International Journal on Digital Libraries
https://doi.org/10.1007/s00799-018-0256-8

Assessing plausibility of scientific claims to support high-quality content in digital collections

José María González Pinto[1] · **Wolf-Tilo Balke**[2]

Received: 2 February 2018 / Revised: 8 October 2018 / Accepted: 13 October 2018

Abstract
This paper presents a formalization and extension of a novel approach to support high-quality content in digital libraries. Building on the concept of *plausibility* used in cognitive sciences, we aim at judging the plausibility of new scientific papers in light of prior knowledge. In particular, our work proposes a novel assessment of scientific papers to qualitatively support the work of reviewers. To do this, our approach focuses on the key aspect of scientific papers: claims. Claims are sentences found in empirical scientific papers that state statistical associations between entities and correspond to the core contributions of the papers. We can find these types of claims, for instance, in medicine, chemistry, and biology, where the consumption of a drug, a substance, or a product causes an effect on some other type of entity such as a disease, or another drug or substance. To operationalize the notion of plausibility, we promote claims as first-class citizens for scientific digital libraries and exploit state-of-the-art neural embedding representations of text and topic models. As a proof of concept of the potential usefulness of this notion of plausibility, we study and report extensive experiments on documents with scientific papers from the PubMed digital library.

Keywords Digital libraries · Information discovery · Plausibility · Empirical claims · Quality assessment

1 Introduction

For years, digital libraries have been a valuable and trustworthy source of information due to the *carefully curated* quality of their content. Since collections over the last decades are continuously growing with steeply increasing publication numbers, the main challenge to preserve content quality lies in the selection of new articles for inclusion in some collection. Today, peer review is the key to assess new articles and thus help digital libraries preserve high-quality content. However, with increasing numbers of publications reviewers are facing the problem of workload scalability: there is less and less time to do this valuable and necessary task. This has also been recognized by the community, for instance, in the work of Price et al. in [39], and while nobody has a perfect solution, there are many approaches to at least aid the process, such as expertise profiling, matching submissions with possible reviewers, or resolving paper biddings. In this work, we aim at supporting peer review not at the process level, but with a clear focus on document level. We aim at assessing a new scientific paper's plausibility at large in light of prior knowledge represented by some digital library collection. With this novel assessment, the question of how many reviewers a new paper actually needs can be adjusted by its respective degree of plausibility: the less plausible it is (i.e., the more its inclusion could potentially hurt a collection's consistency), the more reviewers might be needed to come to a correct decision.

The notion of plausibility in our work is based on the knowledge-fit theory from Cognitive Sciences recently studied by Connell et al. [13]. The theory states that human plausibility judgments consist of two major steps: first, a mental representation of current knowledge is built and secondly, an assessment examines how well a new piece of information fits all prior knowledge. Of course, this is very hard to decide in general settings. Thus, we will focus our work on a particular, yet often occurring type of document

✉ José María González Pinto
pinto@ifis.cs.tu-bs.de

Wolf-Tilo Balke
balke@ifis.cs.tu-bs.de

[1] Technische Universitat Braunschweig, Mühlenpfordtstrae 23, 2.OG Room 232, Braunschweig, Germany

[2] Technische Universitat Braunschweig, Mühlenpfordtstrae 23, 2.OG Room 237, Braunschweig, Germany

Published online: 28 October 2018

to provide first insights on the general feasibility of the idea: in particular, we focus on documents containing empirical claims in the sense of statistical associations between entities. Empirical claims are sentences that express some kind of association between two entities and in what way one affects the other. Indeed, our research shows that these simple types of claims can be found in many scientific papers: consider, for instance, medicine, chemistry, or biology, where the consumption of a drug, a substance, or a product, has an effect on some other type of entity such as a disease, or another drug or substance.

What makes precisely these types of claims so interesting are findings like those reported by nutritional researchers Schoenfeld et al. in [42]. Basically for 50 food ingredients, the researchers performed literature searches using PubMed[1] to obtain articles investigating the association between each ingredient and cancer risk. To their surprise, 80% of the ingredients were indeed related to cancer risk. However, what was even more surprising is that the authors found contradictory findings: out of 264 single-study assessments 191 (72%) associated the tested food with both, an increased ($n = 103$) and a decreased ($n = 88$) risk. What does that say about the concept of plausibility? How can we account for these types of situations and still provide a consistent instantiation of plausibility over digital libraries? Moreover, how many of these empirical types of claims are there in any case?

As opposed to the first two questions, the last one is easy to answer. To estimate this number, we used PubMed search interface. By using the PubMed's search interface, our query benefits from the PubMed's algorithm that uses machine learning to combine over 150 signals that are helpful to find relevant matching results. In summary, it automatically expands our query pattern to account for synonyms, MeSH terms, and medical terms. Our query pattern uses a similar linguistic query pattern analyzed by Ciccarese et al. [12]: *(help AND prevent) OR (lower AND risk) OR (increase OR increment AND risk) OR (decrease OR diminish AND risk) OR (factor AND risk) OR (associated AND risk)*. Even with this simple query on more than 28 million abstracts currently indexed by PubMed, we got more than 1 million articles in PubMed with empirical claims.

As one of our anonymous reviewers suggested, our Boolean query may contain some results that do not fit our definition of claims in scientific documents. Given that manually assessing more than 1 million documents would have not been affordable, we decided to use domain knowledge from the biomedical field to provide more profound insights regarding empirical claims. Our definition of empirical claims (see Sect. 3.1) states that an association between two entities must hold in a given sentence. In particular, how one of the entities influences, causes, manipulates, or affects the other. Thus, we decided to use the Semantic MEDLINE Database (SemmedDB) [24] to investigate the existence of empirical claims in PubMed. SemmedDB contains semantic predications in the form of subject–predicate–object triples extracted from the entire set of PubMed citations using the software tool SemRep [41]. SemRep is a specialized rule-based semantic interpreter of biomedical text. As discussed in [41], SemRep extracts predicates relating to pharmacogenomics (e.g., `AFFECTS`, `AUGMENTS`, `DISRUPTS`), genetic etiology of disease (e.g., `ASSOCIATED_WITH`, `CAUSES`, `PREDISPOSES`), substance interactions (e.g., `INTERACTS_WITH`, `INHIBITS`, `STIMULATES`), and clinical medicine (e.g., `TREATS`, `DIAGNOSES`, `PROCESS_OF`). What is essential about SemRep is that it recognizes concepts and relations from the Unified Medical Language System (UMLS).[2] The program SemRep takes as input a natural language sentence and outputs the semantic predications in the form of the subject, object, and relation that links them from the UMLS semantic network. Consider the following example:

"Dietary salt intake was directly associated with risk of gastric cancer in prospective population studies, with progressively increasing risk across consumption levels."

SemRep will identify the following triple:

- Dietary salt intake (C0425431),
- `PREDISPOSES`,
- Malignant neoplasm of stomach (C0024623).

The example shows the subject and objects arguments of the predication that are concepts from the UMLS Metathesaurus with their corresponding unique identifiers in parenthesis, and their binding relationship (in uppercase) is a relation from the UMLS Semantic Network. For a more detailed description of SemRep, see [41]. Unfortunately, not all relations that SemRep can identify fit our definition of claims. For instance, an `IS_A`, `LOCATION_OF`, or `COEXIST_WITH`, among others, do not fit our definition. To identify the relations fitting our definition, we manually analyzed each of the relations that currently recognizes SemRep and decided to consider the following relations to fit our definition of empirical claims: `CAUSES`, `TREATS`, `USES`, `INHIBITS`, `DISRUPTS`, `PREVENTS`, `STIMULATES`, `AUGMENTS`, `DIAGNOSES`, `PRODUCES`, `INTERACTS_WITH`, `METHOD_OF`, `PREDISPOSES`, `ASSOCIATED_WITH`, `AFFECTS`, `PRECEDES`, `COMPLICATES`, and `PROCESS_OF`. Hereafter let us called the semantic relations mentioned `RELATION_CLAIMS`. For our analysis, we used the table named `PREDICATION` from SemmedDB.

[1] PubMed comprises more than 28 million citations for biomedical literature from MEDLINE, life science journals, and online books.

[2] More information about UMLS in https://www.nlm.nih.gov/research/umls/.

The PREDICATION table contains several attributes including the sentence where the predication triple was identified. Relevant for our experiments are PMID and PREDICATE. We query the PREDICATION table to validate each of the PMIDs that resulted from our Boolean query. We proceeded as follows: we consider a document as correct if it contained at least one of the relations in the set of relations in RELATIONS_CLAIMS. We found that 1,010,668 out of 1,163,953 papers contained at least one sentence with a relation in RELATION_CLAIMS. In summary, 87% of the results of our query indeed fit the definition of empirical claims. We show our findings in Fig. 1.

To tackle the challenge of the first two questions, we develop a data-driven approach relying on a novel integration of state-of-the-art neural embedding representation of text and generative topic models to operationalize the concept of plausibility. Our goal is to provide a way to assess the consistency of each new document with respect to the current knowledge (i.e., the state of the art) so that we can answer questions such as: is a new document consistent with current knowledge? Do we have documents in our collection supporting or contradicting a new document? Can we represent our collection in a way such that we can derive a decision to reflect the consistency of new knowledge in light of current knowledge? To accomplish this, we first operationalize the concept of plausibility. As a proof of concept, we then implement a new architecture integrating these ideas and providing first insights by analyzing empirical claims. In summary, our contributions are:

1. Firstly, a representation of document collections that combines topic modeling with a neural embedding to exploit two relevant metadata elements: conclusions and abstracts.
2. Secondly, a query facility to find semantically similar claims that may support or contradict a new documents claim.
3. And thirdly, a mechanism to finally assess the plausibility of a new document, e.g., to verify its consistency with respect to a collection's representation.

In this paper, we extend our preliminary work in [15] to get plausibility assessment in digital libraries to the next level: after formalizing and operationalizing the notion of plausibility, we introduce a new task: automatic tagging of claims using a supervised machine learning approach, and perform experiments for evaluation.

While one clear motivation is to assist the peer-reviewing process at document level, three other ideas underline the value of operationalizing plausibility. Firstly, our proposed approach may help users to organize the range of findings in a specific scientific field. Secondly, the semantics of detecting something as not being plausible—because it contradicts current knowledge—deserve further attention as an indicator of novelty, i.e., a new finding challenging current beliefs. Thirdly, our approach could assist researchers to find related literature when building up an argumentation in research papers. At this point, all of these aspects stand unexplored and will drive future directions in our research.

Our paper is organized as follows: in Sect. 2, we provide relevant related work. We then propose a general architecture with the formalization of Plausibility in Sect. 3. In Sect. 4, we present the experimental setting to evaluate our proposed solution with a discussion of our findings. Afterward, we present concluding remarks and outline future work in Sect. 5.

2 Related work

Many attempts to model arguments for different purposes exist in the literature. Particularly relevant for our work is the body of research dealing with the semantic annotation of claims of scientific articles in the biomedical domain. For instance, Ciccarese et al. in [12] developed a model for the annotation of scientific hypotheses and claims in natural language using as a case study of Alzheimer's disease. Nanopublications [19,28,45] were promoted by the Concept Web Alliance models core scientific statements with associated context, and it is used in the work of Groth et al. in [20] for data integration across chemical and biological databases. A more detailed model of scientific papers in the biomedical domain is the work of Velterop called micropublications [45]. The model specified as an OWL 2 Vocabulary (the ontology language for the Semantic Web) is developed around the idea that scientific claims are defeasible arguments [43,46]. Thus, they support natural language statements, data, methods, materials specifications, discussion, challenge, and disagreement. In our work, we build on these ideas and represent one core component of scientific papers: claims. Moreover, we attempt to operationalize the notion of Plausibility from Cognitive Sciences. In particular, a Plausibility theory that has been empirically proven by Connell et al. [13] to be strongly correlated with human judgments.

3 Model and approach

In this section, we formalize the notion of Plausibility and the problem we aim to solve. Plausibility in this work builds on the knowledge-fit theory from Cognitive Sciences. The theory states that human judgments consist of two steps: firstly, a mental representation of prior knowledge that allows us to comprehend and make sense of the world; secondly, assess how new knowledge fits this prior knowledge. Therefore, to operationalize Plausibility we need to define formally: a)

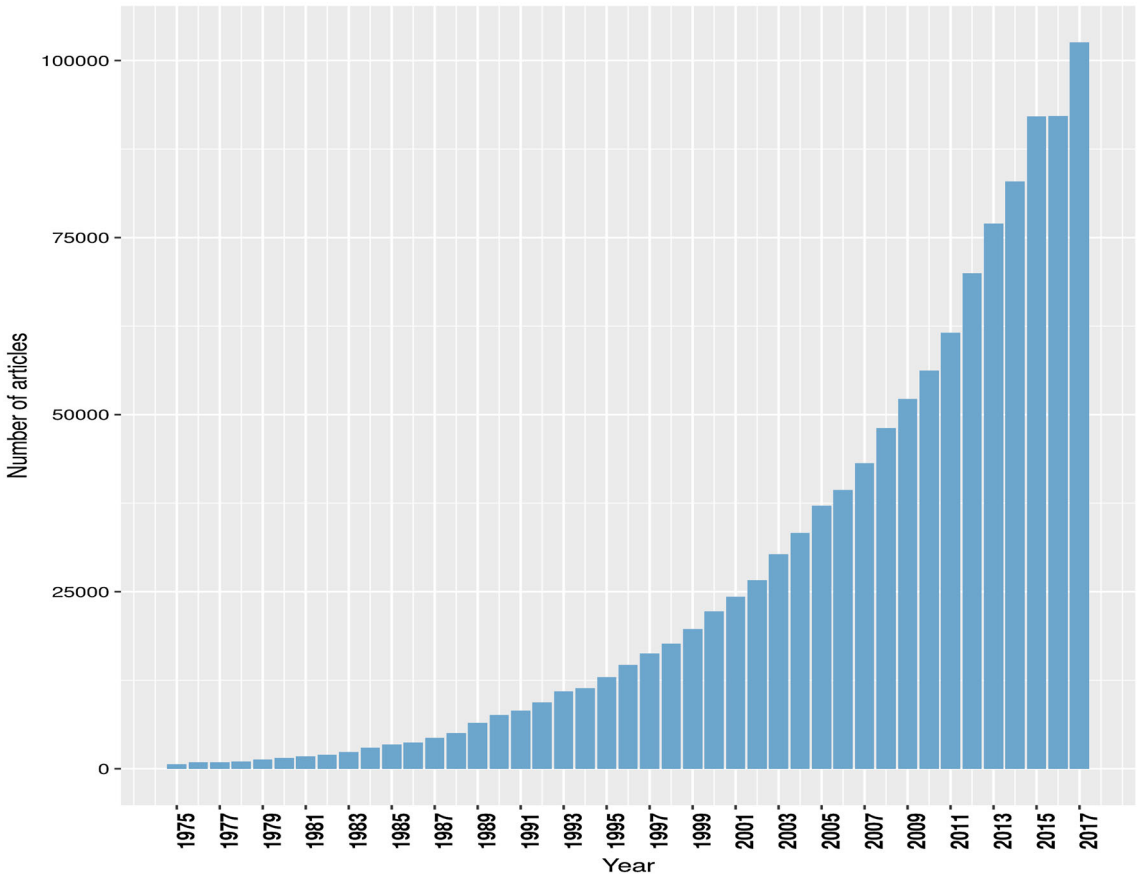

Fig. 1 Number of articles containing empirical claims in PubMed per year

how to represent our current knowledge of a Digital Collection and b) how to determine Plausibility of a new document given a).

Let us then formally define each of these two core parts of our proposed approach.

3.1 Representation of current knowledge of a digital collection

Let us define formally our document model. We consider a Digital Collection of documents $D = \{d_1, ..., d_n\}$. Each document d_i in D is a tuple $(Claim, Context)$. Where:

- *Claim*: a claim is a sentence that contains an association between two entities. In particular, how one of the entities influences, causes, manipulates, or affects the other. In this work, we use the conclusion metadata of each paper to find such sentences.
- *Context*: the sentences that correspond to the abstract of a paper where a given claim exist.

To instantiate these definitions, we further proceed as follows:

Claim embeddings Claims play a core role in our approach. Thus, we represent them using state-of-the-art language models. In particular, we use neural network language models [3,32–34,44]. Embedding language models have shown interesting semantic properties to find related concepts, related paragraphs, analogies, etc. Such word embeddings are generally computed using word-occurrence statistics (observing what words co-occur in sentences or documents), using a variety of techniques, some involving neural networks, others not. In our work, we use the successful word embeddings implementation called word2vec and developed by Tomas Mikolov during his time at Google back in 2013 [30,32]. However, some other word embeddings approaches could actually fit our goal, such as [9,38]. In this work, we rely on such representations to capture claim-specific semantics. The idea to use this type of representation is not only to find semantically similar claims but also to distinguish between claims that express supporting or contradicting positions with respect to the claim of the document we want to assess.

We decided to use the embedding space in our experiments because it benefits our approach to ease the finding of highly

semantic related claims. The intuition is that entities used in similar contexts with respect to the effect on another entity are related and might help in the absence of explicit knowledge. To put into effect this intuition in our experiments, we first train the word embedding in the entire collection of documents over the abstracts. After that, we represent each claim as a weighted point of embedded words. Indeed, it is this representation that is used to query the embedding space and find related claims.

Topic Context Model We use a generative probabilistic model to represent the *Context* of each $d_i \in D$. In particular, the latent Dirichlet allocation (LDA) created by Blei et al. in [8]. This model is an instance of a general family of mixed membership models for decomposing a collection into multiple latent components (topics). In LDA, it is assumed that words of each document arise from a mixture of topics, where each topic is a multinomial over a fixed vocabulary. The topics are shared by all documents in the collection, but the topic proportions vary stochastically across documents, as they are randomly drawn from a Dirichlet distribution, see the article by Blei in [6] for a detailed overview of Topic Models and applications. Algorithm 1 formally shows the generative process behind LDA. In Algorithm 1, we represent the document collection D using bag of words as studied by Blei et al. in text mining tasks [7]. Note that in the model the distribution over the words given the number of k topics is assumed to be a Dirichlet as well as the distribution over the topic proportions.

Data: D, K
Result: `Topic Context Model`
foreach *topic* $t \in K$ **do**
  Draw a distribution over the words β_k
end
foreach *document* $d \in D$ **do**
  Draw a vector of topic proportions θ_d;
  foreach *word* $w \in d$ **do**
    Draw a topic assignment $Z_{d,n}$
  end
end

Algorithm 1: LDA Generative Process

3.2 Plausibility of a new document

In this section, we provide details of how to determine the Plausibility of a new document given the document model previously defined. The notion of Plausibility in our work builds on the ideas introduced in the Cognitive Sciences. What we want to create is an automatic mechanism that mirrors what humans do when contrasted with new knowledge: an assessment of how consistent the new knowledge is concerning what we already know. To accomplish that, we focus on the following idea regarding consistency: consistency of new knowledge in light of current knowledge means to discover whether the new knowledge contradicts what is known or not. To realize this goal in a computer system, we need to perform two steps: firstly, given a new document, we use its claim as a query to find in our document collection documents that are semantically similar to this new document. Secondly, we use the result set of the query to identify which documents may support or/and contradict the new document. Using these two sets of documents, we then proceed to assess the consistency of a new document. Let us proceed to define these ideas and explain the details of how we operationalize these two fundamental steps: how to determine consistency (Sect. 3.2.1) and how to find similar semantic claims (Sect. 3.2.3).

Let d_{new} be a document that is currently not in our collection D and we would like to assess its Plausibility, given our current knowledge. Let $ClaimOf(d_{new})$ be the claim of document d_{new}. Let $DocSimClaim(d_{new})$ be the set of documents dealing with semantically similar claims to $ClaimOf(d_{new})$. Moreover, let $DocsContradict$ and $DocsSupport$ be the documents that contradict and support, respectively, the new document d_{new}.

3.2.1 Consistency

The notion of how to assess `consistency` is a key aspect to operationalize the idea behind Plausibility. Let us formalize `consistency` to clarify what exactly means that a new document d_{new} is consistent concerning a Digital Collection D. In a nutshell, `consistency` means "agreement" of a new document d_{new} with the Digital Collection D.

To implement this idea in an information system, we define a function `checkConsistency` that exploits the tuple representation of documents to assess whether a new document agrees with a collection. The function `checkConsistency` relies on the Possible World Semantics introduced by Dalvi et al. [14] to accomplish its task; to operationalize the Possible World Semantics, we define a Possible World in a Digital Collection as a set of documents which are as similar as possible within the Possible World and dissimilar as possible from documents in any other Possible World. In other words, we rely on the cluster hypothesis, see Manning et al. [31], from information retrieval.

Cluster hypothesis Documents in the same cluster behave similarly with respect to relevance to information needs.

The cluster hypothesis states the fundamental assumption we make when using Possible World semantics for cases where we have documents in our collection that both support and contradict a new document: to assess how consistent a new document is with respect to current knowledge, we should look at other documents that share the same Possible World as given by the context of the documents.

Thus, let PW be a Possible World that represents the most likely topic of a document from the Topic Model learned using the bag-of-words representation of the *Context* of the documents in D [31].

This notion of Possible Worlds is crucial to deal with the most difficult situation that we might find with respect to d_{new}: finding documents that contradict $DocsContradict$ and support $DocsSupport$ d_{new} in our collection D. Thus, this latter situation is now handled as follows: if there exist documents that both contradict and support d_{new}, we use the context of d_{new} to map it to the PW where it should belong to. If in this world all documents agree, and the new document d_{new} also states the same position of this PW, then the new document d_{new} is `plausible`. More formally, we can now define our problem:

Definition 1 *Document Plausibility Problem*: given a document collection $D = \{d_1, ..., d_n\}$, and a new claim in a new document d_{new}, we aim at finding if the claim in d_{new} is consistent with respect to D.

To solve the problem, all that remains to do is to formally define an algorithm to assess whether a new document d_{new} is `plausible` or not, let us do that now.

3.2.2 Plausibility algorithm

To determine the Plausibility of a new document, we apply Algorithm 2 Plausibility Assessment. Two special cases to notice: If $DocsContradict$ and $DocsSupport$ are both empty, then $ClaimOf(d_{new})$ calls for a special assignment of resources to manually assess its value. The same applies for a situation that we have called `controversial`. A controversial situation exists when our algorithm finds `inconsistent` the Possible World where the new document d_{new} will belong. The other cases covered by our algorithm are easier to follow: if our approach can only find documents that support the new document, then the new document is `plausible`. However, if our approach can only find documents that contradict the new document, then the new document is `not plausible`.

Let us consider the following illustrative examples to give an overview of what we seek to achieve.

Example 1 Suppose we have in our document collection D a document d_1 whose claim states "beta-carotene consumption is associated with reduced risk of cancer". Consider a new document d_{new} that states "beta-carotene consumption significantly increases the risk of cancer". Thus, the new document d_{new} contradicts what we currently know and our approach will signal this new document as *not plausible*. To handle these types of cases, careful human assessment will be required.

Data: $DocsContradict$, $DocsSupport$, d_{new}
Result: `plausible`, `not plausible` or `controversial`
if *DocsSupport is not empty and DocsContradict is empty* **then**
  return `plausible`;
end
if *DocsSupport is empty and DocsContradict not empty* **then**
  return `not plausible`;
end
if *DocsSupport not empty and DocsContradict not empty* **then**
  isConsistent =
  checkConsistency(d_{new},DocsSupport,DocsContradict);
  if *isConsistent and* d_{new} *agrees* **then**
    return `plausible`;
  end
  if *isConsistent and* d_{new} *disagrees* **then**
    return `not plausible`;
  end
  return `controversial`;
end

Algorithm 2: Plausibility Assessment

Example 2 Suppose we have in our document collection the same document d_1 of the previous example. Moreover, suppose a new document d_{new} states "beta-carotene consumption has shown to play a major role in the prevention of cancer". In this case, the new document d_{new} agrees with what we know: there exists a beneficial association between beta-carotene and cancer; thus, our approach will signal the new document as `plausible`.

In other words, our proposed decision process is as follows: if a new document agrees with our current view, it is considered `plausible`, otherwise it is `not plausible`. Now consider the difficult situation: if in our Document Collection exist documents with claims that at the same time contradict and support the new claim, we decided to label it as `controversial`.

However, we can push this idea further: to identify whether the documents in our Digital Collection exhibit some characteristics that make them belong to different possible worlds. Thus, if we can find a Possible World that is consistent and the new document will belong to this possible world, then we can proceed to assess the Plausibility of the new document as before. Otherwise, if the Possible world is inconsistent, then we again arrive at a `controversial` situation.

In Fig. 2, we show a diagram that summarizes the interaction between the different components of our proposed methodology.

3.2.3 Finding semantic similar claims

Finding similar claims $DocSimClaim(d_{new})$ in our work for a given d_{new} is a crucial step to make our approach work. This first step involves not only being able to find similar claims $DocSimClaim(d_{new})$ but also to distinguish between claims that support $DocsSupport$ or contradict $DocsContradict$

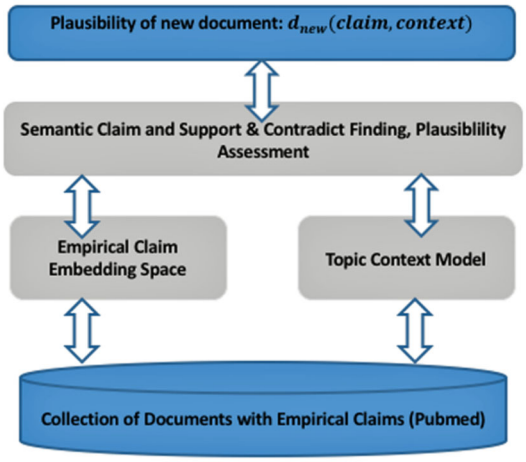

Fig. 2 Architecture of plausibility

d_{new}. Thus, for this task, we proceed as follows: we rely on the distance metric discovered and efficiently implemented by the work of Kusner et al. in [29,36]. This approach called the Word Mover's Distance (WMD) is a method that allows us to assess the distance between two documents using word embeddings in a meaningful way. What makes this distance so useful is that it can find similar documents in the embedding space even when they have no words in common! It exploits vector representation of word embeddings, and it has been shown to outperform several of the state-of-the-art methods in k-nearest neighbors classification. Technically speaking, WMD is inspired by the Earth Mover's Distance and employs a solver of the transportation problem. Because this method has been shown to be very useful to find similar semantic documents in the embeddings space against some other alternatives as shown by Kusner et al. in [29], we rely on it in our first step.

WMD details. In the following paragraphs, we provide details of how the Word Mover's Distance (WMD) works. To understand the approach, let us first start with Linear Programming, then we discuss the Transportation Problem—the primary source of motivation of the Word Mover's Distance.

Linear Programming (LP), see Bertsimas et al. [5] for an in-depth introduction to the field, deals with the problem of maximizing or minimizing a linear function subject to linear constraints. The constraints are equalities or inequalities. Thus, as an illustrative example consider the following problem: find the numbers x_1 and x_2 that maximize the sum x_1+x_2 subject to the constraints $x_1 >= 0$, $x_2 >= 0$, and

$$\begin{aligned} x_1 + 2x_2 &<= 4 \\ 4x_1 + 2x_2 &<= 12 \\ -x_1 + x_2 &<= 1 \end{aligned}$$

One of the original applications of Linear Programming was the so-called Transportation Problem. The idea is the following: a company produces certain goods at m different supply centers, $i = 1, ..., m$. The supply produced at supply center i is S_i. The demand for the good is spread out at n different demand centers $j = 1, ..., n$. The demand at the jth demand center is D_j. The problem of the company is to get goods from supply centers to demand centers at minimum cost. Assume that the cost of shipping is one unit from supply center i to demand center j is c_{ij} and that shipping cost is linear. The problem is to identify the minimum cost shipping schedule. The constraints are that you must (at least) meet demand at each demand center and cannot exceed supply at each supply center. Thus, the objective is to find a transportation plan denoted by T_{ij} to solve:

$$\min \sum_{i=1}^{m} \sum_{j=1}^{n} T_{ij} c_{ij}$$

subject to

$$\sum_{j=1}^{n} S_i \forall i = 1, ..., m$$

and

$$\sum_{i=1}^{m} D_j \forall j = 1, ..., n$$

If one substitutes demands and supplies with words, then we have the Word Mover's Distance (WMD)! Notice that the above optimization problem has specialized solvers as noted by Kusner et al. Thus, all that remains to know is how to compute the cost c_{ij} and how to represent the documents. The cost c_{ij} is defined as the Euclidean distance between word i and word j; thus, we have:

$$c_{ij} = ||x_i - x_j||_2$$

where x_i and x_j are the d-dimensional word embedding representation of words i and j, respectively, from word2vec. As described by Kusner et al. in [29], the idea of "travel cost" between two words gives the building block to finally create the distance between two documents. Let d and d' be the bag-of-words representation of two documents after removing stop words. Then, one can define the distance between the two documents as the minimum weighted cumulative cost required to move all words from d to d'.

In Fig. 3, we show an illustrative example of the semantics captured in the claim embedding space as given by the distance computation named Word Mover's Distance (WMD). In the figure, we show how two sentences that contain entities such as "tomato sauce" and "lycopene" end up close to each

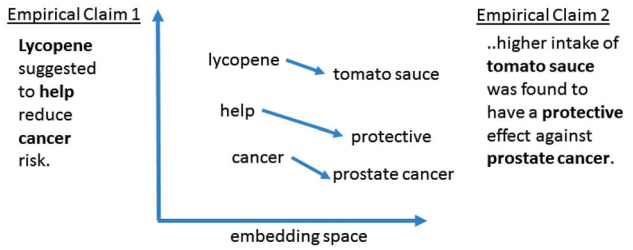

Fig. 3 Illustrative example of the power of the embedding space using WMD distance

other in the embedding space because of the semantics captured by the WMD. WMD tries to match the relevant words in the documents to compare and measures how similar or dissimilar the two sentences are.

Thus, instead of using a fixed list of synonyms, we rely on the WMD method to discover similar entities used in similar contexts. After that, we need to distinguish between supporting $DocsSupport$ or contradicting $DocsContradict$ documents. Because this distinction is a core part of our proposed approach, we focus on claims expressing `increase` vs. `decrease` associations that express contradictory positions. Thus, in our experiments for the claims that we investigate we distinguish these two positions. A simple textual pattern mechanism with synonyms to the words `increase` and `decrease` was used to distinguish between the two. Synonyms in this work are words related to `increase` and `decrease` as captured in the embedding space. We are aware of some other cases such as: "After adjustment for demographic and dietary characteristics, there was *no association* between pancreatic cancer risk and the intake of coffee, beer, red wine, hard liquor or all alcohol combined" or "Recent observations of association of risk with coffee consumption and with use of decaffeinated coffee *require further evaluation*", that need further attention to detect such semantics in the associations. However, in this work we focus only on `increase` and `decrease` and let for future work the detection of such other cases.

4 Experiments and findings

4.1 Exploratory experiments

To demonstrate and evaluate our proposed Plausibility indicator, we performed experiments with two primary goals. Firstly, we wanted to gather valuable insight into the notion of finding similar semantic claims in our corpus that either support or contradict a new document's claim. Because we do not have ground truth, we manually observed the claims and set a threshold that to the best of our understanding can lead to highly related claims that may or may not support each other concerning a disease. After this experimentation, we set a threshold of 0.50 for the experiments that we report here. Secondly, as a proof of concept, we needed to compare our approach to experts' work. In particular, we chose the results reported by Schoenfeld et al. in [42] that we mentioned in Introduction. Thus, we first retrieved all the documents related to two of the ingredients investigated by Schoenfeld et al. "salt" and "lycopene". To retrieve the documents, we used the query pattern mentioned in our Introduction. We chose these two cases to acknowledge the scope of our tool since they represent different situations: salt was found to be one of the few exceptions of the analysis regarding its risk effect that was not the subject of controversy due to contradicting findings. However, lycopene represents a situation that cannot be plausible because there are documents that on the one hand conclude that an increased effect exists regarding cancer and on the other hand other documents conclude the opposite. Thus, the goal of studying these two cases was to see whether we could find a suitable explanation. For every document, we extracted the empirical claims contained in the conclusions section when available. Unfortunately, not all the documents contained this valuable metadata. Thus, the collection of documents used for these experiments consist of 87k documents. We used this collection to train our embedding representation using default parameters as given by the open-source project Gensim [40]. For every experiment, we first selected one document at random and considered it as the new document and proceeded to assess its Plausibility.

The first case that we report here is the association of "salt" and "cancer". The title of the document to evaluate was "Salt intake and gastric cancer risk according to helicobacter pylori infection, smoking, tumour site and histological type" by Peleteiro et al. [37]. The claim of the paper that we used to query our semantic embedding space is a very simple one: "our results support the view that salt intake is an important dietary risk factor for gastric cancer, and confirms the evidence of no differences in risk according to h. pylori infection and virulence, smoking, tumour site and histological type." After querying our semantic embedding space, we retrieved some related claims in other papers. Some examples are:

- "Dietary salt intake was directly associated with risk of gastric cancer in prospective population studies, with progressively increasing risk across consumption levels."
- "Improved dietary habits, reducing salt consumption and eradication of h. pylori infection may provide protection against gastric cancer in Turkey."
- "These data suggest that high intake of salt and smoked and pickled food may be associated with a high risk of gastric cancer, and this association could be due to intragastric formation of nitrosamines"

In this particular case, the new claim finds support in our current knowledge and our approach states a `plausible` situation. This is an example of how our approach could help a reviewer to assess whether the new document is consistent with the current body of knowledge. Basically, it could allow the reviewer to find similar studies dealing with the specified entities probably in similar ways.

Next, we report on a second experiment. In this second experiment, we take "lycopene" as one of the ingredients where there was evidence of being in a situation that we call `controversial`. Remember that `controversial` means that "lycopene" was found in an increased and decreased risk association in different research papers. In this particular case, we found 197 documents related to the association within our collection. We selected a document with the following claim: "this study does not support a role for lycopene in prostate cancer prevention" reported by Kristal et al. [27].

We found in our collection claims that both support and contradict the new document's claims. That leads to a `controversial` situation as defined in our methodology section. However, we can take our second step to make a final decision. In this case, the new document fits better in a possible world that is not `consistent`. Thus, we conclude that this is a controversial situation in need of human experts to look into the new document carefully. One should notice the level of complexity of this case. For instance, in the community-curated archive Wikipedia entry for lycopene, the FDA (US Food and Drug Administration), in rejecting manufacturers' requests in 2005 to allow "qualified labeling" for lycopene and the reduction in various cancer risks, stated:

> "...no studies provided information about whether lycopene intake may reduce the risk of any of the specific forms of cancer. Based on the above, FDA concludes that there is no credible evidence supporting a relationship between lycopene consumption, either as a food ingredient, a component of food, or as a dietary supplement, and any of these cancers."

Furthermore, two more experts of medical panels cited in the entry of the Wikipedia page also confirmed this situation.

To get a better assessment of the potential of our approach, we performed simulated experiments with a selection of the 80 most recent meta-analyses found in our collection with respect to three other diseases: hypertension, diabetes, and asthma in addition to cancer. A meta-analysis is a systematic review that uses statistics analysis to combine several research papers on a particular topic. One characteristic of meta-analyses is that it may never be possible to include all the papers that deal with a particular phenomenon. Usually, researchers query a digital library using keywords to get a candidate set of papers and after that, they manually decide which candidates can be included in the analysis. Depending on the methodology chosen by the researchers, the final number of articles varies.

In this set of experiments, we proceeded as follows: we took out of our collection the meta-analysis, and then, we queried our representation using the claim of the meta-analysis. If we could agree with the claim of the meta-analysis in at least one possible world, then we consider that as a positive outcome. After our experimentations, our best result was a kappa of 0.7746 with a 95% confidence interval (0.7875, 0.9549). Notice that because the criteria that the experts use to include and/or exclude some papers in a meta-analysis are beyond our current text mining processing, we included all papers as given by our query pattern. However, one caveat of these types of experiments is the training time of the embedding and the LDA hyperparameters. In this particular setting, we trained the word embeddings with 100 dimensions and LDA with 8000 iterations with a fixed 300 topics in a collection of 315k documents.

To provide insights of our results, let us look at one of the cases where our approach failed. Consider the findings of the work of Zhao et al. reported in [49], where the study of alcohol regarding prostate cancer was analyzed. As stated in the paper, a total of 340 studies were found in the exploratory search, but only 27 satisfied the inclusion criteria of the researchers (manual assessment). For this case, we found a `controversial` situation. In other words, our proposed approach did not agree with the meta-analysis in any possible world. More specifically, all the possible worlds were inconsistent and our tool stated a `controversial` situation. Moreover, Zhao et al. reported "Our study finds, for the first time, a significant dose response relationship between level of alcohol intake and risk of prostate cancer starting with low volume consumption". Of course, this is an expert assessment and our tool is not aiming at replacing a decision but instead helping to detect situations that may require a better administration of reviewers, especially in cases of controversy where clearly major care should be taken.

4.2 Quantitative experiments

In this section, we test our approach on the entire collection of documents from our query pattern concerning the following diseases: cancer, hypertension, and asthma. For each disease, we provide in Table 2 the statistics (January 2018) with the number of documents in the collection with and without a conclusions metadata.

We observed in our previous study the absence of the conclusions metadata in many of the documents. This metadata is crucial for a more realistic measure of the accuracy of what we have proposed. We have defined claims as sentences that link entity pairs in a relationship with the constraint that they should be part of the major contribution of the research paper.

Table 1 Statistics of the number of sentences in our document collection

Type	Sentences
No claims	4,433,879
Claims	426,794

Table 2 Statistics of the number of documents per disease

Disease	Number of docs	Docs. with no metadata
Cancer	236,909	130,205
Hypertension	81,571	41,059
Asthma	16,432	7,779

Previously, we just ignored documents that did not contain the conclusions metadata and proceeded with our approach. In these experiments, we aim at using all the papers available and test our approach on a significant scale. Thus, we introduce to our pipeline a new task to recover the missing metadata. After that, we proceed to test our approach.

4.2.1 Automatic tagging of claims

Let us then first describe the new task that we introduced in this paper: recovering for each document retrieved from our query pattern, the metadata needed for our approach: sentences that correspond to the contribution of a paper. Since we want to obtain a representative function that at the sentence level outputs whether a sentence is part of the contribution of a paper, we exploit the large number of abstracts that include the conclusions metadata in PubMed. We assume that papers with the conclusions metadata fit the idea of finding the contribution of a paper.

To put things into perspective, we show in Table 1 the statistics of the number of sentences found in the document collection of our query pattern. In Table 1, `Claims` are the number of sentences found in the conclusion metadata, and `No claims` are the number of sentences found in the abstract metadata. We can observe that we have enough data to accomplish the task according to Goodfellow et al. [16]. Formally, the task consists in assigning a specific label to each sentence in an abstract of a document: either the sentence represents a contribution or not. Here contribution means that the sentence should be in the conclusions metadata.

Thus, let $Sentences(d_i) = \{s_1, ..., s_n\}$, be the sentences of a document d_i , and let L denote the labels `claim`, `noclaim`. We attempt to learn from the observed data, a function that assigns one of the two labels to each sentence. Thus, we perform a binary classification task. In particular, a supervised machine learning approach. In Table 2, we show the impact of this task for our approach: half of the documents do not have the conclusions metadata. Thus, without the proposed solution we would have missed a lot of information.

Approaches used for text classification Text classification has recently seen deep learning approaches to surpass more classical machine learning approaches such as SVMs or logistic regression. In general, researchers have used two types of deep learning algorithms for text classification: recurrent neural networks and convolutional neural networks. The curious reader in a proper introduction to Neural Networks' terminology and Deep Learning in general, please see the authoritative work of Goodfellow et al. [16]. In the following paragraphs, we briefly describe the models we used in our experiments.

Recurrent neural networks (RNN) The intuition behind models based on recurrent neural networks is to process input data such as sentences, in a similar way that humans do: we handle each sentence word by word, but we keep memories of what came before. More formally, a recurrent neural network process sequences by iterating through the sequence elements and maintaining a state containing information relevant to what it has seen so far and using a loop in the architecture to avoid forgetting what the network has learned. Even though in theory RNNs should be able to retain information about inputs seen many timesteps before, such long-term dependencies are just impossible to learn. Why? Well, researchers called it the vanishing gradient problem.

This problem has the effect of making the network become untrainable. The details of the theory behind it were studied by Hochreiter et al. [4]. Moreover, the algorithm developed by Hohreiter and Schmidhuber [22] represents a milestone in research concerning the vanishing gradient problem. One of the most widely successfully used recurrent neural networks is the LSTM. Researchers have used it to solve several time-series or sequence data problems, such as sentence embeddings for information retrieval investigated by Palangi et al. [35], speech recognition from audio data analyzed by Graves et al. [17] or the translation of sentences into different languages proposed by Bahdanau et al. [2]. As with other deep-learning-based approaches, tuning the parameters of such networks is challenging. Practitioners have to tailor the model to the specific problem. What this means in practice is a lot of time to try-measure-repeat. Fortunately as researchers understand better what might be good starting points, one can rely on guidelines to explore if any of these models can solve a particular problem. For instance, Greff et al. [18] performed an extensive empirical study of different variations of LSTMs performance and tuning of its parameters. Due to the successful application of these types of networks, we implement three variants of the LSTMs: a BiLSTM, a two stack of LSTMs and a combination with a convolution neural network (CNN). The architectures used in this work were all implemented in the open-source library Keras developed by Chollet et al. [11]

using the work of Abadi et al. Tensorflow [1] as the backend engine.

Convolutional neural networks (CNN) Researchers [25,47,48] tailored and successfully applied Convolutional Neural Networks on sentence classification. What makes CNN so useful is the ability of the network to learn linguistic patterns on distributed representations of words but without asking for it. Goodfellow, I. et al. [16] have emphasized that three 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 learning automatically—no feature engineering required—linguistic n-grams patterns; parameter sharing influences computation storage requirements; equivariant representation provides for robustness in the patterns learned regarding the position in the sentence. In contrast to LSTMs, CNNs are faster to train than LSTMs but can perform less accurately than LSTMs. To combine the best of the two models, we combine a CNN with an LSTM layer to get a CNN LSTM-based topology.

Hyperparameter details All of these deep learning variations that we try have the same downside: trying them to account for all the hyperparameters is just not reasonable, it involves a highly intensive computational task. As pointed out by Chollet [10], deep learning methods lack a theory to tell in advance what one should do to solve a problem optimally. Thus, to select the specific hyperparameters used in this work, we mainly iterate: we start with a small network, gradually increase its hyperparameters capacity and we keep on doing this until the validation score no longer improved. To train all these models, we used ADAM optimizer [26] using binary cross-entropy as a loss function and regularize—to avoid overfitting—by early stopping and a rather high drop-out rate (0.50) following Hinton et al. findings [21]. Note that the final network hyperparameters were the following: layer size to 300, vocabulary size to 30,000, training epochs set to 20, and a batch size of 128. Furthermore, 80% of the data was used for model training and 20% for testing. Note that we have more examples of sentences that do not correspond to `Claims`, thus an instance of an imbalance problem. To deal with this problem, we used Random Under-sampling: majority class instances are randomly excluded until positive and negative class instances become equal.

We also apply to our task a well-known baseline model: the Support Vector Machine (SVM) [23].

Results. In Table 3, we show the results of our experiments. In the table, `SVM` stands for a Support Vector Machine model using bag of words. `BiLSTM` stands for a bidirectional recurrent model. `CNN with LSTM` stands for a model that combines one layer of a convolutional neural network and one layer of a recurrent neural network. `Stack of two LSTM` stands for a model with two recurrent layers. All deep-learning-based approaches outperform the `SVM` approach with a subtle difference. The `stack of two LSTM` turned out to fit this particular binary classification problem better. The model based on the combination of the CNN and the LSTM turned out to be the second best approach by a minor margin. In summary, the best model obtained an overall high accuracy of 84% at the sentence level. Overall, we consider the performance we obtained to be in the range of other reported results in similar tasks.

Table 3 Results of the task to tag claims automatically

Method	Accuracy
SVM	0.78
BiLSTM	0.81
CNN with LSTM	0.82
Stack of two LSTM	0.84

Table 4 Summary of results per disease

Disease	Accuracy	Claims evaluated
Cancer	0.80	1997
Hypertension	0.82	277
Asthma	0.89	298

4.2.2 Analysis of results

For the experiments reported in this section, we first applied the best model of the automatic tagging classification task described in the previous section to recover the claims of documents that had no conclusions metadata. We focus on three diseases: cancer, hypertension, and asthma. To measure how our approach performs we will proceed as follows: for each meta-analysis, we will apply our Plausibility algorithm. If our algorithm can determine that this new document is plausible, or if it agrees with the Possible World where the new document fits, then we will consider the meta-analysis as correctly classified by our approach. Thus, we will report accuracy as our metric. We will report the results per disease. We show in Table 4 the performance of our proposed approach. In Table 4, column `Claims evaluated` refers to the number of claims found in documents with meta-analysis. Overall, we achieve an average accuracy of more than 80%. Thus, we consider our results very promising. The most challenging set of claims corresponds to cancer. Our findings confirm expert assessment reported by Schoenfeld et al. [42] where researchers found how difficult it is to interpret findings regarding cancer. Nevertheless, our results confirm that we can move forward and test our approach in other domains. We anticipate the challenge of evaluating our proposed approach in domains where the notion of meta-

analysis probably does not exist. However, the potential to apply our assessment of how a new document fits current knowledge motivates us to continue our work and alleviate the workload of reviewers in a real scenario.

4.3 Discussion

Our results look promising, and there are some issues that we noticed during our experiments. Firstly, the assessment of the degree of association between the claims is something that only domain experts can properly adjust. For instance, the idea that "tomato sauce" and "lycopene" can be considered similar enough to retrieve claims that associated both of them with cancer depends on what the experts would consider as related. Moreover, the idea of considering or not considering related types of a disease, such as prostate cancer, lung cancer, or gastric cancer, in the retrieval of related claims is again questionable. In our experiments, we did notice a difference when we filtered the results to restrict the retrieval to the specified entities. Nevertheless, we envision an application where the reviewer can actually experiment with this feature of our approach. Secondly, the final decision of `controversial` with the idea of the "possible world" explanation did help to some extent but stayed below expectations in the experiments. One possible explanation is the criteria of inclusion or exclusion of articles in a meta-analysis and the methodology used to assess its conclusion.

These two aspects are of course beyond our approach capacities and not in the scope of what we want to achieve. And third, our approach could accurately find controversial situations as confirmed by the meta-analysis experiments. However, this was only possible when we did not restrict the entities to exact matches but instead expanded them.

5 Conclusions

We introduced a novel approach to assess the Plausibility of a new document to support peer review not at the process level, but with a clear focus at the document level. Our results look promising toward the goal of different management of resources in peer review. In particular, how to adjust the number of reviewers for a new paper given its Plausibility. Of course, our experiments also reveal future work that is needed to crystallize our vision. For instance, assuming that "tomato sauce" and "lycopene" can be considered similar enough to retrieve papers that associate both of them with "cancer" depends on the goal of the analysis. Moreover, this type of assessment is something that domain experts can adequately adjust. Thus, this should be a feature that users should tune to personalize the degree of associations between the claims.

The new set of experiments that we introduced confirmed the potential of our approach with the current restriction of the types of claims considered in our work. Thus, the model of claims that we currently have must be extended to cope with other domains. To do that, we will need to account for more advanced representation of arguments in scientific papers. We are aware that the incipient field of Argumentation Mining in the last few years has shown tremendous potential to envision more powerful applications. We will also explore that line of research in future work.

We also introduced a new task to deal with papers that according to our definition of a claim did not have the metadata needed. We showed through our experiments that a deep-learning-based approach solved the problem with an overall accuracy of 80%. The outcome of this task allowed us to recognize claims automatically from more papers than our previous experiments.

Finally, three other ideas can increase the value of our current work on Plausibility. Firstly, Plausibility could help to organize the range of findings in a specific scientific field. Secondly, our approach could help to detect novelty, for example, a new claim that challenges our current beliefs. Thirdly, our approach could assist a researcher to find related literature when addressing an argument in a research paper. All of these aspects stand unexplored and drive future directions in our research.

References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Kaiser, L., Kudlur, M., Levenberg, J., Man, D., Monga, R., Moore, S., Murray, D., Shlens, J., Steiner, B., Sutskever, I., Tucker, P., Vanhoucke, V., Vasudevan, V., Vinyals, O., Warden, P., Wicke, M., Yu, Y., Zheng, X.: TensorFlow: large-scale machine learning on heterogeneous distributed systems. arXiv:1603.04467v2 p. 19 (2015). URLhttp://download.tensorflow.org/paper/whitepaper2015.pdf
2. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: International Conference on Learning Representations, pp. 1–15 (2015). https://doi.org/10.1146/annurev.neuro.26.041002.131047
3. Bengio, Y., Ducharme, R., Vincent, P., Janvin, C.: A neural probabilistic language model. J. Mach. Learn. Res. **3**, 1137–1155 (2003). https://doi.org/10.1162/153244303322533223
4. Bengio, Y., Simard, P., Frasconi, P.: Learning long-term dependencies with gradient descent is difficult. IEEE Trans. Neural Netw. **5**(2), 157–166 (1994). https://doi.org/10.1109/72.279181
5. Bertsimas, D., Tsitsiklis, J.N.: Introduction to Linear Optimization. Athena Scientific, Belmont (1997)
6. Blei, D.M.: Probabilistic topic models. Commun. ACM **55**(4), 77 (2012)
7. Blei, D.M., Lafferty, J.D.: Topic models. In: Srivastava AN, Sahami M (eds) Text Mining: Classification, Clustering, and Applications, chap. 4. Data Mining and Knowledge Discovery Series, Chapman & Hall/CRC, pp. 71–89 (2009). https://doi.org/10.1145/1143844.1143859

8. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003). https://doi.org/10.1162/jmlr.2003.3.4-5.993
9. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information **5**, 135–146 (2016). DOI 1511.09249v1. arXiv:1607.04606
10. Chollet, F.: Deep Learning with Python, 1st edn. Manning Publications, Shelter Island (2017)
11. Chollet, F., others: Keras. (2015) https://github.com/keras-team/keras
12. Ciccarese, P., Wu, E., Wong, G., Ocana, M., Kinoshita, J., Ruttenberg, A., Clark, T.: The SWAN biomedical discourse ontology. J. Biomed. Inform. **41**(5), 739–751 (2008). https://doi.org/10.1016/j.jbi.2008.04.010
13. Connell, L., Keane, M.T.: A model of plausibility. Cognit. Sci. **30**(1), 95–120 (2006). https://doi.org/10.1207/s15516709cog0000_53
14. Dalvi, N., Ré, C., Suciu, D.: Probabilistic databases: diamonds in the dirt. Commun. ACM **52**(7), 86–94 (2009). https://doi.org/10.1145/1538788.1538810
15. González Pinto J.M.; Balke, W.T.: Can plausibility help to support high quality content in digital libraries? In: TPDL 2017 21st International Conference on Theory and Practice of Digital Libraries. Thessaloniki, Greece (2017)
16. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning, vol. 521(7553). MIT Press, Cambridge (2016). https://doi.org/10.1038/nmeth.3707
17. Graves, a., Mohamed, A., Hinton, G.: Speech recognition with deep recurrent neural networks. In: 38th International Conference on Acoustics, Speech, and Signal Processing, pp. 6645 – 6649 (2013). https://doi.org/10.1109/ICASSP.2013.6638947
18. Greff, K., Srivastava, R.K., Koutnik, J., Steunebrink, B.R., Schmidhuber, J.: LSTM: a search space odyssey (2016). https://doi.org/10.1109/TNNLS.2016.2582924
19. Groth, P., Gibson, A., Velterop, J.: The anatomy of a nanopublication. Inf. Serv. Use **30**(1–2), 51–56 (2010). https://doi.org/10.3233/ISU-2010-0613
20. Groth, P., Loizou, A., Gray, A.J.G., Goble, C., Harland, L., Pettifer, S.: API-centric linked data integration: the open PHACTS discovery platform case study. J. Web Semant. **29**, 12–18 (2014). https://doi.org/10.1016/j.websem.2014.03.003
21. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R.: Improving neural networks by preventing co-adaptation of feature detectors (2012). arXiv:1207.0580
22. Hochreiter, S., Urgen Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997). https://doi.org/10.1162/neco.1997.9.8.1735
23. Joachims, T.: Text categorization with support vector machines: Learning with many relevant features. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **1398**, 137–142 (1998). https://doi.org/10.1007/s13928716
24. Kilicoglu, H., Shin, D., Fiszman, M., Rosemblat, G., Rindflesch, T.C.: SemMedDB: a PubMed-scale repository of biomedical semantic predications. Bioinformatics **28**(23), 3158–3160 (2012). https://doi.org/10.1093/bioinformatics/bts591
25. Kim, Y.: Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP, pp. 1746–1751 (2014). https://doi.org/10.3115/v1/D14-1181. arXiv:1408.5882
26. Kingma, D.P., Ba, J.L.: Adam: a method for stochastic optimization. Int. Conf. Learn. Represent. **2015**, 1–15 (2015)
27. Kristal, A.R., Till, C., Platz, E.A., Song, X., King, I.B., Neuhouser, M.L., Ambrosone, C.B., Thompson, I.M.: Serum lycopene concentration and prostate cancer risk: results from the prostate cancer prevention trial. Cancer Epidemiol. Biomark. Prev. **20**(4), 638–646 (2011). https://doi.org/10.1158/1055-9965.EPI-10-1221
28. Kuhn, T., Barbano, P.E., Nagy, M.L., Krauthammer, M.: Broadening the scope of nanopublications. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 7882 LNCS, pp. 487–501 (2013). https://doi.org/10.1007/978-3-642-38288-8-33
29. Kusner, M.J., Sun, Y., Kolkin, N.I., Weinberger, K.Q.: From word embeddings to document distances. In: Proceedings of The 32nd international conference on machine learning vol. 37, pp. 957–966 (2015)
30. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. International Conference on Machine Learning - ICML 2014, vol. 32, pp. 1188–1196 (2014). https://doi.org/10.1145/2740908.2742760
31. Manning, C.D., Raghavan, P.: An introduction to information retrieval (2009). https://doi.org/10.1109/LPT.2009.2020494. URLhttp://dspace.cusat.ac.in/dspace/handle/123456789/2538
32. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. Nips pp. 1–9 (2013). https://doi.org/10.1162/jmlr.2003.3.4-5.951
33. Mikolov, T., Corrado, G., Chen, K., Dean, J.: Efficient estimation of word representations in vector space. In: Proceedings of the International Conference on Learning Representations (ICLR 2013) pp. 1–12 (2013). https://doi.org/10.1162/153244303322533223. arXiv:1301.3781v3.pdf
34. Mikolov, T., Yih, W.t., Zweig, G.: Linguistic regularities in continuous space word representations. In: Proceedings of NAACL-HLT, June, pp. 746–751 (2013)
35. Palangi, H., Deng, L., Shen, Y., Gao, J., He, X., Chen, J., Song, X., Ward, R.: Deep Sentence embedding using long short-term memory networks: analysis and application to information retrieval. IEEE/ACM Trans. Audio Speech and Language Process. **24**(4), 694–707 (2016). https://doi.org/10.1109/TASLP.2016.2520371
36. Pele, O., Werman, M.: Fast and robust earth mover's distances. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 460–467 (2009). https://doi.org/10.1109/ICCV.2009.5459199
37. Peleteiro, B., Lopes, C., Figueiredo, C., Lunet, N.: Salt intake and gastric cancer risk according to Helicobacter pylori infection, smoking, tumour site and histological type. British Journal of Cancer **104**(1), 198–207 (2011). https://doi.org/10.1038/sj.bjc.6605993. URLhttp://www.nature.com/doifinder/10.1038/sj.bjc.6605993
38. Pennington, J., Socher, R., Manning, C.: Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014). https://doi.org/10.3115/v1/D14-1162. URLhttp://aclweb.org/anthology/D14-1162
39. Price, B.Y.S., Flach, P.A.: Computational support for academic peer review: a perspective from artificial intelligence. Commun. ACM **60**(3), 70–79 (2017)
40. Rehurek, R., Sojka, P.: Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks pp. 45–50 (2010). https://doi.org/10.13140/2.1.2393.1847
41. Rindflesch, T.C., Fiszman, M.: The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. J. Biomed. Inform. **36**(6), 462–477 (2003). https://doi.org/10.1016/j.jbi.2003.11.003
42. Schoenfeld, J.D., Ioannidis, J.P.A.: Is everything we eat associated with cancer? A systematic cookbook review. Am. J. Clin. Nutr. **97**(1), 127–134 (2013). https://doi.org/10.3945/ajcn.112.047142
43. Toulmin, S.: The uses of argument. Ethics **70**(1), vi, 264 (1958). https://doi.org/10.2307/2183556

44. Turian, J., Ratinov, L., Bengio, Y.: Word representations: a simple and general method for semi-supervised learning. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp 384–394 (2010)
45. Velterop, J.: Nanopublications: the future of coping with information overload. LOGOS: J. World Book Community **21**, 3–4 (2010)
46. Verheij, B.: The toulmin argument model in artificial intelligence. In: Rahwan I (ed) Argumentation in Artificial Intelligence, pp. 219–238. Springer (2009). https://doi.org/10.1007/978-0-387-98197-0
47. Wang, P., Xu, J., Xu, B., Liu, C.l., Zhang, H., Wang, F., Hao, H.: Semantic clustering and convolutional neural network for short text categorization. In: Proceedings ACL 2015 pp. 352–357 (2015). https://doi.org/10.1016/j.neucom.2015.09.096
48. Zhang, Y., Wallace, B.: 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, pp. 253–263 (2017). arXiv:1510.03820
49. Zhao, J., Stockwell, T., Roemer, A., Chikritzhs, T., Bostwick, Dea: Is alcohol consumption a risk factor for prostate cancer? A systematic review and metaanalysis. BMC Cancer **16**(1), 845 (2016). https://doi.org/10.1186/s12885-016-2891-z