HHS Public Access

Author manuscript

J Biomed Inform. Author manuscript; available in PMC 2020 July 13.

Published in final edited form as:

J Biomed Inform. 2019 March ; 91: 103123. doi:10.1016/j.jbi.2019.103123.

Confirm or Refute?: A Comparative Study on Citation Sentiment Classification in Clinical Research Publications

Halil Kilicoglu[a,*], **Zeshan Peng**[a], **Shabnam Tafreshi**[b], **Tung Tran**[c], **Graciela Rosemblat**[a], **Jodi Schneider**[d]

[a]Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, MD, 20894

[b]Department of Computer Science, George Washington University, Washington, DC, 20052

[c]Department of Computer Science, University of Kentucky, Lexington, KY, 40506

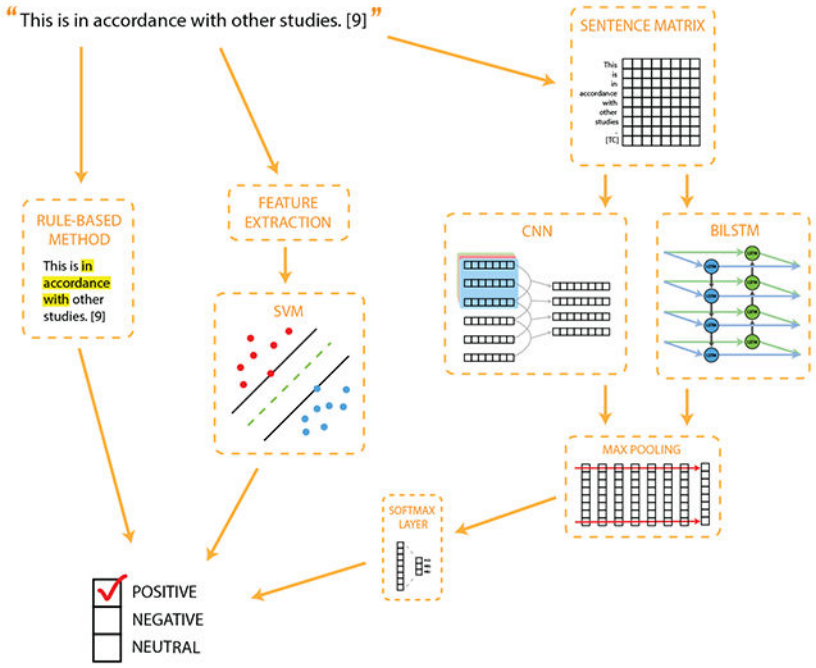
[d]School of Information Sciences, University of Illinois, Champaign, IL, 61820

Abstract

Quantifying scientific impact of researchers and journals relies largely on citation counts, despite the acknowledged limitations of this approach. The need for more suitable alternatives has prompted research into developing advanced metrics, such as h-index and Relative Citation Ratio (RCR), as well as better citation categorization schemes to capture the various functions that citations serve in a publication. One such scheme involves *citation sentiment*: whether a reference paper is cited positively (agreement with the findings of the reference paper), negatively (disagreement), or neutrally. The ability to classify citation function in this manner can be viewed as a first step toward a more finegrained bibliometrics. In this study, we compared several approaches, varying in complexity, for classification of citation sentiment in clinical trial publications. Using a corpus of 285 discussion sections from as many publications (a total of 4,182 citations), we developed a rule-based method as well as supervised machine learning models based on support vector machines (SVM) and two variants of deep neural networks; namely, convolutional neural network (CNN) and bidirectional long short-term memory (BiLSTM). A CNN model augmented with hand-crafted features yielded the best performance (0.882 accuracy and 0.721 macro-F_1 on held-out set). Our results show that baseline performances of traditional supervised learning algorithms and deep neural network architectures are similar and that hand-crafted features based on sentiment dictionaries and rhetorical structure allow neural network approaches to outperform traditional machine learning approaches for this task. We make the rule-based method and the best-performing neural network model publicly available at: https://github.com/kilicogluh/clinical-citation-sentiment.

Graphical Abstract

*Corresponding author.



Keywords

citation analysis; sentiment analysis; natural language processing; supervised machine learning; neural networks

1. Introduction

Citation-based metrics, such as *journal impact factors* [1], *g-index* [2], and *h-index* [3] are widely used to quantify and evaluate scientific impact of journals and scientists [4]. Academic institutions use such metrics as proxies for the quality of a scientist's output and they contribute to decisions about tenure and promotion. Funding agencies also take these metrics into account for awarding grants and justifying renewals. Citation-based metrics primarily rely on citation counts, and assume that each article contains the same level of scientific contribution and that each citation is equally important [5]. The flaws of these assumptions have been widely discussed, and a strong consensus has emerged in recent years against an overreliance on such metrics for measuring scientific impact [6, 7]. However, their use remains common practice. In recent years, more sophisticated new metrics, such as *Relative Citation Ratio (RCR)* [7], have been proposed, even though such alternatives also rely on citation frequency, and are thus unable to account for the qualitatively different ways in which an article may be cited. For example, a publication can be cited as support for findings, or its findings can be disputed; it can be cited as a key point of departure for the current work or as "homage to pioneers" [5]. Making such distinctions can not only provide the basis for better evidence assessment and synthesis, but also help to develop more accurate bibliometric measures.

Citation content analysis research focuses on describing and classifying semantic relationships between the citing and the reference (or cited) works [8, 9], providing a more nuanced, qualitative view of how an article is cited. This has the potential to enhance bibliometric measures and allow stakeholders to make more informed decision about scientific contribution and impact. Many classification schemes to describe the relationships between the citing and the reference papers have been developed over time. Relatively coarse-grained classifications involve categorizing citations based on whether or not they are significant for the citing work [10, 11] or based on the sentiment expressed in the citing paper towards the reference: whether it is being criticized or its findings refuted (negative), cited in a positive light or its findings confirmed (positive), or cited neutrally [12, 13]. More fine-grained classification schemes involve *citation function*, that is, the role that the citation plays in citing paper, for which many classifications with varying degrees of complexity have been developed [14, 15, 16, 17, 18]. For example, Teufel et al.'s [16] citation function classification includes categories, such as PBAS (author uses reference work as starting point) and COCO-(unfavorable comparison/contrast with citing work). Some of these classification schemes are purely descriptive [14, 15], others have been used for corpus annotation [16, 17, 18], and a limited number of them for automatic citation content analysis [17, 18, 19].

Most citation content analysis research has focused on the computational linguistics literature ([12, 16, 19, 20], among others). In recent years, the biomedical literature has also garnered some attention [13, 17, 21]. In one of these studies, Xu et al. [13] presented a corpus of clinical trial articles annotated for citation sentiment, and experimented with support vector machine (SVM) models and a variety of features to identify these categories. In addition to enhancing bibliometric measures, they proposed that citation sentiment analysis can enable citation bias analysis [22] and help detect non-reproducible studies. While drawing such conclusions based on citation sentiment analysis alone may not be straightforward, three-way citation sentiment classification (positive/negative/neutral) strikes a balance between the widely-used quantitative metrics and complex citation function schemes, providing a qualitative assessment at a granularity that is easier to annotate and reach consensus on than citation function categorizations [23]. Furthermore, sentiment polarity is often implicit in fine-grained citation function categories [23]. For example, the COCO- (unfavorable comparison/contrast with the citing work) category in Teufel et al.'s [16] classification is implicitly negative.

In this paper, we focus on developing methods to automatically recognize citation sentiment in biomedical literature. We take Xu et al.'s [13] work as the starting point and use their corpus of clinical trial articles annotated for citation sentiment for training and evaluation. We experiment with methods of varying complexity: a relatively simple rule-based method, a traditional supervised machine learning method similar to that employed by Xu et al. [13], and two deep neural network variants. We also conduct experiments to assess the effect of *citation context* (i.e., the passage in the citing paper that discusses the reference paper) and *citation subject matter* (i.e., the part of the citation context that is about the topic of the reference paper) on citation sentiment analysis. Additionally, we experiment with several sampling techniques to address corpus imbalance, and we compare the generalizability of the methods, using a portion of the dataset as a held-out test set. A convolutional neural network (CNN) model that relies on dependency-enhanced word as well as part-of-speech

(POS) embeddings and hand-crafted phrase-level features yields the best classification performance. Our results also indicate that taking into account manual citation context and subject matter annotations degrades the performance, and that undersampling instances belonging to the majority class (neutral sentiment) has limited effect on classification.

2. Related Work

In this section, we discuss some of the relevant research in automatic citation content analysis (see recent reviews [21, 24] for in-depth analyses of this topic).

Most automatic citation analysis research has focused on the computational linguistics literature, owing to the availability of a corpus of full-text articles (ACL Anthology Network Corpus) [25]. This corpus has been used to study *citation significance* [10], *citation sentiment* [12, 26], as well as *citation context* [20, 27]. Citation significance and sentiment have been addressed as classification tasks, while citation context recognition has been studied both as a classification and a sequence labeling task. Conventional supervised machine learning techniques with various lexical, syntactic, and citation-specific features have been applied. For example, in his study on citation significance, Athar [10] used as features the similarity of the citation sentence (i.e. the sentence containing an explicit citation) with the title of the article as well as number of sentences with acronyms, with formal citation to the paper, and to the author's name, and achieved the best results with a Naive Bayes classifier (0.55 F_1). In an earlier study on citation sentiment recognition [12], he used an SVM classifier with n-gram and dependency relation features from the citation sentence, obtaining a macro-F_1 score of 0.76. Qazvinian and Radev [27] studied detection of *citation context*, the span of text that discusses the reference paper in the citing paper. Their approach, based on Markov Random Fields, used sentence similarity and lexical features to detect the neighboring sentences that are part of the citation context, and improved extractive summarization due to better citation context recognition. From a different angle, Abu-Jbara and Radev [20] investigated citation contexts that are shorter than the full sentence. They compared several classification and sequence labeling approaches, achieving the best results (0.87 F_1) with a Conditional Random Fields (CRF) model augmented with rule-based post-processing. The effect of accurate, multi-sentence context identification on citation sentiment has also been investigated [26, 28], with contradictory results. While Abu Jbara et al. [28] found that citation sentiment recognition performance improved with ground truth context information (specifically for polarized citations), Athar and Teufel [26] found that classification without multi-sentence context (i.e., citation sentence only) performed better, attributing this to data sparsity resulting from larger citation contexts.

Several studies involved other smaller corpora, while still focusing on computational linguistics articles. In early work, Teufel et al. [16] developed a fine-grained *citation function* classification consisting of 12 categories (e.g., WEAK (weakness of the cited approach), PBAS (author uses cited work as starting point)) and presented a corpus annotated with these categories. Inter-annotator agreement was found to be 0.72 (Cohen's κ). In a follow-up study [19], they used a memory-based learning algorithm to recognize these categories, achieving Cohen's κ of 0.57. Features used included cue phrases in the citation sentence, position of the citation, and whether the citation was a self-citation. More recently,

Zhu et al. [11] focused on the notion of *citation influence* (similar to significance), building a corpus by surveying authors about the most influential citations in their papers. Based on this corpus, they built SVM models using features based on counts, similarity, position, and context. Their results indicate that the number of times the reference paper is cited is the most predictive feature of citation influence. Hernández-Alvarez et al. [18] proposed a composite classification scheme based on four dimensions (function, polarity, influence, and aspects), the latter referring to key phrases indicating citation use. They suggested that this approach simplified annotation while providing adequate information for qualitatively measuring the impact of a reference paper in a citing paper. They reported inter-annotator agreement of 0.86 and 0.91 for citation function and polarity, respectively. They obtained F_1 scores of 0.89 and 0.93 in recognizing these dimensions, using SVM models based on aspect (i.e., key phrase) features.

In the biomedical domain, citation content analysis is a relatively understudied research area. Few existing studies focus on different corpora and tasks, also relying primarily on supervised machine learning approaches. In early work, Agarwal et al. [17] annotated a corpus of 43 biomedical articles with eight citation function categories (Background/Perfunctory, Contemporary, Contrast/Conflict, Evaluation, Explanation of results, Material/Method, Modality, Similarity/Consistency), obtaining moderate inter-annotator agreement (κ=0.63). Using n-gram features with SVM and Naive Bayes models, they achieved a macro-F_1 score of 0.75. Xu et al. [13] investigated citation sentiment, annotating the discussion sections of 285 clinical trial articles. They trained SVM models using as features n-grams as well as sentiment lexicons and higher-level phrase categories indicating rhetorical structure (e.g., CITING_WORK, NEGATION, CONTRAST), obtaining a macro-F_1 score of 0.72. Jia [21] developed a two-level annotation scheme, the top level concerning citation polarity, and the bottom-level involving eight citation function categories (e.g., Confirmation, Contrast/Conflict). She annotated a corpus of 1823 citation sentences, obtaining moderate intra-annotator agreement (κ=0.71). Experimenting with various supervised machine learning techniques and feature combinations, she achieved the best results with POS tags, n-grams, and dependency features and linear SVM models (0.73 F_1 score for citation function and 0.78 F_1 score for citation polarity). It is worth noting that the schemes proposed by Agarwal et al. [17] and Jia [21] both address citation sentiment to some extent, as they incorporate categories of Confirmation and Contrast, subsumed by the positive and negative sentiment classes of Xu et al. [13], Abu Jbara et al. [28] and others.

3. Materials and Methods

In this section, we first briefly describe the corpus that we used in this study, developed in previous work [13]. We also describe additional annotation we performed on this corpus to better understand the role of citation context in citation sentiment classification. Next, we provide details about the methods we used: a rule-based method that is based on scoring key phrases, SVM models with a variety of features discussed in earlier work, and two deep neural network variants: a convolutional neural network (CNN) model and a bidirectional long short-term memory (BiLSTM) model, both being increasingly used for natural language processing tasks.

3.1. Annotated Corpus

For our experiments, we used a corpus of 285 clinical trial articles with 4,182 citations annotated for citation sentiment (Positive, Negative, and Neutral) [13]. We obtained the corpus from the primary author of the corresponding publication via personal communication[1]. In this corpus, only the discussion sections of the articles have been annotated, since these sections have a fairly significant rhetorical component and are thus more likely to contain opinionated citations. The annotation scheme was presented as a decision tree (reproduced in Supplementary File 1). The corpus was double-annotated and adjudicated. The inter-annotator agreement was reported to be 0.504 (Cohen's κ), indicating moderate agreement. The basic statistics of the corpus is shown in Table 1. Similar to other citation sentiment corpora, this corpus is also imbalanced, with most citations being Neutral.

It is important to reiterate at this point that in the context of citation analysis, sentiment is broadly used to include agreement (positive sentiment) or disagreement (negative) of the findings in the citing paper with those of reference paper, rather than indicating a value judgment about the reference paper (good or bad). The examples for sentiment classes given in Table 2 illustrate this point. Key sentiment-related phrases are underlined.

3.2. Citation context and subject matter annotation

One of our goals in this study was to more accurately assess the effect of citation context on classification and isolate the effect of citation subject matter. The effect of context size on classification, generally in terms of number of sentences surrounding the citation sentence, has been considered in previous work (e.g., [27, 26]), with inconclusive results. In principle, it also seems plausible that discussion of subject matter of the reference paper would lead to noisy features, as citation sentiment is often indicated with a few rhetorical phrases that are distinct from the clauses discussing biomedical topics such as treatments and adverse effects. This intuition also underpinned earlier work in unsupervised learning of rhetorical structure in scientific writing [29]. Consider the example with Negative sentiment in Table 2. The citation sentence does not contain any explicitly negative phrase, while it seems crucial to recognize that the previous sentence (*These results differ from two previous studies.*) is part of this citation's context. Furthermore, the subject matter (discussion of glucose control and morbidity/mortality) seems essentially irrelevant for citation classification.

In view of these considerations, we also conducted an exploratory annotation of citation context and subject matter on this corpus. We defined *citation context* as "the spans of text that are relevant to understanding the contribution of a particular citation to the article in consideration". We also stipulated that citation context should be interpretable in isolation and that it can consist of one sentence, a fragment of a sentence, or a set of sentences (either contiguous or non-contiguous). *Citation subject matter* was defined as "the main substance/argument/claim of the reference paper that the citing paper refers to or compares/contrasts with". It is the same as or is subsumed by the citation context and is often a factual/neutral statement attributed to the author of the reference paper. In the Negative example in Table 2, the subject matter was annotated as *patients in a cardiovascular ICU [5] demonstrated a*

[1]While the corpus is currently not publicly available, it can be requested from the authors of that paper.

decrease in morbidity and mortality, whereas in the Neutral example, the subject matter is the same as the context, the full sentence.

In the first step of annotation, four authors (HK, ST, GR, JS) annotated and reconciled seven articles, and developed annotation guidelines. Next, 30 articles were annotated by three authors (HK, GR, JS) according to the guidelines and reconciled by two (HK, GR). Then, the rest was annotated by a single annotator (GR). On the 30 articles that were annotated independently based on the guidelines, we calculated inter-annotator agreement using F_1 measure, when one set of annotations was taken as the ground truth. We considered both exact and partial matches of annotations. With exact match, mean average agreement was low (context: 0.56, subject matter: 0.33). Considering partial matches, we observed much higher agreement (0.83 for both context and subject matter), which suggests that the main difficulty in annotating these elements is identifying their precise boundaries. The resulting set of annotations can be considered a "silver" standard, since it was not fully adjudicated.

3.3. Rule-based method

Our rule-based method generates a *cumulative sentiment score* for each citation c ($cscore(c)$) and uses thresholding to determine the sentiment class for the citation. Sentiment scores are generated using a list of key phrases, each of which (p) is assigned a pre-computed score based on their occurrences in positive and negative citation contexts ($pscore(p)$). An initial phrase list was constructed, based on the dictionary developed by Teufel [5] and categorized into various relevant classes (e.g., Comparison, Contrast, Good, Negation) in her work on citation analysis. We augmented this list with positive and negative word lists used by Xu et al. [13], resulting in a total of 570 phrases. We manually reduced this list to 189 phrases, selecting those explicitly indicating positive or negative assessment. We then calculated the number of times each of these phrases occurred in the context of citations with positive vs. negative sentiment in the training set. If an explicit negation marker (e.g., *no, not*) appeared in a window size of 2 preceding the phrase p, the opposite sentiment was recorded. Citation context was taken to be the single sentence containing the citation. We calculated a score for the phrase with the formula given below:

$$pscore(p) = (n_{POS}(p) - n_{NEG}(p))/(n_{NEU}(p) + 1)$$

$n_{POS}(p)$ indicates the number of times the phrase p appears in a positive context, $n_{NEG}(p)$ the number of times it appears in a negative context, and $n_{NEU}(p)$ the number of times it appears in a neutral context. 41 terms were assigned a sentiment score of zero, leaving us with 148 phrases.

We then calculated a cumulative sentiment score for each citation ($cscore(c)$), summing up the scores associated with the relevant phrases in the context. The scores were adjusted in three ways:

- If the relevant phrase p appears as sentence-initial, and is a contrastive discourse marker (e.g., *however, nevertheless, despite*), it is excluded from the calculation.

- If there is an explicit negation marker in a window size of 2 preceding *p*, we reverse the sign of its score ($-pscore(p)$).
- If the score resulting from the phrase is less than zero, we apply a weight to this score by a factor based on the distance between the phrase and the citation. This factor is given as: $log_2 d(p, c)$, where $d(p, c)$ indicates the distance between *p* and *c* in terms of lexical units.

The citation sentiment value is determined from the resulting cumulative score $cscore(c)$ as follows. The thresholds for POS and NEG classes were determined using the training set.

$$sentiment(c) = \begin{cases} \text{POS}, & \text{if } cscore(c) > 1 \\ \text{NEG}, & \text{if } cscore(c) < -0.1 \\ \text{NEU}, & \text{otherwise.} \end{cases}$$

3.4. Support vector machines (SVM)

The goal of SVM experiments was four-fold: a) to reproduce the results of Xu et al. [13], b) to experiment with additional features, c) to use additional citation context and subject matter annotations and assess their effect on citation sentiment classification, and d) to address class imbalance using undersampling techniques.

To reproduce the results of Xu et al. [13], we re-implemented their features based on their description. They reported three types of features, briefly explained below. All features consider the elements within the citation context, which they defined as "the citation sentence plus the next sentence, unless the next sentence contains another citation or starts with a contrastive discourse marker".

- *n-gram features*
  - All unigram, bigram, and trigrams in the context
  - Total number of negation phrases in the context
  - For a word within negation scope (i.e., word within a window of 2 words following a negation term), *NOT_word* features
- *Sentiment lexicon features* based on positive and negative lexicons derived from the corpus (53 and 46 phrases, respectively)
  - The presence/absence of a sentiment phrase with its polarity (opposite polarity if the phrase is within negation scope)
- *Structure features* based on rhetorical phrase lists for citing work (e.g., *Our data*), comparatives, negation, sentiment, and contrastive markers
  - Unigrams, bigrams, and trigrams of the phrase types recognized in the context (e.g., CITINGWORK_CONTRAST_POS trigram)
  - If there is a CONTRAST phrase, the direction of this phrase with respect to the other phrases (e.g., CITINGWORK_CONTRAST and CONTRAST_POS for the trigram above)

Like Xu et al. [13], we used LIBLINEAR [30] implementation of SVM with default parameters (*L2*-regularized, *L2*-loss classification, regularization parameter $C = 1$). We used Weka release 3.8.2 in our experiments.

To improve classification, we also considered additional features proposed in prior work on citation sentiment, including features based on POS tags and syntactic dependency relations. Syntactic dependency features were found to be useful in Athar et al.'s [12] and in Jia et al.'s work [21], while POS tags were found useful in Jia's work [21]. We computed syntactic dependencies using the Stanford Core NLP toolkit [31][2], and represented each syntactic dependency in the citation context as a feature. We generated two types of POS features. First, POS n-grams (unigram, bigram, and trigrams) used POS tag sequences as features. Secondly, we concatenated token n-gram features with their POS tags, to address ambiguity to some extent.

We augmented the resulting set of features in two ways. First, for features involving tokens (n-gram, n-gram-POS pair, syntactic dependency features), we used lemmas instead of tokens. Secondly, we used the prediction made by the rule-based method as an additional feature. Features and their illustration on an example instance are shown in Table 3. The surrogate [TC] indicates the citation under consideration, and [OC] indicates other citation in the context.

We used citation context/subject matter annotations in our experiments as follows: first, instead of using the heuristics for determining the context as proposed by Xu et al. [13], we used the "silver" citation contexts annotations. Second, we simply substituted the text spans corresponding to subject matter annotations with the surrogate [SUBJECT_MATTER] before computing the features described above.

3.4.1. Undersampling—Since the majority of instances in the dataset are neutral in sentiment, we also conducted experiments that undersampled those instances. In one experiment, we randomly selected a subset of neutral examples to reach a predefined ratio of positive and negative examples versus neutral instances to use for training. Considering that the proportion of the total number of POS and NEG instances to the number of NEU instances in the dataset is 1:3.15, we used ratios of 1:1 and 1:2 in undersampling. In another experiment, we used the Easy Ensemble method for undersampling [32], in which we divided the neutral instances randomly into five subsets, trained five models by augmenting each neutral subset with all the positive and negative instances, and used majority voting among these trained models to predict the sentiment class.

3.5. Deep neural network models

In recent years, approaches based on deep neural network (NN) architectures in conjunction with distributed text representations have dominated the NLP field, including text classification tasks [33]. To evaluate these approaches for citation sentiment classification, we experimented with two well-known NN architectures, convolutional neural networks (CNNs) [34] and bidirectional long short-term memory networks (BiLSTMs) [35]. We also

[2]We also used this toolkit for tokenization and POS tagging.

evaluated several distributed text representations and their combination as input to these models: word embeddings [36], dependency-based word embeddings [37], and POS embeddings. In addition, we investigated the effect of augmenting the models with predictive hand-crafted features that were used for SVM models.

Below, we first briefly describe the distributed text representations used in this study, and then provide a high-level overview of the deep neural network models.

3.5.1. Distributed text representations—In this work, NN models use the sentence in which a given citation mention occurs as the input context for that mention. To generate input for these models, each citation sentence is tokenized and represented as a *sentence matrix*, where the rows correspond to dense vector representations of each token in the sentence. For a sentence that consists of s tokens and a vector dimensionality of d, the dimensionality of the sentence matrix is $s \times d$.

Most commonly, word embeddings learned from text corpora using unsupervised methods such as word2vec [38] are used as distributed text representations. The core idea underlying word embeddings has also been generalized to generate other types of embeddings, including character, POS, or dependency-based embeddings. We considered word, POS, and dependency-based embeddings in this work, based on our observation that such information proved useful as features in SVM experiments.

As word embeddings, we used the publicly available biomedical word embeddings pre-trained on PubMed abstracts and PubMed Central articles using word2vec [39][3]. These embeddings consist of 200-dimensional vectors ($d = 200$) and were generated using the skip-gram model with a window size of 5, hierarchical softmax training, and a frequent word subsampling threshold of 0.001. During our training process, word vectors were further modified via backward propagation (i.e., non-static embeddings).

In the absence of similarly pre-trained part-of-speech vectors, we generated POS embeddings ourselves from the full-text of 2200 clinical trial articles retrieved from PubMed Central in addition to the original training data for citation sentiment analysis. To generate these embeddings, POS tags associated with the tokens were used instead of the tokens themselves. The gensim[4] implementation of the word2vec algorithm is used to generate 30-dimensional part-of-speech vectors using the skip-gram model, a window size of 5 and a sample count of 5 for negative sampling.

As dependency-enhanced word embeddings, we used 300-dimensional vectors pre-trained on English Wikipedia corpus [37]. These embeddings were generated using the skip-gram model with a sample count of 15 for negative sampling. When so indicated, these dependency-enhanced embeddings substitute the pre-trained word embeddings in our experiments. As with word embeddings, both POS and dependency-enhanced word embeddings are further modified during training.

[3] http://bio.nlplab.org/
[4] https://radimrehurek.com/gensim/

3.5.2. Convolution neural networks—Convolutional networks (CNNs) are a type of feed-forward neural network, originally developed for image recognition tasks [34, 40]. More recently, they have also been successfully applied to various text classification tasks [33, 41, 42]. In a text classification scenario, a CNN essentially identifies informative localized patterns in the text, and combines them to produce a fixed size feature representation of the text. To achieve this, CNNs use *convolution* and *pooling* operations. Convolution operation involves a *convolution filter*, which is applied to a window of words in the sentence (i.e., filter region) to produce a feature map representation for that window. 1-max pooling operation then identifies the most important feature corresponding to each filter. These features are combined and passed to a fully connected softmax layer whose output is the probability distribution over labels. We refer the reader to [33, 41] for a more formal and in-depth discussion of the CNN architecture.

Kim [41] proposed a CNN with a single convolution layer (one-layer CNN) that achieved state-of-the-art results on several well-known text classification datasets. Owing to its relative simplicity and good performance, one-layer CNN architecture has been proposed as a drop-in replacement for traditional supervised machine learning approaches, such as SVM [43]. We adopted this one-layer CNN architecture in this study. The following model configuration was used in training the CNN model: activation function (rectified linear units), filter region size (3,4,5), feature maps (200), dropout rate (0.5), and mini-batch size (16). A random 20% of the training data serves as the development set. We trained the model for a maximum of 20 epochs while checkpointing on each epoch. The final network parameters correspond to the checkpoint that performed best on the development set. We illustrate our CNN architecture in Fig. 1.

3.5.3. Bidirectional long short-term memory—While representations derived from CNNs offer some sensitivity to word order, they are restricted to mostly local patterns, and cannot account for the order of patterns that are far apart in a sequence. Recurrent neural networks (RNNs) [44] have been proposed specifically for sequential data and can learn arbitrarily long-distance dependencies. Accordingly, they have been primarily applied to sequence modeling tasks in NLP, such as language modeling, named entity recognition, and relation extraction [33]. However, some text classification studies also demonstrated superior performance with RNN-based architectures [45]. Therefore, we considered such an architecture in this work, as well.

RNNs employ cyclical connections to recurrently compose context vectors of a sentence from left to right, accumulating information from the previously seen word sequence. The input to an RNN at each time step is a vector representation of the next word in text and the output is a vector representation of the context, including the current word and all the previous words. The output of the hidden layer is fed back to itself at consecutive time steps.

Long short-term memory (LSTM) network [35] is a type of RNN that aims to address some of the limitations associated with the basic RNN architecture, primarily the vanishing gradients problem. An LSTM network substitutes the conventional hidden layer of an RNN with an LSTM unit, which consists of a *memory cell* and three gates that control the information flow into and out of the cell: *input gate*, *forget gate*, and *output gate*. The

memory cell remembers values over arbitrary time intervals. The input gate limits the extent of information from the next input to incorporate into the memory, the forget gate controls how much of the current memory is to be forgotten, and the output gate determines how much of the information from the current memory cell to propagate to the output state.

A modification of the LSTM architecture, called bidirectional LSTM (BiLSTM) builds on the notion that words following the current word may also be useful for prediction tasks, allowing to look arbitrarily far at both the past and the future. A BiLSTM network consists of two separate RNNs, each using its own parameters and capturing the context from one direction (left-right or right-left). To generate the output for each time step, output vectors from each RNN are typically concatenated. We refer the reader to [33] for a more formal and in-depth discussion of RNN and related architectures.

In this work, we use the BiLSTM variant described in Kavuluru et al. [46]. This variant augments the BiLSTM with a subsequent 1-max pooling layer, which generates one feature vector for each sentence, and a fully-connected softmax layer that outputs probability distributions. The following experimental setup was used in training the BiLSTM model: hidden size (200), dropout rate (0.5), and mini-batch size (16). Unlike Kavuluru et al. [46], we did not include position vectors (encoding the distance of a word to the citation mention) as input, as we did not observe any benefit from their inclusion. The overall training procedure of the BiLSTM model mirrors that of the CNN model. The BiLSTM architecture is illustrated in Fig. 2.

3.5.4. Additional high-level features—

To determine whether NN models can be augmented with high-level contextual features (as opposed to word-level features) to improve performance further, we concatenated a feature representation derived from some features used by the SVM model to the feature representation derived from neural models before the fully connected softmax layer. The features incorporated in this way are: part-of-speech n-grams (POSGRAM), negation count, negated unigrams, sentiment lexicon features (SENT), structure features (STRUCT), and rule-based output. These feature sets corresponded to a total of 4539 features in SVM training. Using principal component analysis (PCA), we reduced the dimensionality of this feature representation to 200 to ensure that these features do not overwhelm the features derived using neural architectures.

3.6. Evaluation

We performed two types of evaluation. First, we measured the performance of SVM and NN models using 10-fold cross-validation, following Xu et al. [13]. Second, we measured performance of these models using a 80-20 training/test split of documents. This was done to a) provide a fair comparison to the rule-based method, which was based on an analysis of the 80% of the corpus documents, and b) to assess the generalizability of SVM/NN models and various features to unseen data. The evaluation metrics used were overall accuracy and macro-F_1 scores, as well as precision, recall, and F_1 scores for each sentiment class. Unlike Xu et al. [13], we do not report micro-F_1, since trends with this metric were the same as those observed with accuracy.

Since the deep NN training process is highly stochastic in nature, different models with different randomly initialized parameters typically result in slightly different performance. One solution is model averaging, where the final class is predicted based on the average of probability estimates of multiple models trained using the same architecture with different parameter initializations with the goal of reducing variance and improving performance. In this work, we use model averaging to report NN performance. To do this, we trained a pool of 20 models each with different random parameter initializations and the same 80-20 split for the training and development set. The final reported evaluation measures are based on averaging the individual performance of 20 "runs" where each "run" corresponds to an ensemble of ten models randomly sampled from the pool. The ensembling aspect improves stability, while reporting mean-performance over "runs" allows us to draw comparisons based on average model behavior.

4. Results

In this section, we first provide the results of 10-fold cross-validation experiments. Results obtained with SVM models are presented and compared to earlier results reported in Xu et al. [13]. We also examine the effect of citation context/subject matter annotations and various sampling approaches on performance of the best SVM model. Next, we present the results obtained with NN models. We conclude this section by providing a comparison of rule-based, SVM, and NN methods on a held-out test set.

4.1. SVM experiments

The results of SVM experiments with 10-fold cross-validation are provided in Table 4. To a large extent, we were able to reproduce the results of Xu et al. [13], matching their accuracy, while obtaining slightly better results for NEG and NEU classes. On the other hand, the replication results were very slightly lower for macro-F_1 and the pos class. These minor differences are likely due to possible differences in the pre-processing steps, such as tokenization, or in phrase lists, as well as the precise cross-validation folds used.

Using lemmas instead of tokens (NGRAML vs. NGRAM) improved overall accuracy and macro-F_1 , as well as the performance for POS and NEU classes, while leading to a slight degradation in performance for the NEG class.

The effect of dependency features (DEP) and the rule-based method output (RULE) on classification performance was positive, with increases in overall measures, as well as F_1 scores for all sentiment classes. In particular, we matched the best SVM accuracy (0.878) and obtained the best SVM POS class performance (0.744 F_1) with the addition of these features.

The addition of part-of-speech information helped mainly with the NEG class performance and with macro-F_1, the main performance criterion in this study. Using n-gram-POS pairs (NGRAML_POS) and POS sequences (POSGRAM) as additional features, we obtained the best NEG class performance (0.556 F_1 score), as well as the best macro-F_1 score (0.741), which represents a 3.1% improvement over the score reported in Xu et al. [13].

With macro-F_1 as the main evaluation criterion, we used the best feature combination (NGRAML_POS + POSGRAM + SENT + STRUCT + DEP + RULE) as the basis of experiments that incorporated "silver" citation context and subject matter annotations as well as those that considered undersampling. The results of these experiments are presented in Table 5.

Incorporating citation context and subject matter degraded the results across the board. A similar overall degradation in performance was also observed with undersampling, although we observed some improvements for polarized classes (POS and NEG). For a balanced dataset (1:1 ratio), we obtained the best recall and F_1 score for the POS class. On the other hand, the Easy Ensemble method led to the best precision for this class as well as best recall for the NEG class. However, the performance for the majority class (NEU) suffered in all cases.

4.2. NN experiments

The results of NN experiments are provided in Table 6. The performance of the baseline BiLSTM model is somewhat higher than that of the baseline CNN model (0.668 macro-F_1 vs. 0.635), but neither baseline model performs any better than baseline SVM model that use simple n-gram features (macro-F_1 of 0.665– full data not shown). However, with additional features and embeddings, the CNN model outperforms the corresponding BiLSTM model (0.757 macro-F_1 vs. 0.732). We did not observe any improvement due to simply substituting word vectors in the baseline model with POS and dependency-enhanced word vectors.

Augmenting neural network models with additional high-level features improved results across the board. Simply adding these features to baseline models increased the macro-F_1 score by more than 13% and 10% for CNN and BiLSTM, respectively. Interestingly, while using POS and dependency-enhanced word vectors with the baseline models degraded overall results slightly, when both these vectors were used as input in conjunction with high-level features, they yielded the best overall results (a 5.6% improvement in macro-F_1 and an increase of 3% in accuracy over replication of Xu et al. [13]). There was also improvement when either of these vectors was used in conjunction with high-level features (full data not shown), though their combination performed best. We note that this additional performance improvement due to POS and dependency-enhanced word embeddings was only observed for CNN (macro-F_1 of 0.757 vs. 0.719), and not for BiLSTM (macro-F_1 of 0.732 vs. 0.731).

4.3. Comparison of Three Approaches

To provide a fair comparison to the rule-based method, we also used a 80-20 training/test document split for classification. In Table 7, we report the results on the test split for the rule-based method and the SVM model that uses the features in Xu et al. [13], as well as the best-performing SVM and NN models.

The performance of the rule-based method did not reach that of the supervised learning methods. It particularly performed poorly for the NEG class. We also observed that both Xu et al.'s SVM model and the best SVM model incur a significant performance hit for the NEG class (38% and 31% degradation in F_1 score, respectively) on the held-out test set. On the other hand, the performances for other classes were reasonably close when we compare the rule-based method with the SVM models and the SVM models to one another. Finally, our

best SVM model performs better than that of Xu et al. [13], as with 10-fold cross-validation, though the improvement observed on the held-out set is slightly larger (3.6% in macro-F_1).

The NN model, on the other hand, seems to generalize better than other approaches. The performance drop for the NEG class in this case is about 9% (macro-F_1 of 0.552 vs. 0.497) on the held-out set. Overall macro-F_1 drop is 4.8%, compared to 11.7% and 12.2% for the Xu et al. model and the best SVM model, respectively. Lastly, the accuracy degrades by 1.6% for the best NN model, compared with 3.2% for the best SVM model and 2.7% for the Xu et al. model.

5. Discussion

5.1. SVM models

We were able to improve upon earlier work by using additional features based on part-of-speech tags, dependency relations, and rules. These features as well as other high level sentiment features proposed in earlier work also proved useful within a CNN architecture, leading to our best results. In all experiments, the performance on the NEG class lagged behind that on pos and NEU classes, confirming the earlier findings that criticism toward or disagreement with a reference paper is often not very clear-cut, while praise or confirmation of the reference paper is expressed more explicitly [5, 13, 26].

The experiment that considered “silver” citation context annotation yielded poorer performance, consistent with the findings in Athar and Teufel [26], who explain the degradation by data sparsity with larger citation contexts. Using subject matter annotations further degraded performance, indicating that successful supervised models for this task learn to classify using features from the topical language surrounding the citation, possibly leading to overfitting. While undersampling improved results for the minority classes, this came at the expense of some overall degradation. Thus, it seems safe to conclude that undersampling should be considered only in cases in which recall of positive and negative sentiment is the most critical concern.

5.2. NN models

The majority of studies comparing NN architectures and traditional supervised learning algorithms for text classification find that NN architectures outperform such algorithms in a baseline configuration [43]; however, our results contrast with this general finding. This may be due to several reasons. First, our training data may not be large enough to benefit from NN architectures. Secondly, the n-gram-based SVM model may be overfitting the data, as suggested above. Finally, citation sentiment analysis task may be a more challenging problem, compared to the text classification tasks considered in earlier studies.

In NN experiments, using part-of-speech and dependency-enhanced word vectors did not yield any improvement over simple word vectors. This may be partly due to the fact that we trained POS vectors on our relatively small corpus and we used dependency-enhanced word vectors pre-trained on a general English corpus (Wikipedia). In contrast, word vectors were pre-trained on a very large set of biomedical publications, and thus, may better capture the regularities of the corpus under investigation. The BiLSTM model outperformed the CNN

model in the baseline case; however, with the additional features and embeddings, the situation was reversed. This suggests that, as part of its training process, the BiLSTM model may already be learning information contributed by these features and embeddings and, therefore, may not require explicit representations of such information. However, it seems clear that high-level features are overall beneficial for citation sentiment classification. In summary, a comparison of CNN and BiLSTM architectures shows a slight edge to CNN models in citation sentiment classification. When used simply as a drop-in for a baseline supervised learning model, neither performs better than that model for this task. They both benefit from more sophisticated input and architectural design choices, though not in the same way, lending support to the view that CNNs and RNNs provide complementary information for text classification tasks [47]. One key advantage of the CNN model over BiLSTM (or any RNN model) is that it is significantly faster to train (in our case, about 3 times faster).

5.3. Comparison of Three Approaches

The rule-based method, relying simply on 148 phrases, did not perform as well as other methods, although the effect of its use as one of the features in SVM and NN models seems largely positive. It particularly performed poorly for the NEG class, pointing to the difficulty of relying solely on rhetorical cues for classifying this class, which, as mentioned earlier, is often expressed more implicitly. SVM models also incurred significant performance hits for the NEG class, suggesting that features learned for this class are based on topical language and may not generalize well. The performance hit for this class with the NN model was smaller, which may be partly attributed to using pre-trained vectors, which are less likely to overfit than corpus-based n-grams alone. The performances for other classes are reasonably close for the rule-based method, indicating that features and phrase lists for the pos class are more robust.

5.4. Limitations

Our work has several limitations. First, in NN experiments, word embeddings may not be directly comparable to POS or dependency-enhanced word embeddings, since they are generated from different corpora (though similarly modified on the training data). Our preliminary experiments showed that word vectors pre-trained on biomedical corpora performed slightly better than those pre-trained on general English corpora for our task (data not shown), which suggests that POS and dependency-enhanced word vectors trained on much larger biomedical corpora could have improved performance even further. Rather than attempting syntactic processing of large-scale corpora, we took a simpler route by training POS vectors on a small corpus and using dependency-enhanced word vectors pre-trained on Wikipedia. Future work should isolate the effect of different kinds of input vectors by using the same corpora in their training.

Another limitation is that we did not experiment with more complex NN architectures involving more hidden layers, since, in this study, we were primarily interested in comparing distinct approaches, rather than variants of the same architecture. One such model could outperform our best NN model. In addition, while we experimented with different model configuration parameters (filter region size, mini-batch size, etc.), we did not perform an

exhaustive search for the optimal parameters and mostly followed best practices for NN training [43].

Finally, the results of 10-fold cross-validation experiments should be treated as indicators of general trends regarding the effect of features, citation context/subject matter and undersampling, and relative performance of SVM models and NN models, rather than indicators of absolute performance. This is because we did not exclude sentiment-related and other relevant phrases from feature extraction based on their absence in the training folds in cross-validation. This was to enable comparison to Xu et al. [13] and to avoid overpenalizing our models. Our results on the held-out test set (Table 7) address this issue, since, for those experiments, we trained using phrases that appeared in the training documents (80%) only. Therefore, we consider the results on the held-out test as the primary results of our work.

5.5. Error analysis

We analyzed the instances in which all three methods (rule-based method as well as the best SVM and NN models) failed to recognize the correct sentiment class. In the majority of these cases, a citation was classified as neutral by all methods, whereas it expressed positive or negative sentiment (36.5% and 49%, respectively).

Inability of the methods to identify the precise citation context often led to classification errors. In the example below, all methods classified the citation as neutral, probably because they failed to consider the following sentence, which expresses the positive sentiment, as part of the context.

1. Merlin et al. [22] also reported that the route of naloxone administration (IN or IV) made no difference to the effect on respiratory rate. Our findings supported this conclusion.

Failure to recognize a citation context shorter than the full citation sentence may lead to errors, too, as illustrated in the sentence below. The citation in this sentence was classified as neutral by the CNN model and negative by the other two, possibly because of the subordinate clause *In contrast to the negative findings of these two studies*, which in fact refers to the citations discussed in the previous sentence.

2. In contrast to the negative findings of these two studies, Cals and co-workers showed that the use of CRP testing significantly reduced, antibiotie prescribing for LRTI without decreasing the quality of care and the outcome of treatment [25].

Another significant source of errors seems to be comparative statements, which can indicate a discrepancy between the findings of the citing paper and those of the reference paper. In such cases, sentiment is often negative and the prediction neutral. On the surface, the statements seem objective, which may explain the challenge they present. An example is provided below.

3. We found a greater fluoride concentration 12 hours following the use of both mouthwashes compared to investigations performed by Pessan et al. [8] , Vogel et al. [10] and Whitford et al. [9].

Some other errors seem to be due to inadequacy of the sentiment phrase lists used in the study or the inability of the supervised learning methods to learn such phrases from the corpus. In the example below, the matrix clause *It is known that* and the subordinate clause *which may be the case in this cohort* seem to contribute to the positive sentiment, but it was not in the rule-based method phrase list and was probably not learned as a predictive feature by the other methods.

4. It is known that chronically inflamed eyes are at higher risk of graft failure [27] , which may be the case in this cohort.

All the examples above provide support to the view that sentiment in the context of citations is expressed more subtly than that in other contexts, such as movie reviews.

In the only case (shown below) where all three methods classified the citation with the opposite polarity of the actual sentiment (negative instead of positive), we were unable to determine the rationale for the annotation.

5. Consistently, in the study of Basu et al., daily use of cranberry juice for 8 weeks caused a significant decrease in both ox-LDL and MDA versus placebo treatment (−33% versus −17% and −50% versus +7%, resp.) 30]; however, in this study, the baseline MDA values were higher than those of our subjects and the duration of intervention was longer.

Since the inter-annotator agreement is moderate for the corpus, it is conceivable that this citation was incorrectly annotated, particularly when we consider that most cases involving comparative statements like this one had negative polarity (as noted above).

6. Conclusions

We presented and compared several methods for citation sentiment analysis in discussion sections of clinical research publications. A convolutional neural network model that uses part-of-speech and dependency-enhanced word vectors as input and incorporates additional high-level features yielded the best performance and seemed to generalize better than traditional machine learning (SVM) and rule-based methods. However, training a NN model is orders of magnitude slower than training traditional learning algorithms (including SVM), and whether the difference in performance justifies the additional cost may be debatable. In our case, the difference in performance in terms of macro-F_1 on unseen data was about 11%; however, hand-crafted features incorporated into the CNN model accounted for some of this improvement. Considering also the fact that baseline NN models did not perform better than the baseline SVM model, it seems reasonable to conclude that the features automatically learned through NN training were less useful than the hand-crafted sentiment lexicon, rhetorical structure, and rule-based features considered for this task.

NN architectures are known to benefit from large amounts of annotated data, and the relatively small size of the corpus is probably the biggest reason behind the somewhat

modest results we obtained with the baseline NN models. We explored leveraging the predictions of the rule-based method on unlabeled data as a kind of semi-supervised learning, but this did not yield any noticeable improvement for NN models (data not shown). However, more sophisticated methods to leverage unlabeled data or distant supervision based on external citation databases may be promising avenues to explore.

Our intuition that providing more complete information about the context of the citation and isolating its topical content would lead to more informative, less noisy features was not borne out by our experiments. However, since our annotation of context/subject matter was exploratory and our method of using these annotations for learning was relatively simple, we believe more work is needed to understand their role in sentiment classification of citations in biomedical publications.

Other future work will focus on more fine-grained citation analysis, particularly citation function. Challenges in this task include identifying a set of citation function classes appropriate for biomedical publications and the development of a manually annotated corpus based on these categories. The citation sentiment classification models developed in this work are likely to be useful as assistance for manual annotation and they can also inform approaches developed for recognizing citation function categories.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

HK, ZP, and GR were supported by the intramural research program at the U.S. National Library of Medicine, National Institutes of Health. ST and TT took part in this study during their participation in the Lister Hill National Center for Biomedical Communications (LHNCBC) Research Program in Medical Informatics for Graduate students at the U.S. National Library of Medicine. JS was supported by an appointment to the NLM Research Participation Program, administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the National Library of Medicine Research Participation Program.

References

[1]. Garfield E, Citation analysis as a tool in journal evaluation, Science 178 (4060) (1972) 471–479. [PubMed: 5079701]

[2]. Egghe L, Theory and practise of the g-index, Scientometrics 69 (1) (2006) 131–152.

[3]. Hirsch JE, An index to quantify an individual's scientific research output, Proceedings of the National Academy of Sciences of the United States of America 102 (46) (2005) 16569–16572. doi:10.1073/pnas.0507655102. [PubMed: 16275915]

[4]. Waltman L, A review of the literature on citation impact indicators, Journal of Informetrics 10 (2) (2016) 365–391.

[5]. Teufel S, The Structure of Scientific Articles: Applications to Citation Indexing and Summarization, Center for the Study of Language and Information (CSLI), 2010.

[6]. Seglen PO, Why the impact factor of journals should not be used for evaluating research, BMJ 314 (7079) (1997) 497. doi:10.1136/bmj.314.7079.497.

[7]. Hutchins BI, Yuan X, Anderson JM, Santangelo GM, Relative Citation Ratio (RCR): A New Metric That Uses Citation Rates to Measure Influence at the Article Level, PLOS Biology 14 (9) (2016) 1–25. doi:10.1371/journal.pbio.1002541.

[8]. Swales J, Citation analysis and discourse analysis, Applied linguistics 7 (1) (1986) 39–56.

[9]. Zhang G, Ding Y, Milojević S, Citation content analysis (CCA): A framework for syntactic and semantic analysis of citation content, Journal of the Association for Information Science and Technology 64 (7) (2013) 1490–1503.

[10]. Athar A, Sentiment analysis of scientific citations, Tech. Rep. UCAM-CL-TR-856, University of Cambridge, Computer Laboratory (2014).

[11]. Zhu X, Turney PD, Lemire D, Vellino A, Measuring academic influence: Not all citations are equal, CoRR abs/1501.06587.

[12]. Athar A, Sentiment analysis of citations using sentence structure-based features, in: Proceedings of the ACL 2011 Student Session, 2011, pp. 81–87.

[13]. Xu J, Zhang Y, Wu Y, Wang J, Dong X, Xu H, Citation sentiment analysis in clinical trial papers, in: AMIA Annual Symposium Proceedings, 2015, pp. 1334–1341.

[14]. Moravcsik MJ, Murugesan P, Some results on the function and quality of citations, Social studies of science 5 (1) (1975) 86–92.

[15]. Spiegel-Rosing I, Science studies: Bibliometric and content analysis, Social Studies of Science 7 (1) (1977) 97–113.

[16]. Teufel S, Siddharthan A, Tidhar D, An annotation scheme for citation function, in: Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue, SigDIAL ‘06, 2006, pp. 80–87.

[17]. Agarwal S, Choubey L, Yu H, Automatically classifying the role of citations in biomedical articles, in: AMIA Annual Symposium proceedings, Vol. 2010, 2010, pp. 11–15.

[18]. Hernández-Alvarez M, Soriano JMG, Martínez-Barco P, Citation function, polarity and influence classification, Natural Language Engineering 23 (4) (2017) 561–588.

[19]. Teufel S, Siddharthan A, Tidhar D, Automatic classification of citation function, in: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, EMNLP ’06, 2006, pp. 103–110.

[20]. Abu-Jbara A, Radev D, Reference scope identification in citing sentences, in: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT ’12, 2012, pp. 80–90.

[21]. Jia M, Citation function and polarity classification in biomedical papers, Master’s thesis, University of Western Ontario, Canada (2018).

[22]. Greenberg SA, How citation distortions create unfounded authority: analysis of a citation network, BMJ 339 (2009) b2680. doi:10.1136/bmj.b2680. [PubMed: 19622839]

[23]. Yu B, Automated citation sentiment analysis: What can we learn from biomedical researchers, Proceedings of the American Society for Information Science and Technology 50 (1) (2013) 1–9. doi:10.1002/meet.14505001084.

[24]. Hernández-Alvarez M, Gómez JM, Survey about citation context analysis: Tasks, techniques, and resources, Natural Language Engineering 22 (3) (2016) 327–349.

[25]. Radev DR, Muthukrishnan P, Qazvinian V, The ACL Anthology Network Corpus, in: Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries, NLPIR4DL ’09, 2009, pp. 54–61.

[26]. Athar A, Teufel S, Context-enhanced citation sentiment detection, in: Proceedings of the 2012 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies, Association for Computational Linguistics, 2012, pp. 597–601.

[27]. Qazvinian V, Radev DR, Identifying non-explicit citing sentences for citation-based summarization, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010, pp. 555–564.

[28]. Abu-Jbara A, Ezra J, Radev D, Purpose and polarity of citation: Towards NLP-based bibliometrics, in: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2013, pp. 596–606.

[29]. Ó Séaghdha D, Teufel S, Unsupervised learning of rhetorical structure with un-topic models, in: Proceedings of the 25th International Conference on Computational Linguistics (COLING-14), Dublin, Ireland, 2014, pp. 2–13.

[30]. Fan R-E, Chang K-W, Hsieh C-J, Wang X-R, Lin C-J, LIBLINEAR: A library for large linear classification, Journal of Machine Learning Research 9 (2008) 1871–1874.

[31]. Manning CD, Surdeanu M, Bauer J, Finkel J, Bethard SJ, Mc-Closky D, The Stanford CoreNLP natural language processing toolkit, in: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2014, pp. 55–60.

[32]. Liu X-Y, Wu J, Zhou Z-H, Exploratory undersampling for class-imbalance learning, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 39 (2) (2009) 539–550.

[33]. Goldberg Y, Neural network methods for natural language processing, Synthesis Lectures on Human Language Technologies 10 (1) (2017) 1–309.

[34]. LeCun Y, Bottou L, Bengio Y, Haffner P, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (11) (1998) 2278–2324.

[35]. Hochreiter S, Schmidhuber J, Long short-term memory, Neural computation 9 (8) (1997) 1735–1780. [PubMed: 9377276]

[36]. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J, Distributed representations of words and phrases and their compositionality, in: Advances in neural information processing systems, 2013, pp. 3111–3119.

[37]. Levy O, Goldberg Y, Dependency-based word embeddings, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Vol. 2, 2014, pp. 302–308.

[38]. Mikolov T, Chen K, Corrado G, Dean J, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781.

[39]. Chiu B, Crichton G, Korhonen A, Pyysalo S, How to train good word embeddings for biomedical nlp, in: Proceedings of the 15th Workshop on Biomedical Natural Language Processing, 2016, pp. 166–174.

[40]. Krizhevsky A, Sutskever I, Hinton GE, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.

[41]. Kim Y, Convolutional neural networks for sentence classification, arXiv preprint arXiv:1408.5882.

[42]. Kalchbrenner N, Grefenstette E, Blunsom P, A convolutional neural network for modelling sentences, arXiv preprint arXiv:1404.2188.

[43]. Zhang Y, Wallace B, A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification, arXiv preprint arXiv:1510.03820.

[44]. Elman JL, Finding structure in time, Cognitive science 14 (2) (1990) 179–211.

[45]. Tang D, Qin B, Liu T, Document modeling with gated recurrent neural network for sentiment classification, in: Proceedings of the 2015 conference on empirical methods in natural language processing, 2015, pp. 1422–1432.

[46]. Kavuluru R, Rios A, Tran T, Extracting drug-drug interactions with word and character-level recurrent neural networks, in: 2017 IEEE International Conference on Healthcare Informatics (ICHI), IEEE, 2017, pp. 5–12.

[47]. Yin W, Kann K, Yu M, Schütze H, Comparative Study of CNN and RNN for Natural Language Processing, CoRR abs/1702.01923. arXiv:1702.01923.

Highlights:

- We present a detailed comparison of three approaches to citation sentiment analysis in clinical trial publications: a rule-based method, SVM and two neural network variants: a convolutional neural network model and a bidirectional long short-term memory model.
- Convolution neural network model augmented with hand-crafted features yields the best performance both in 10-fold cross-validation and generalizes better to unseen data.
- Hand-crafted phrase-based features and high-level contextual features are more predictive for this task than n-gram or embedding-based features.

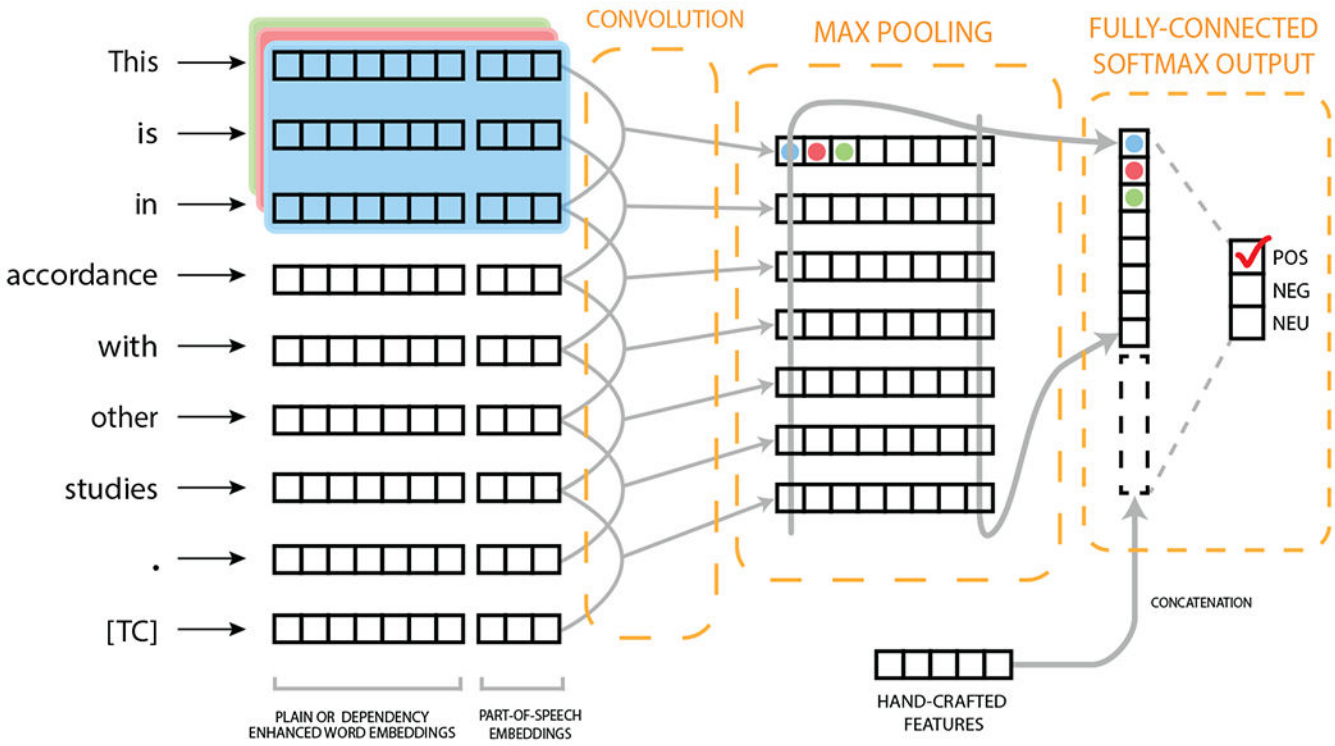


Figure 1:

A high-level overview of the CNN architecture. The input is the sentence *This is in accordance with other studies*. [9].

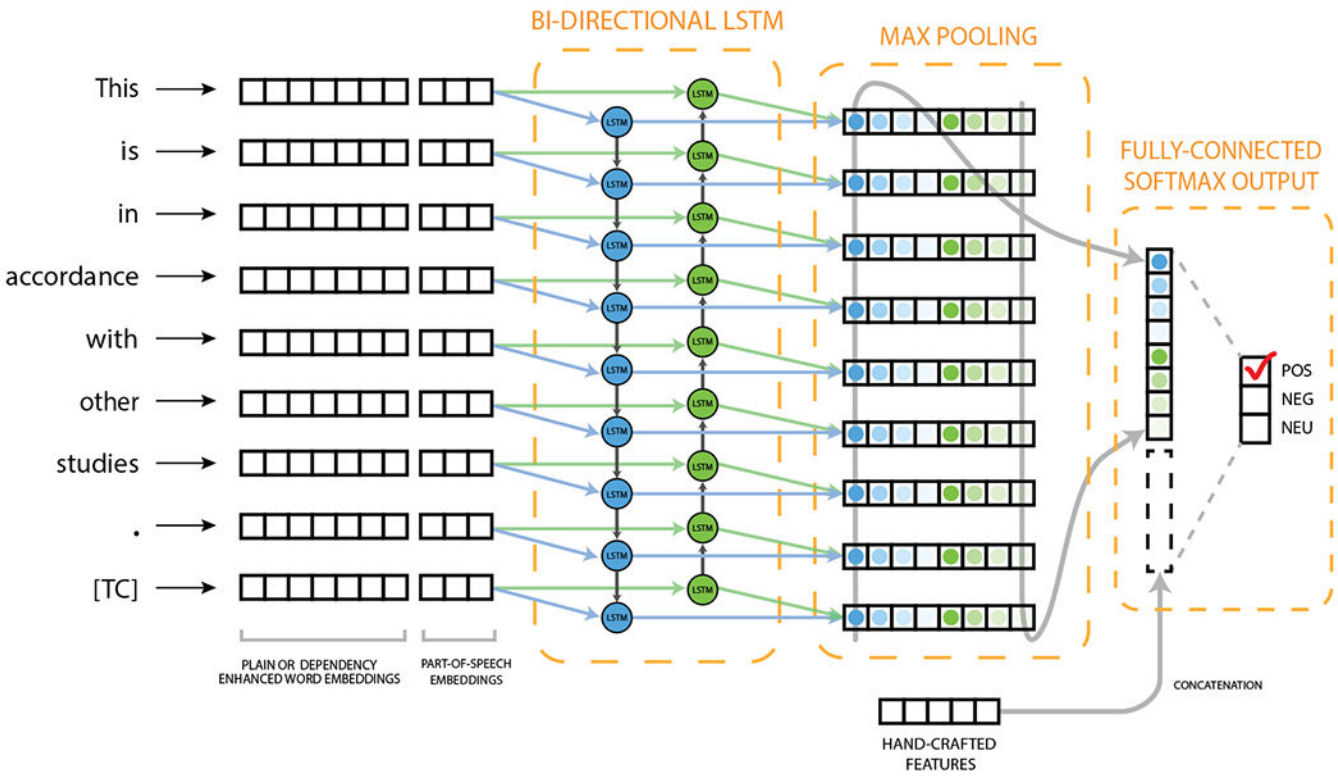


Figure 2:

A high-level overview of the BiLSTM architecture. The input is the sentence *This is in accordance with other studies*. [9].

Table 1:

Basic corpus statistics

Documents	285
Citations	4,182
Positive (POS) citations	702 (16.8%)
Negative (NEG) citations	308 (7.4%)
Neutral (NEU) citations	3,172 (75.8%)

Table 2:

Citation sentiment examples from the annotated corpus

Class	Citation	Context
POS	*[17]*	*Consistent with other reports of behavioral treatments [17]. adherence to yoga and exercise interventions in this clinical trial was significantly correlated with baseline variables including depression, fatigue, and physical aspects of quality of life.*
NEG	*[5]*	*We found that intensive glucose control did not reduce the morbidity or the mortality of patients admitted to a mixed medical/surgical ICU with medical problems, non-cardiovascular surgeries or trauma. These results differ from two previous studies. The first one with patients in a cardiovascular-surgical ICU [5] demonstrated a decrease in morbidity and mortality.*
NEU	*[27]*	*Deterioration of glycemic regulation often first appears as compensatory hyperinsulinemia seen post-prandial to a meal to maintain glucose uptake into cells and normal homeostasis [27].*

Table 3:

Illustration of features on the sentence *Our data are generally consistent with that of other studies [4–7], but not with studies where a single dose of paracetamol was administered, which reported no significant impact on any reactions [8,9].* The citation under consideration for this example is *[4–7]* and the citation context is the current sentence only.

Feature Type	Features
n-gram features	
n-grams (NGRAM)	[OC], [TC], administered, which_reported, are_generally_consistent, …
Lemma n-grams (NGRAML)	[OC], [TC], administer, which_report, be_generally_consistent, …
POS n-grams(POSGRAM)	CC, DT, NN, CC_RB, DT_IN, CC_RB_IN, …
POS-augmented lemma n-grams (NGRAML_POS)	[OC], [TC], administer_VB, which_WD__report_VB, be_VB__generally_RB__consistent_JJ …
Negation count	2 (due to *no, not*)
Negated unigrams	NOT_impact, NOT_with, NOT_study, NOT_significant
Sentiment lexicon features (SENT)	
POS sentiment	TRUE (due to *consistent with*)
NEG sentiment	FALSE
Any sentiment	TRUE
Structure features (STRUCT)	
n-grams	[OC], [TC], CITINGWORK, POS, CONTRAST, CITINGWORK_POS, CITINGWORK_POS_[TC], …
Direction	CITINGWORK_CONTRAST, [TC]_CONTRAST, …
Additional features	
Dependency features (DEP)	advcl_consistent_administer, neg_consistent_not, nsubj_consistent_datum, …
Rule-based output (RULE)	POS

Table 4:

The results of SVM experiments with various feature combinations. The best performance for individual metrics are underlined. The evaluation is based on 10-fold cross-validation.

Experiment	Overall		Per Category			
	Accu.	MacroF_1	Cat	Pr.	Rec.	F_1
Results from Xu et al. [13]	0.870	0.719	POS	0.823	0.644	0.723
			NEG	0.711	0.399	0.511
			NEU	0.886	0.966	0.924
Reproducing Xu et al. (NGRAM+SENT+STRUCT)	0.870	0.717	POS	0.835	0.618	0.710
			NEG	0.717	0.403	0.516
			NEU	0.883	0.971	0.925
NGRAML+SENT+STRUCT	0.873	0.720	POS	0.854	0.631	0.726
			NEG	0.709	0.396	0.508
			NEU	0.884	0.973	0.927
NGRAML+SENT+STRUCT+DEP+RULE	0.878	0.735	POS	0.846	0.664	0.744
			NEG	0.733	0.419	0.533
			NEU	0.891	0.970	0.929
NGRAML+SENT+STRUCT+POSGRAM+RULE	0.876	0.738	POS	0.808	0.672	0.734
			NEG	0.739	0.442	0.553
			NEU	0.895	0.963	0.927
NGRAML_POS+POSGRAM+SENT+STRUCT+DEP+RULE	0.877	0.741	POS	0.828	0.667	0.739
			NEG	0.751	0.442	0.556
			NEU	0.891	0.966	0.927
NGRAML_POS+SENT+STRUCT+RULE	0.878	0.737	POS	0.836	0.660	0.737
			NEG	0.724	0.435	0.544
			NEU	0.893	0.969	0.930

Table 5:

Results of experiments with “silver” citation context/subject matter annotations and undersampling. All experiments use the best feature combination (NGRAML_POS + POSGRAM + SENT + STRUCT + DEP + RULE). The best results are underlined. The evaluation is based on 10-fold cross-validation.

Experiment	Overall		Per Category			
	Accu.	MacroF$_1$	Cat	Pr.	Rec.	F$_1$
Base case	0.877	0.741	POS	0.828	0.667	0.739
			NEG	0.751	0.442	0.556
			NEU	0.891	0.966	0.927
“Silver” citation context	0.868	0.718	POS	0.818	0.638	0.717
			NEG	0.709	0.403	0.513
			NEU	0.885	0.965	0.923
“Silver” context with normalized subject matter	0.849	0.690	POS	0.736	0.642	0.686
			NEG	0.547	0.412	0.470
			NEU	0.891	0.937	0.913
1:1 ratio (POS +NEG=NEU)	0.773	0.726	POS	0.801	0.738	0.768
			NEG	0.687	0.513	0.587
			NEU	0.774	0.876	0.822
1:2 ratio	0.824	0.709	POS	0.808	0.641	0.715
			NEG	0.675	0.425	0.522
			NEU	0.840	0.948	0.891
Easy Ensemble	0.839	0.701	POS	0.903	0.517	0.658
			NEG	0.438	0.695	0.537
			NEU	0.891	0.924	0.907

Table 6:

Results obtained with NN models, with 10-fold cross validation and model averaging. The scores that indicate an improvement over SVM models in Table 4 are underlined. WORD, POS, and DEP indicate word, part-of-speech and dependency-enhanced word embeddings, respectively. FEAT indicates additional high-level features (POSGRAM + SENT + STRUCT + RULE).

Experiment	Overall		Per Category			
	Accu.	MacroF_1	Cat	Pr.	Rec.	F_1
CNN(WORD)	0.858	0.635	POS	0.830	0.579	0.680
			NEG	0.780	0.192	0.305
			NEU	0.863	0.983	0.919
CNN(POS+DEP)	0.850	0.632	POS	0.818	0.559	0.661
			NEG	0.674	0.212	0.318
			NEU	0.861	0.975	0.914
CNN(WORD)+FEAT	0.887	0.719	POS	0.919	0.655	0.764
			NEG	0.957	0.310	0.464
			NEU	0.882	0.994	0.935
CNN(POS+DEP)+FEAT	0.896	0.757	POS	0.914	0.689	0.783
			NEG	0.898	0.404	0.552
			NEU	0.896	0.990	0.940
BiLSTM(WORD)	0.857	0.668	POS	0.785	0.651	0.711
			NEG	0.551	0.291	0.375
			NEU	0.882	0.958	0.918
BiLSTM(POS+DEP)	0.837	0.644	POS	0.747	0.610	0.670
			NEG	0.456	0.303	0.356
			NEU	0.876	0.938	0.906
BiLSTM(WORD)+FEAT	0.883	0.731	POS	0.845	0.695	0.761
			NEG	0.827	0.378	0.509
			NEU	0.892	0.973	0.931
BiLSTM(POS+DEP)+FEAT	0.874	0.732	POS	0.798	0.708	0.747
			NEG	0.706	0.426	0.523
			NEU	0.900	0.952	0.925

Table 7:

Results of experiments on the held-out test set. The best performing SVM model uses the features NGRAML_POS + POSGRAM + SENT + STRUCT + DEP + RULE. Best NN model is a CNN which uses POS and dependency-enhanced word embeddings as input and incorporates high-level features. Best results are underlined.

Experiment	Overall		Per Category			
	Accuracy	MacroF$_1$	Cat	Pr.	Rec.	F$_1$
Rule-based method	0.831	0.604	POS	0.780	0.568	0.657
			NEG	0.389	0.187	0.252
			NEU	0.859	0.951	0.903
Xu et al. SVM model	0.847	0.633	POS	0.783	0.576	0.664
			NEG	0.548	0.227	0.321
			NEU	0.868	0.966	0.914
Best SVM model	0.850	0.656	POS	0.740	0.616	0.672
			NEG	0.600	0.280	0.382
			NEU	0.877	0.956	0.915
Best NN model	0.882	0.721	POS	0.783	0.681	0.728
			NEG	0.930	0.341	0.497
			NEU	0.895	0.982	0.937