Scientific Claims Characterization for Claim-Based Analysis in Digital Libraries

José María González Pinto and Wolf-Tilo Balke

pinto@ifis.cs.tu-bs.de balke@ifis.cs.tu-bs.de
Institut für Informationssysteme
Mühlenpfordstrasse 23 28106 Braunschweig, Germany

Abstract. In this paper, we promote the idea of automatic *semantic characterization* of scientific claims to explore entity-entity relationships in Digital collections. Our proposed approach aims at alleviating time-consuming analysis of query results when the information need is not just one document but an *overview* over a set of documents. With the semantic characterization, we propose to find what we called "dominant" claims and rely on two core properties: the consensual support of a claim in the light of the collection's previous knowledge as well as the authors' assertiveness of the language used when expressing it. We will discuss useful features to efficiently capture these two core properties and formalize the idea of finding "dominant" claims by relying on Pareto dominance. We demonstrate the effectiveness of our method regarding quality by a practical evaluation using a real-world document collection from the medical domain to show the potential of our approach.

Keywords: pareto semantics, scientific claims, skyline

1 Introduction

With the exponential growth in the number of publications, information needs are not always easy to satisfy. Consider Julia who wants to write an overview of research findings regarding 'Ibuprofen' and 'headaches'. This type of query is widespread in scientific digital libraries. For instance, Islamaj et al. [14] analyzed one month of log data from PubMed, consisting of more than 23 million user sessions and more than 58 million user queries. Indeed, the authors found that most PubMed user's type in very few terms (3.54) and more than 30% are token pairs. In fact, more than 55% of the queries are *entity-based* queries such as disease, gene, drug, chemical substance, protein and medical procedure. Moreover, the size of the returned result sets for those queries can be rather difficult to manage: over 10K. Thus, today, Julia would have a time-consuming analysis of the results of the query to build an information space with possible relationships between the entities, find representative papers for each relation and then argue in an informed way to finally accomplish her goal. In other words, Julia is interested in writing a summary of the results of the query and not in a particular document.

How can we help? In this work, we promote the idea of distinguishing between 'dominant' and 'dominated' *scientific claims* to help users that share Julia's information needs.

In a nutshell, scientific claims [9–11] are sentences that contain any association between entities, where one entity has some influencing, manipulating, or even causal relationship to another entity with the additional constraint that they represent the primary contribution(s) of a scientific paper. Indeed, the use of claims as metadata can help to build the information space needed for Julia: each claim contains the relation between the entities! However, little attention is paid to enabling a *semantic characterization* of claims to ease Julia's journey.

In this paper we focus on the discovery of properties to *characterize* claims. To do so, we argue that we should focus on the following properties; a) commitment of the authors: the certainty in the results; b) overall agreement and disagreement between authors concerning current knowledge. Consider the following example to clarify what we mean: *"[…] we performed a preliminary non-randomized clinical trial with 10 participants and our limited data suggests that Ibuprofen can alleviate headaches […]"*. In this example the phrase 'our limited data suggests' directly expresses a rather weak assertiveness of the sentence and –generally speaking- the sentence's structure correspond to specific stylistic and linguistic features casting doubt on the information contained. The second aspect in our running example is the expert assessment of the perceived strength of the relationship between the pair of entities judging the context for the claim given in the document (e.g., slightly weak characteristics of some clinical trial).

Our proposed methodology aims to identify what we have called "dominant" and "dominated" claims. The identification of these two types of claims can ease the creation of a summary in two complementary ways. Firstly, "dominated" claims can help to discover specific aspects of associations between entities that need the development of new hypothesis and the design of new experiments. Secondly, "dominant" claims can help to identify documents which represent aspects that are more certain regarding the association between entities.

We approach the problem of automatically annotating claims as "dominant" and "dominated" in a data-driven fashion. Therefore, we proceed as follows with four necessary steps: first, for a given pair of entities $< e_1, e_2 >$ we rely on high-quality content from a digital library and retrieve a set of documents relevant to the query. Second, from this set of candidate documents, we extract all scientific claims linking the entities. Thirdly, we then proceed to operationalize the properties above to characterize each scientific claim. Then, to formalize the idea of identifying "dominant" and "dominated" claims, we found that the notion of Pareto *dominance* fits naturally. That means: given a choice between two scientific claims, with one claim being better concerning at least one property but at least equal concerning all other properties, the first claim should always be preferred over the second one (the first claim is said to 'dominate' the second claim). Thus, finally, we use this simple but powerful intuitive concept to annotate "dominant" and "dominated" claims. Our contributions are as follows:

- We introduce a novel approach to annotate scientific claims as "dominant" and "dominated" to serve as high-quality building blocks for claim-based summary analysis in Digital Libraries.
- We provide an entirely data-driven approach to operationalize the concept of "dominant" claims.
- We investigate, with evaluation in the context of a real-world digital library, the scope and limitations of our proposed approach.

2 Related Work

Our work draws motivation from the following two areas: the field of argumentation mining and from the research efforts on credibility analysis in social media.

Argumentation mining has a clear focus on modeling and extracting argument structures for different purposes. Currently, efforts to identify argument components, to find evidence for claims, and to predict arguments structures exist. In particular, the work of Habernal et al. [13] is related to our approach. Here, the convincingness of Web arguments is analyzed and the authors show that it is indeed possible to predict the convincingness for a given argument pair concerning some given topic. For this task researchers annotated a large dataset of pairs of arguments over different topics using crowdsourcing. Then the authors used different features and machine learning algorithms to perform two tasks: to determine from a given pair of arguments which one is more convincing (a classification task) and to rank them by convincingness (a regression task). Our work differs from Habernal et al. [13] in three aspects. Firstly, we target scientific digital libraries with clear-cut claims as first-class citizens. Secondly, we use machine learning algorithms only as part of our pipeline but then rely on the semantics of the skyline operator that captures the intuition of "dominance" more naturally. Moreover, we aim at using all the claims to annotate them as "dominant" or "dominated" and let the user use this semantic filter instead of discarding more convincing claims for argumentation. Also related to our work is the ongoing effort to realize argumentation machines of IBM Watson's Debater [29]. Currently, IBM's system relies on handcrafted argumentation structures created by expert users, see Lippi et al. [19].

Credibility analysis. Efforts that account for the credibility of online communities are also relevant to our work. For example, the work of Mukherjee et al. [23] uses a probabilistic graphical model to account for user trustworthiness, language objectivity and credibility of postings in online communities. In fact, we build on the idea of using lexicons that are related to bias in language from their work. In a similar line of thought in Mukherjee et al. [22], researchers studied the credibility of news articles jointly modeled with expert-level users judgments and the trustworthiness of the sources. The work of Castillo et al. [5] focuses on analyzing the credibility of news propagated on Twitter. The authors model credibility as a binary classification task (credible vs. not credible) based on features extracted from user behavior (re-tweeting), the presence of URLs and citations in the tweets to external sources. Kumar et al. [15] studied the credibility of Wikipedia articles to detect false information. Here, researchers addressed the auto-

matic discovery of articles that have fabricated (hoax) entities and events as a classification task with the goal of stopping false articles to remain in Wikipedia. The authors also investigated for how long such articles usually survive, then discussing their impact.

3 Problem Definition

In this section, we formalize our goal to annotate scientific claims as "dominant" and "dominated" in a Digital library.

Definition 1. A **scientific claim** is a natural language sentence in a scientific paper that expresses a specific *relationship* between 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, *'cigarettes'* and *'lung cancer'* are the entities, and the relationship between them is *'increase the risk'.*

A scientific claim involves a specific relationship between entities. The set of relationships considered as relevant is domain-dependent. Let $\mathbb{R} = \{r_1, \dots, r_n\}$ be the set of relevant relations of a given domain of study. For instance: alleviates, causes, and treats.

Pareto Semantics. Here we follow the terminology used by Lofi et al. in [20]. Borzsony et al. in [3] proposed Skyline queries to fill the gap between set-based SQL queries and rank-aware database retrieval. Skyline queries rely on the notion of Pareto semantics from the field of Economics discussed by Gabbay et al. in [7]: some object $o_1$dominates an object o_2, if and only if o_1 is preferred over o_2 with respect to some attribute and o_1 is preferred over or equivalent to o_2 with respect to all other attributes. Formally, the dominance relationship is denoted as $o_1 \succ o_2$. This simple concept has been used in the data base community to implement an intuitive, personalized data filter as dominated objects can be safely excluded, resulting in the so-called *skyline set* of the query.

More formally, let us define dominance relationships following Pareto semantics for every database relation $R \subseteq D_1 \times \dots \times D_m$ over m attributes as follows:

$$o_1 \succ o_2 \Leftrightarrow \exists\, i \in \{1,..,m\}: o_{1i} \succ o_{2i} \wedge \forall\, i \in \{1,..,\mathrm{m}\}: o_{1i} \succcurlyeq o_{2i} \quad (1)$$

Where $o_{j,i}$ denotes the *i*-th component of the database tuple o_j

The skyline set is the set of all non-dominated objects of the database instance R. Let $A = \{a_1, \dots, a_m\}$ be the set of attributes used to characterize claims. Let $\Sigma Claims$ be the set of claims found in a collection of documents D for a given pair of entities $< e_1, e_2 >$. In summary, note that we use the Pareto semantics to identify "dominant" and "dominated" claims to achieve our final goal in this paper: semantic annotation of the claims.

We assume that the claims in each document of a given collection D are part of the metadata available or were found using the adaptation of the TextRank algorithm using embedding representations of sentences by González et al. in [11].

Definition 2. Claim Skyline Set. Let $Claims_{e_1,e_2}$ be an m-dimensional dataset that represents the $\Sigma Claims$ of the entity pair $< e_1, e_2 >$, where greater values are preferred. Then, a claim p in $Claims_{e_1,e_2}$ dominates claim q iff claim p is better than or equal to claim q in all attributes A and is strictly better than claim q in at least one of the dimensions. Now we are ready to define our problem:

Definition 3. Finding Dominant Claims. Given $\Sigma Claims$ of the entities $< e_1, e_2 >$, we attempt to find the Claim Skyline Set under a set of attributes A with respect to an explicit set of relations $\mathbb{R}$.

Once we have found the dominant claims, we can proceed to the semantic annotation of the claims $\Sigma Claims$ for a given entity pair $< e_1, e_2 >$.

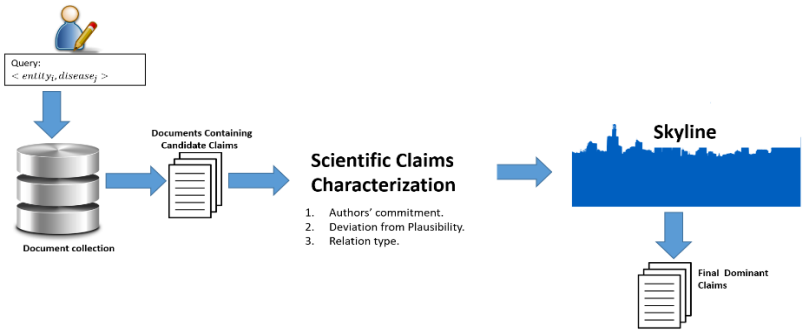

Fig. 1. Approach overview to find dominant claims for a given entity pair

4 Scientific Claim Characterization

In this section we present our approach for determining the attributes A to characterize the claims $\Sigma Claims$ of a given entity pair $< e_1, e_2 >$. We propose to model three core content-based properties to build the following set of attributes: author's commitment; deviation from or similarity to a given plausible claim, and the relation type expressed in the claim. In short, a dominant claim is one that a) expresses high commitment from the authors of the papers: - it is highly objective, and it is unbiased, b) shares similarity concerning a claim that fits the current body of knowledge, and c) expresses a relation type that is mutually exclusive with respect to other relations between the given pair of entities. To this end, we consider a set of attributes to capture these properties. Let's provide the details of how we do it.

Authors' commitment. To capture commitment from the authors, we rely on a set of lexicons from the Natural Language community that proved to model two different stylistic features in written language: concreteness introduced by Brysbaert et al. in [4] and bias introduced by Recasens et al. in [26]. The concreteness lexicon contains ratings of more than 35,000 English words and more than 2,500 bigrams. We hypothesize that this dataset can help our approach to detect the degree of concreteness in the language used by the authors. The lexicon of Recasens et al. in [26] is a set of words that includes

"factive verbs", "implicative", "hedges", and "subjective intensifiers". Recasens et al. argue that the identification of unbiased language is a requirement for reliable sources such as scientific articles or encyclopedia. This lexicon was developed and used to remove bias from Wikipedia articles. For our task, we found hedges –word unigrams and bigrams- useful to model the "strength" of the context of scientific claims. Recasens et al. found that the use of hedges reduces one's commitment to the truth of a proposition. Some instances of this lexicon are: might, likely, may, and perhaps.

Deviation from Plausibility. We build on the notion of plausibility based on the knowledge-fit theory from cognitive sciences by Connell et al. in [6] that we proposed for high-quality content preservation in Digital Libraries in González et al. in [9].

For the sake of completeness, we provide here the intuition of plausibility. The theory from cognitive sciences states that human plausibility judgments consist of two steps: firstly, a mental representation of current knowledge is built and secondly, an assessment examines how a new piece of information fits all prior knowledge. The operationalization of plausibility in an information system is extremely hard in general settings. However, we show in González et al. in [9] that for some specific type of collections where scientific claims are first-class citizens, the approach proved to be useful to support quality content in Digital Libraries. Given that we share the same domain in this work, we rely on the approach to find a document with a plausible claim as detailed in our previous work for each entity pair used in our experiments.

Relation types. To define the relation types worth modeling, we first perform a manual exploration of our dataset. Manual exploration led us to distinguish between relations relevant to the domain, yet incomparable. In particular, we focus on the following 'semantic types': *beneficial*, *non-beneficial*, *no-effe*ct and *unknown*. The following examples will clarify the meaning of each type:

- Example 1*: "Recent studies suggest that occasional drinking of coffee might offer protection from pancreatic cancer."* Example 1 will be mapped to the 'beneficial' semantic type, because -as stated in the claim- drinking coffee offers protection from pancreatic cancer.
- Example 2: *"Coffee consumption may weakly increase the risk of pancreatic cancer."* Example 2 will be mapped to the 'non-beneficial' semantic type because according to the claim, coffee consumption increases the risk of cancer.
- Example 3: *"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."* This example corresponds to the 'no-effect' semantic type.
- Example 4: *"Recent observations of association of risk with coffee consumption and with use of decaffeinated coffee require further evaluation."* This example corresponds to the 'unknown' semantic type.

In summary, given a set of claims $\Sigma Claims$ of a given pair of entities $< e_1, e_2 >$, we aim at automatically identifying the semantic type of the relationship between the entities. In addition to the relation type, the following are the specific attributes A that we used to characterize scientific claims $\Sigma Claims$ to obtain $Claims_{e_1,e_2}$ for a given entity pair $< e_1, e_2 >$. In the features 1-3, $context_i$ of a claim $claim_i$ refers to the abstract of the paper that contains the corresponding claim.

1. Concreteness score: for each claim $claim_i$, we consider the sum of the ratings of each word in the concreteness lexicon occurring in its context $context_i$.
2. Bias score: for each claim $claim_i$, we consider its context $context_i$ to compute the relative frequency of words in the bias lexicon occurring in $context_i$, thus obtaining the bias score:
   $bias = num(words\ in\ lexicon\ present\ in\ context_i)/length(context_i)$.
3. The similarity concerning a plausible claim: cosine similarity of a Topic Model [2] representation of $context_i$, with respect to the plausible claim found. We rely on this popular latent Dirichlet allocation algorithm that assigns probabilistically $context_i$ to different topics in an unsupervised manner. Let $T = \{t_1, \dots t_n\}$ be the set of different topics. Then $context_i$ is represented as a vector representation of these n topics. This representation is used to compute cosine similarity with respect to the $context$ representation of a plausible claim.
4. Distance concerning a plausible claim: word mover's distance of $claim_i$ with respect to the plausible claim found. In other words, given a plausible claim, we compute for each claim $claim_i$ a semantic distance using word mover's distance that has been shown to outperform other approaches using the semantics of word embeddings, see the details of the work of Kusner et al. in [16].

To find a plausible claim and thus compute features 3 and 4, we proceed as follows: let $contradict(claim_i, D)$ and $support(claim_i, D)$ be functions that compute a cumulative sum of similarities concerning the documents whose corresponding claims contradict and support $claim_i$ respectively. We select as plausible the document with the claim that has the highest similarity score difference between supported and contradicted documents.

We chose to work only with these four attributes for two reasons: firstly, to avoid one of the drawbacks of the Skyline operator: the skyline size. Researchers have shown that with datasets with five up to 10 attributes, the skyline set can contain 30% or more of the entire dataset [1, 3, 8]. Secondly, to interpret the results and evaluate the potential of our proposed approach. However, the approach can be applied considering some other aspects of documents, such as citations counts, the prestige of authors, or altmetrics see for instance the work of Priem in [24]. Regardless of the attributes used, the approach can be applied keeping in mind a manageable size of the skyline set for the task at hand.

Finally, to find the Claim Skyline Set within the set of claims and being able to annotate all the claims semantically, we performed two steps: firstly we generated a dataset where each claim is a data point with the features outlined above. Secondly, we applied the skyline operator on the dataset.

5 Evaluation

In this section, we report the evaluation of two aspects of our approach. Firstly, the semantic type detection of scientific claims. In other words, whether a given claim corresponds to one of the four semantic types, we defined in Section 4. Secondly, we eval-

uate to determine the degree of success of our proposed approach to distinguish between scientific claims that are "dominant" or "dominated". Thus, in the following, we detail the document collection, the algorithms, and metrics used in our evaluation.

Document collection. Firstly, to find documents with claims relevant to a pair of entities, we relied on PubMed as our primary data source and used the following query pattern [9] for the entity pair: $< entity, disease >$:

(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) AND (entity AND disease).

To evaluate our proposed approach on a real-world dataset, we used twenty entity pairs from work in nutritional sciences of [9, 27] linking entities investigated in researchers papers concerning their impact on cancer. Therefore, we used the following entities: coffee, tea, salt, lycopene, wine, milk, sugar, potato, pork, onion, olive, lemon egg, corn, cheese, carrot, butter, bread, beef, and bacon. We used the new "Best matches" algorithm from PubMed to retrieve relevant documents to our queries. As stated on PubMed's website, the new algorithm uses machine learning to combine over 150 signals that are helpful to find matching results. In summary, it automatically expands our query pattern to account for synonyms, MeSH terms, and medical terms. The

Table 1. Results of semantic relation types

	P	R	F
Logi BoW	0.84	0.85	0.84
SVM BoW	0.85	0.85	0.84
Logi Embed	0.72	0.72	0.71
SVM Embed	0.66	0.64	0.64

Table 2. Per-class results of the best model

	P	R	F
Beneficial	0.89	0.91	0.85
Non-beneficial	0.82	0.88	0.78
No-effect	0.78	0.78	0.78
Unknown	0.80	0.57	0.67

Table 3. Statistics of the data used to assess performance of the different models

#sentences per category	Number of samples
Beneficial	260
Non-beneficial	161
No-effect	84
Unknown	71

final collection size consisted of 12,616 research papers. The text mining techniques used only the abstracts of the documents retrieved. There were two reasons behind our decision: first, not all retrieved research papers featured full-text access. Thus, to be fair and avoid bias, we decided to rely only on the abstracts. Second, we assume that the abstracts convey enough information and the relevant context to summarize research papers accurately.

a) Semantic relation-type detection. To detect the semantic relation-type automatically, we annotated a set of claims to build a supervised learning system that can predict the semantic type in unseen claims. As basic machine learning algorithms we used logistic regression and support vector machines with the following features: word n-grams (unigrams, bigrams, and trigrams). We call this set of features 'Bag of

Words' features. We compare this approach to the respective word embedding representation of the claims. In particular, we use word embedding based on the models presented in [17, 21]. These vector representations are obtained in an unsupervised fashion from a large textual corpus. They have been used in many applications in text classification systems and serve to support more advanced deep learning models, see [18, 28]. Due to their potential usefulness in similar tasks, we used them to compare to more traditional approaches such as 'Bag of Words'.

In our experiments, we relied on the word2vec vectors trained by Pyysalo et al. [25] on a combination of all publication abstracts from PubMed and all full-text documents from the PubMed Central Open Access subset. Word2vec was run using the skip-gram model with a windows size of 5, hierarchical softmax training, and a frequent word subsampling threshold of 0.001 to create 200-dimensional vectors. Given that these vectors come from a representative collection of the domain that we study, we decided to use them to represent each sentence in our dataset. We refer to the sum representation of the word vectors of each claim using word2vec as 'Embedding' features. This representation of sentences corresponds to the one used in the experiments reported by Lev et al. in [18] that achieved comparable results with more time-consuming deep learning models in different classification tasks. We evaluate the performance of our learned model regarding overall accuracy as well as per-class precision, recall, and F-measure. The results are calculated using 10-fold cross-validation. In Table 1 we summarize the results; 'Logi BoW' and 'SVM BoW' are the models using logistic regression and support vector machines with Bag of Words features respectively; 'Logi Embed' and 'SVM Embed' correspond to the logistic regression and support vector machines models trained with Embedding features. We employ SVM with RBF kernel with one-vs-rest decision function. In Table 2 we show the detailed per-class results of the best model: a support vector machine trained with Bag of Words features. Finally, Table 3 shows statistics of the data used for this first task.

Discussion of the results: To our surprise the Embedding features models were outperformed by its counterparts using simple Bag of Words features. That may due the size of the data that is rather small compared to the number of examples needed according to Goodfellow et al. [12] required to bring significant improvements over more shallow traditional approaches. Nevertheless, the data confirms that traditional machine learning approaches can be used to model relations to guarantee a certain degree of success.

We also observed one limitation of our current approach: we assumed that each claim must fit in one out of the four classes that we have previously defined. This assumption is the reason behind our single-label multi-class problem approach. However, some claims require considering a multi-class multi-label classification approach. For instance, consider the following claim: "*Risk of pancreatic cancer decreased with increasing tea consumption but was unrelated to coffee consumption."* We leave a richer feature engineering approach to solve this issue for future work.

b) Finding dominant and dominated claims. To evaluate our overall approach we performed a Crowdsourcing task on CrowdFlower[1]. We randomly selected 22 claim-

[1] https://www.crowdflower.com/

pairs. Each pair consisted of a claim in the skyline set, in other words, a dominant claim. The other member of each pair was a claim not in the skyline set. For each pair we asked workers to decide which claim was more convincing. Five workers evaluated each pair of claims; we took majority vote as the final judgment and we consider that as our ground truth. To further control quality of the workers we set a minimum of 70% of correct answer concerning the gold standard questions we provided.

The idea of evaluating a pair of argumentative units –in our case scientific claims- in a crowdsourcing environment have been shown to deliver high-quality results. In particular, in the work of Habernal et al. [13], workers evaluated pairs of arguments to decide "which one is more convincing." Motivated by the results of the authors, we considered a similar approach to evaluation. However, two fundamental challenges in our setting are the use of scientific collections instead of web collections and the fact that 'convincingness' is just an approximation of what we are trying to accomplish. Remember that in our setting we hypothesize that a dominant scientific claim has properties that include but are not limited by stylistic features.

When we considered the crowdsourcing results as ground truth, our approach achieved 86% of accuracy. The results of the experiment look indeed promising.

Discussion of the results. We manually examined some of the pairs that were difficult to assess for the workers. Indeed, we found that it was very challenging for the workers because of two reasons: firstly, they were asked not to use any external source to assess which of the pair of claims was more convincing. Secondly, our scientific claims characterization goes beyond the claim itself and these properties, as well as the reasoning behind them, was not available to the workers. This latter observation explains that examples such as the following pair were difficult to evaluate:

- Claim 1: *"In a prospective study of coffee intake with the largest number of pancreatic cancer cases to date, we did not observe an association between total, caffeinated, or decaffeinated coffee intake and pancreatic cancer."*
- Claim 2: *"Based on an analysis of data from the European Prospective Investigation into Nutrition and Cancer cohort, total coffee, decaffeinated coffee, and tea consumption are not related to the risk of pancreatic cancer."*

Nevertheless, the data shows that the attributes we proposed can approximate human interpretation up to a certain degree of accuracy.

6 Conclusions and Future Work

In this work, we promote the idea of finding "dominant" and "dominated" scientific claims to annotate them semantically. We devised a set of features that focused on the claim and its context (as given by the respective abstract of the underlying research paper). We then relied on the intuitive semantics of Pareto dominance and thus applied the skyline operator on the real-world datasets used in our experiments to subsequently derive the annotation for scientific claims. We evaluated our data-driven approach using a set of crowdsourcing tasks and achieved an accuracy above 80% that proves the usefulness of our proposed approach.

For the near future, we foresee work to use our current approach in an application setting to facilitate users the discovery of topics of research in need of more experiments and new hypothesis given our semantic annotation of claims.

References

1. Balke W-T, Zheng JX, Güntzer U (2005) Approaching the Efficient Frontier: Cooperative Database Retrieval Using High-Dimensional Skylines. In: Zhou L, Ooi BC, Meng X (eds) Database Syst. Adv. Appl. 10th Int. Conf. DASFAA 2005, Beijing, China, April 17-20, 2005. Proc. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 410–421
2. Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet Allocation. J Mach Learn Res 3:993–1022. doi: 10.1162/jmlr.2003.3.4-5.993
3. Borzsony S, Kossmann D, Stocker K (2001) The Skyline operator. Proc 17th Int Conf Data Eng 1–20. doi: 10.1109/ICDE.2001.914855
4. Brysbaert M, Warriner AB, Kuperman V (2014) Concreteness ratings for 40 thousand generally known English word lemmas. Behav Res Methods 46:904–911. doi: 10.3758/s13428-013-0403-5
5. Castillo C, Mendoza M, Poblete B (2011) Information credibility on twitter. Proc 20th Int Conf World wide web - WWW '11 675. doi: 10.1145/1963405.1963500
6. Connell L, Keane MT (2006) A model of plausibility. Cogn Sci 30:95–120. doi: 10.1207/s15516709cog0000_53
7. Gabbay DM, Guenthner F (2002) Handbook of Philosophical Logic. Springer Netherlands
8. Godfrey P (2004) Skyline Cardinality for Relational Processing. In: Seipel D, Turull-Torres JM (eds) Found. Inf. Knowl. Syst. Third Int. Symp. FoIKS 2004 Wilheminenbg. Castle, Austria, Febr. 17-20, 2004 Proc. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 78–97
9. González Pinto, J.M.; Balke W-T (2017) Can Plausibility Help to Support High Quality Content in Digital Libraries? TPDL 2017 – 21st Int. Conf. Theory Pract. Digit. Libr.
10. González Pinto JM, Balke WT (2017) Result Set Diversification in Digital Libraries Through the Use of Paper's Claims. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 10647 LNCS:225–236. doi: 10.1007/978-3-319-70232-2_19
11. González Pinto JM, Balke WT (2017) Offering answers for claim-based queries: A new challenge for digital libraries. In: Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). pp 3–13
12. Goodfellow I, Bengio Y, Courville A (2016) Deep Learning. MIT Press 521:800. doi: 10.1038/nmeth.3707
13. Habernal I, Gurevych I (2016) Which argument is more convincing? Analyzing and predicting convincingness of Web arguments using bidirectional LSTM.

In: ACL. pp 1589–1599
14. Islamaj Dogan R, Murray GC, Névéol A, Lu Z (2009) Understanding PubMed® user search behavior through log analysis. Database. doi: 10.1093/database/bap018
15. Kumar S, West R, Leskovec J (2016) Disinformation on the Web: Impact, Characteristics, and Detection of Wikipedia Hoaxes. Www 591–602. doi: 10.1145/2872427.2883085
16. Kusner MJ, Sun Y, Kolkin NI, Weinberger KQ (2015) From Word Embeddings To Document Distances. Proc 32nd Int Conf Mach Learn 37:957–966.
17. Le Q, Mikolov T (2014) Distributed Representations of Sentences and Documents. Int Conf Mach Learn - ICML 2014 32:1188–1196. doi: 10.1145/2740908.2742760
18. Lev G, Klein B, Wolf L (2015) In defense of word embedding for generic text representation. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 9103:35–50. doi: 10.1007/978-3-319-19581-0_3
19. Lippi M, Torroni P (2016) Argumentation Mining: State of the Art and Emerging Trends. ACM Trans Internet Technol 16:10. doi: 10.1145/2850417
20. Lofi C, Balke W-T (2013) On Skyline Queries and How to Choose from Pareto Sets. In: Catania B, Jain LC (eds) Adv. Query Process. Vol. 1 Issues Trends. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 15–36
21. Mikolov T, Corrado G, Chen K, Dean J (2013) Efficient Estimation of Word Representations in Vector Space. Proc Int Conf Learn Represent (ICLR 2013) 1–12. doi: 10.1162/153244303322533223
22. Mukherjee S, Weikum G (2015) Leveraging Joint Interactions for Credibility Analysis in News Communities. In: Proc. 24th {ACM} Int. Conf. Inf. Knowl. Manag. . pp 353--362
23. Mukherjee S, Weikum G, Danescu-Niculescu-Mizil C (2014) People on drugs: credibility of user statements in health communities. KDD '14 Proc 20th ACM SIGKDD Int Conf Knowl Discov data Min 65–74. doi: 10.1145/2623330.2623714
24. Priem J (2014) Altmetrics. In: Beyond Bibliometr. Harnessing Multidimens. Indic. Sch. Impact. pp 263–287
25. Pyysalo S, Ginter F, Moen H, Salakoski T, Ananiadou S (2013) Distributional Semantics Resources for Biomedical Text Processing. Proc. LBM 2013
26. Recasens M, Danescu-Niculescu-Mizil C, Jurafsky D (2013) Linguistic Models for Analyzing and Detecting Biased Language. Proc 51st Annu Meet Assoc Comput Linguist 1650–1659.
27. Schoenfeld JD (2013) Is everything we eat associated with cancer ? A systematic. Am J Clinincal Nutr 97:127–134. doi: 10.3945/ajcn.112.047142.1
28. Zhang Y, Wallace B (2015) A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification. 253–263.
29. IBM Debating Technologies. http://researcher.watson.ibm.com/researcher/view_group.php?id=5443. Accessed 11 Oct 2017