mining: State of the Art and Emerging Trends. *ACM Transactions on Internet Technology* 16, 2 (2016), 1–10. https://doi.org/10.1145/2850417

[27] Christopher D. Manning and Prabhakar Raghavan. 2009. An Introduction to Information Retrieval. https://doi.org/10.1109/LPT.2009.2020494

[28] Tomas Mikolov, Greg Corrado, Kai Chen, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of the International Conference on Learning Representations (ICLR 2013)*. arXiv, Scottsdale, Arizona USA, 1–12. https://doi.org/10.1162/153244303322533223

[29] Meenakshi Nagarajan, Jacques J. Labrie, Sam Regenbogen, Christie M. Buchovecky, Curtis R. Pickering, Linda Kato, Andreas M. Lisewski, Ana Lelescu, Houyin Zhang, Stephen Boyer, Griff Weber, Angela D. Wilkins, Ying Chen, Lawrence Donehower, Scott Spangler, Olivier Lichtarge, Benjamin J. Bachman, Ilya B. Novikov, Shenghua Bao, Peter J. Haas, MarÃŋa E. Terrón-Díaz, Sumit Bhatia, and Anbu K. Adikesavan. 2015. Predicting Future Scientific Discoveries Based on a Networked Analysis of the Past Literature. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*. ACM, Sydney,Australia, 2019–2028. https://doi.org/10.1145/2783258.2788609

[30] Hamid Palangi, Li Deng, Yelong Shen, Jianfeng Gao, Xiaodong He, Jianshu Chen, Xinying Song, and Rabab Ward. 2016. Deep Sentence embedding using long short-term memory networks: Analysis and application to information retrieval. *IEEE/ACM Transactions on Audio Speech and Language Processing* 24, 4 (2016), 694–707. https://doi.org/10.1109/TASLP.2016.2520371

[31] Leonidas Papachristopoulos, Michalis Sfakakis, Nikos Kleidis, Giannis Tsakonas, and Christos Papatheodorou. 2016. Exploiting Network Analysis to Investigate Topic Dynamics in the Digital Library Evaluation Domain. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries (JCDL '16)*. ACM, New York, NY, USA, 267–268. https://doi.org/10.1145/2910896.2925464

[32] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1532–1543. https://doi.org/10.3115/v1/D14-1162

[33] Vinodkumar Prabhakaran, William L Hamilton, Dan McFarland, and Dan Jurafsky. 2016. Predicting the Rise and Fall of Scientific Topics from Trends in their Rhetorical Framing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, 1170–1180. https://doi.org/10.18653/v1/P16-1111

[34] Sampo Pyysalo, Filip Ginter, Hans Moen, Tapio Salakoski, and Sophia Ananiadou. 2013. Distributional Semantics Resources for Biomedical Text Processing. In *Proceedings of the 5th International Symposium on Languages in Biology and Medicine*. Database Center for Life Science, Tokyo, Japan, 39–43. http://bio.nlplab.org/pdf/pyysalo13literature.pdf

[35] Thomas C. Rindflesch and Marcelo Fiszman. 2003. The interaction of domain knowledge and linguistic structure in natural language processing: Interpreting hypernymic propositions in biomedical text. *Journal of Biomedical Informatics* 36, December 2003 (2003), 462–477. https://doi.org/10.1016/j.jbi.2003.11.003

[36] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. 1986. Learning representations by back-propagating errors. *Nature* 323, 6088 (1986), 533–536. https://doi.org/10.1038/323533a0

[37] Angelo A Salatino, Francesco Osborne, and Enrico Motta. 2018. AUGUR: Forecasting the Emergence of New Research Topics. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries (JCDL '18)*. ACM, New York, NY, USA, 303–312. https://doi.org/10.1145/3197026.3197052

[38] Jonathan D Schoenfeld. 2013. Is everything we eat associated with cancer ? A systematic. *American Journal of Clinincal NUtrition* 97 (2013), 127–134. https://doi.org/10.3945/ajcn.112.047142.1

[39] Scott Spangler, Jeffrey N. Myers, Ioana Stanoi, Linda Kato, Ana Lelescu, Jacques J. Labrie, Neha Parikh, Andreas Martin Lisewski, Lawrence Donehower, Ying Chen, Olivier Lichtarge, Angela D. Wilkins, Benjamin J. Bachman, Meena Nagarajan, Tajhal Dayaram, Peter Haas, Sam Regenbogen, Curtis R. Pickering, and Austin Comer. 2014. Automated hypothesis generation based on mining scientific literature. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14*. ACM, New York, New York, USA, 1877–1886. https://doi.org/10.1145/2623330.2623667

[40] Yi Sun, Xiaogang Wang, and Xiaoou Tang. 2013. Deep convolutional network cascade for facial point detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, Portland, OR, USA, 3476–3483. https://doi.org/10.1109/CVPR.2013.446

[41] Janus Wawrzinek and Wolf Tilo Balke. 2017. Semantic facettation in pharmaceutical collections using deep learning for active substance contextualization. In *International Conference on Asian Digital Libraries*, Vol. 10647 LNCS. Springer, Bangkok, Thailand, 41–53. https://doi.org/10.1007/978-3-319-70232-2{_}4

[42] Ye Zhang and Byron Wallace. 2017. A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification. In *Proceedings of the The 8th International Joint Conference on Natural Language Processing*. Asian Federation of Natural Language Processing, Taipei, Taiwan, 253âĂŞ263. http://arxiv.org/abs/1510.03820