ClaimDistiller: Scientific Claim Extraction with Supervised Contrastive Learning

Xin Wei*
Md Reshad Ul Hoque*
xwei001@odu.edu
mhoqu001@odu.edu
Old Dominion University
Norfolk, Virginia, USA

Jian Wu
Jiang Li
j1wu@odu.edu
jli@odu.edu
Old Dominion University
Norfolk, Virginia, USA

ABSTRACT

The growth of scientific papers in the past decades calls for effective claim extraction tools to automatically and accurately locate key claims from unstructured text. Such claims will benefit content-wise aggregated exploration of scientific knowledge beyond the metadata level. One challenge of building such a model is how to effectively use limited labeled training data. In this paper, we compared transfer learning and contrastive learning frameworks in terms of performance, time and training data size. We found contrastive learning has better performance at a lower cost of data across all models. Our contrastive-learning-based model ClaimDistiller has the highest performance, boosting the F1 score of the base models by 3–4%, and achieved an F1=87.45%, improving the state-of-the-art by more than 7% on the same benchmark data previously used for this task. The same phenomenon is observed on another benchmark dataset, and ClaimDistiller consistently has the best performance. Qualitative assessment on a small sample of out-of-domain data indicates that the model generalizes well. Our source codes and datasets can be found here: https://github.com/lamps-lab/sci-claim-distiller.

CCS CONCEPTS

• **Computing methodologies** → **Information extraction**.

KEYWORDS

Scientific Claim Extraction, Word Embedding, Deep Learning, Transfer Learning, Contrastive learning

ACM Reference Format:
Xin Wei, Md Reshad Ul Hoque, Jian Wu, and Jiang Li. 2023. ClaimDistiller: Scientific Claim Extraction with Supervised Contrastive Learning. In *Proceedings of ACM/IEEE JOINT CONFERENCE ON DIGITAL LIBRARIES (JCDL '23).* ACM, New York, NY, USA, 11 pages. https://doi.org/XXXXXXX.XXXXXXX

*Both authors contributed equally to this research.

JCDL '23, June 26 - 30, 2023, Santa Fe, New Mexico, USA

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Because of the rapid increase of scientific papers indexed by digital libraries [1] [2], there is an emergent need to help readers to efficiently grasp the main ideas of research papers. This can be achieved by development of algorithms to extract and aggregate key information from unstructured scholarly text. Existing machine learning methods have been developed to extract metadata, such as title, authors, year, venue, e.g., [3], non-textual content such as figures and tables, e.g., [4], and high-level semantic information such as keywords, e.g., [5]. However, scientific claims, conveying key findings and contributions from unstructured text remains challenging because scientific ideas could be conveyed in a more complicated way than general text as used in news papers and Wikipedia articles. Although deep learning has shown promising results for open domain extractive summarization and key sentences identification, e.g., [6, 7], it is still challenging to train robust deep learning models on scientific papers [8] because of the lack of large-scale training data. Obtaining such training data usually requires domain knowledge, which regular crowdsourcing workers may not possess. Identifying key claims from scientific papers can also be time-consuming for domain experts. In addition, mining claims from scientific papers has shown to be an important step to automatically assessing reproducibility in social and behavioral sciences and other domains, e.g., [9, 10], which is investigated in DARPA's Systematizing Confidence in Open Research and Evidence (SCORE) program [11].

We define a *scientific claim* as a sentence that provides the core findings of a scientific paper. One example is given in Figure 1. Existing datasets with annotated claims are scarce and not available in all domains. Current datasets on claim extraction include CoreSC dataset [12] with 265 articles in physical chemistry and biochemistry. The Dr. Inventor dataset [13] contains claims extracted from 40 computer graphics articles. Another dataset used in a recent paper [14] contains claims extracted from 1,500 scientific abstracts in the biomedical domain. Due to data scarcity, it is important to develop models that efficiently use existing data. In a recent paper [14] the authors introduced transfer learning to perform scientific claim extraction. In this paper, we explore alternative ways for this task.

Transfer learning uses the knowledge extracted from one or more *source tasks*, which usually have a high amount of resources, to accomplish a *target task*, which usually has a lower amount of resources. Transfer learning works by pretraining a neural model using data for the source tasks. The model is retrained by freezing the weights of a portion of a neural network and learning the

weights of the other portion of the same neural network [15]. Transfer learning has been adopted in computer vision (CV) and natural language processing (NLP) tasks, e.g., [16] [17].

Transfer learning relaxes the i.i.d. (independent and identically distributed) requirement for training and testing datasets. To be specific, the classes in source data does not necessarily need to be the same with target data. This is usually fullfilled by the extremely large sizes of source datasets, such as ImageNet-21k dataset with 14.2 million images [18]. Source data used in NLP (Natural Language Processing) is usually in the magnitude of tens of Mega bites and even more. Data size is a limit for claim extraction and as a result transfer learning does not delivery enough power. In this paper, we introduce contrastive learning framework which uses significantly less training data and achieves comparable or better performance.

Self-supervised contrastive learning, a type of self-supervised representation learning, efficiently leverages limited training data and has demonstrated promising results in multiple CV and NLP tasks, e.g., [19, 20]. This method puts similar samples close to each other while pushing 'negative' samples far apart in the feature space [21]. For example, in image classification, data can be augmented by cutting and rotation. We can adjust the loss function and make the augmented samples from the same image close to each other and augmented samples from different images far away. In this way, the model can learn the features without looking at labels. The drawback of self-supervised contrastive learning is that the correlation of features between images belonging to the same class is ignored. This could be mitigated by leveraging label information, which is the *supervised contrastive learning* [22].

In this paper, we compared transfer learning and supervised contrastive learning frameworks in terms of performance, time and training data size. We found contrastive learning has better performance at a lower cost of data across all models on both datasets. We propose a contrastive-learning-based model ClaimDistiller, the backbone of which is a recurrent neural model with supervised contrastive learning. We demonstrate that the supervised contrastive learning mechanism improves the model performance by a significant margin with less training samples and training time.

Our best model achieves F1=87.45% when trained and tested on **SciCE**. We further trained the model on another benchmark dataset **SciARK**, and contrastive learning methods obtained better performance across all models than transfer learning. ClaimDistiller consistently outperforms all other models.

The contributions of the paper are as follows:

(1) We proposed using supervised contrastive learning for scientific claim extraction. The results show that SCL achieves a comparable or better performance than transfer learning with significantly less training data and training time. The best model achieves an F1=87.45% on the SciCE dataset.
(2) We compared 10 commonly used methods of text augmentation for training SCL in the context of scientific claim extraction. All methods exhibit a marginal effect on the model performance.
(3) Our best model was trained and evaluated on a standard benchmark in the biomedical domain. The model exhibited reasonably well generalizability when it is tested in the computer science domain.

Title: Calpain-mediated ABCA1 degradation: post-translational regulation of ABCA1 for HDL biogenesis

Claim sentence: Pharmacological inhibition of the calpain-mediated ABCA1 degradation results in the increase of the ABCA1 activity and HDL biogenesis in vitro and in vivo, and potentially suppresses atherogenesis.

Non-claim sentence: This article is part of a Special Issue entitled Advances in High Density Lipoprotein Formation and Metabolism: A Tribute to John F. Oram (1945-2010).

Figure 1: An example of claim extraction dataset.

2 RELATED WORK

Scientific claim extraction is closely related to extractive document summarization and argumentation mining, which are more explored in literature. The goal of extractive document summarization is to extract text that is much shorter than the original documents and deliver the main idea of the given documents [23]. A survey on extractive document summarization for *scientific papers* can be found in [24]. The text output by extractive document summarization may contain several key sentences that provide a high-level description of the original text. These sentences may not necessarily describe the core findings. Therefore, the methods cannot directly be used for extracting scientific claims.

Argument mining automatically extract the structure of inference and reasoning presented in natural language text [25]. In argument mining, premises were extracted from news [26], social media [27], scientific article [28], and Wikipedia [29]. Existing argument mining methods include heuristic methods [30, 31] and classical machine learning methods [32]. Recently, deep learning methods, including weak supervision and transfer learning mechanisms, have been proposed [33].

There are limited publications on scientific claim extraction. Dernoncourt et al. [34] developed a scientific discourse dataset **PubMed-RCT**, in which sentences were labeled into five classes, namely, background, introduction, method, result, and conclusion. However, claims were not explicitly labeled in this dataset. Recently, a human-annotated scientific claim extraction dataset in biomedical domains was published [14]. Existing methods used for scientific claim extraction include rule-based and deep learning methods. Rule-based methods were used to extract claims from scientific papers in Jansen et al. [30]. Achakulvisut et al. [14] proposed a model consisting of a bidirectional long short-term memory (BiLSTM) network stacked with a conditional random field (CRF) model trained in a transfer learning framework. They trained their model on the **PubMed-RCT** dataset and then fine-tuned the model on their in-house **SciCE** dataset.

Table 1: A comparison of performances with different data augmentation methods.

Dataset	Labels	Size *	Domain	Utility ***
SciCE	Claim, Non-claim	11702	Biomedical	C-pre, C-tune, T-tune
SciARK	Claim, Evidence, Non-claim	9055	SDG **	C-pre, C-tune, T-tune
Pubmed	Objective, Introduction, Method, Result, Conclusion	2.3 million	Biomedical	T-pre

* Measured in number of sentences.
** Six SDG (Sustainable Development Goals) domains set by the United Nations (UN).
*** "C" means contrastive learning, "T" means transfer learning, "pre" means pre-training, "tune" means fine-tuning.

3 DATA

The claims to be extracted should be absolute, independent, core findings of the paper. A conclusion may not necessarily be a claim, but a claim is highly likely to be a conclusion. Claims may appear in the abstracts and the body text, but in our research task, we focus on extracting claims from abstracts, assuming that authors should put the core findings of the paper in the abstracts.

The data used in this paper includes three corpora. The first corpus was built by Achakulvisut et al. [14], which is the largest dataset so far for scientific claim extraction. For convenience, we call it the scientific claim extraction (**SciCE**) dataset.

Specifically, the dataset labels three types of claims:

Type 1: A statement that declares something is better;
Type 2: A statement that proposes something new;
Type 3: A statement that describes a new finding or a new cause-effect relationship.

The corpus contains 1,500 scientific abstracts in the biomedical domain. Each sentence in the abstracts was labeled by domain experts into two categories, namely, claim and non-claim. An example of a claim sentence and a non-claim sentence, in an abstract, is shown in Figure 1. Each abstract contains 5 to 10 sentences (Figure 2). One abstract may contain more than one claim (Figure 3). The majority of the abstracts contain 1–2 claims and about half of the dataset contains only 1 claim in an abstract. The dataset contains in total 2276 claims and 9426 non-claims. For an even comparison, we adopt the split of the original dataset in which the numbers in training, test, and validation samples are 750, 375, and 375, respectively.

The second corpus is the **Pubmed-RCT** dataset [34], designed for the discourse prediction task, which was to predict the discourse types for a sequence of sentences in one abstract. In our paper, it is used as the *source dataset* for transfer learning. **Pubmed-RCT** is a larger dataset consisting of 20,000 abstracts, including 2.3 million sentences selected from the MEDLINE/PubMed Baseline Database published in 2016. The abstracts are in biomedical and life sciences domains, and particularly in randomized controlled trials (RCTs). The discourse type for each sentence is one of the five classes, *Objective*, *Introduction*, *Method*, *Result*, and *Conclusion*. The *Method* and *Result* classes contain one-third of all labeled sentences, respectively. The remaining one-third contains sentences labeled as the other three classes. The number of sentences in an abstract is between 3 and 51, with an average of 11.6. This dataset will be used for pre-training in transfer learning.

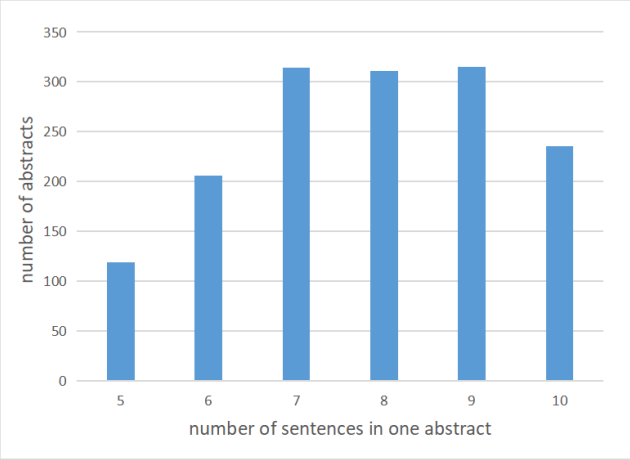

Figure 2: Distribution of the number of sentences in an abstract in the claim extraction dataset (SciCE). An abstract has at least 5 sentences, and at most 10 sentences.

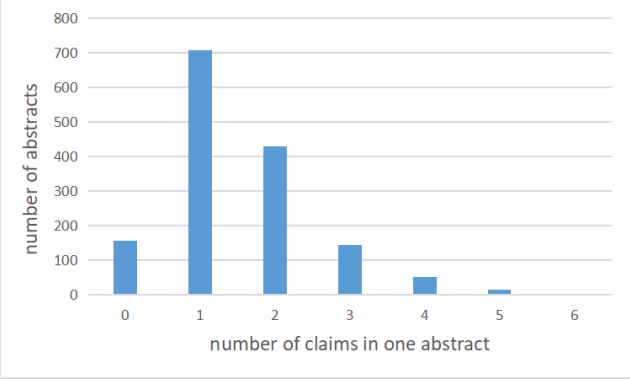

Figure 3: Distribution of the number of claims in an abstract in the claim extraction dataset (SciCE). One abstract can have multiple claims and the maximum number of claims is 6. Most frequently there are 2 claims in one abstract.

A third dataset **SciARK** was introduced in a recent work [35]. It is a relatively small dataset composed of abstracts from 689 academic papers with 9055 sentences. The number of abstracts in

training, testing, and validation samples are 350, 269, and 70, respectively, as split by the authors. Each sentence is annotated as *Claim*, *Evidence*, or *Nonetype*. Unlike SciCE and Pubmed, this dataset is multidisciplinary with abstracts of scientific publications related to a broad spectrum of Sustainable Development Goals (SDG) domains. When using the dataset, we merge the "Evidence" and "Nonetype" as "non-claim" and treat it as a binary-class dataset (claim vs. non-claim).

4 PROPOSED FRAMEWORK: CLAIMDISTILLER

We formulate the claim extraction task as a classification problem on a sequence of sentences, where the model predicts a class label claim or non-claim for each sentence. In regular classification models, text is represented in the form of vectors and training a good representation is essential for classification. We improve the models by adopting supervised contrastive learning to generate better representations. We propose a framework called ClaimDistiller for extracting scientific claims from abstracts.

4.1 Supervised Contrastive Learning

Self-supervised contrastive learning [20] methods can be used to generate representations for non-labeled data. It treats each sample in the dataset as a class and compares them pairwise after data augmentation to obtain "apparent similarities", and further generates representations for each sample. Supervised contrastive learning [22] methods introduce this framework for labeled data. The key idea is to train a representation that pulls together the same class while simultaneously pushing apart different classes in the embedding space. This step helps to create more accurate embeddings and thus subsequent classification based on it can achieve better performance than regular supervised learning.

In self-supervised contrastive learning each sample is considered a class, while in supervised contrastive learning each label is considered a class. As a result, in self-supervised contrastive learning the training process requires $2N$ augmented samples for the N samples in training data, but in supervised contrastive learning, the model could be trained by either N or $2N$ augmented samples. In our task we use supervised contrastive learning to train the model. We tried both N and $2N$ augmented samples. The Supervised Contrastive Loss function is defined as:

$$SCL = \sum_{i \in I} \frac{-1}{|C(i)|} \sum_{c \in C(i)} \log \frac{\exp(z_i * z_c/\tau)}{\sum_{a \in A(i)} \exp(z_i * z_a/\tau)} \quad (1)$$

Here i is the index of an arbitrary sample in the augmented dataset I. $C(i)$ is the set of samples in the same class with i except sample i. $A(i)$ is the set of samples in the augmented dataset except sample i. z_i, z_c and z_a stand for the representations of the anchor, positive, and negative samples respectively. τ is the temperature parameter, which adjusts the distance of different classes in the embedding space.

4.2 Framework Architecture

Our proposed framework is based on supervised contrastive learning. The architecture of the framework is shown in Figure 4. The SCL can be implemented in two stages. In the first stage, we augment each labeled sentence into two sentences with similar semantics. This augmented dataset is fed into the encoder and supports the **Stage 1** training. The encoder along with the projection head, which is composed of several dense layers, minimizes the supervised contrastive loss to obtain the optimal embeddings in order to group positive samples together and push negative samples far away. In **Stage 2**, we keep the encoder and freeze the weights in its dense layers, and add two more dense layers for classification. The classifier is trained to minimize the cross-entropy loss function.

4.3 Data Augmentation

Data augmentation is an essential part in contrastive learning methods, which creates the dataset used for pre-training by sentences with similar semantics. We investigate five types of methods and their variants to augment text given a labeled sentence.

(1) **Round Trip Translation (RTT) [36].** This method first translates the sentence from English to French and then translates it back to English. Translation is based on Google translation services as well as Amazon translate [36].
(2) **Wordnet Synonym Replacement [36].** This method replaces words with their synonyms in the sentence. Replaceable words such as verbs, nouns are selected from a sentence using a part-of-speech tagger. Then a number of words are selected out of them following a Geometric distribution and replaced by their synonyms, which are given by a synonym library provided by WordNet.
(3) **EDA (Easy Data Augmentation) Synonym Replacement [37].** Randomly pick a word (not stop words) from the sentence and then replace the word with one of its synonyms chosen at random.
(4) **EDA Random Deletion [37].** Randomly remove any word in the sentence with a probability you can specify. We use the default probability value 0.2.
(5) **EDA Random Insertion [37].** Find a random synonym of a random word (not a stop word) in the sentence and then insert the synonym into any position in the sentence randomly.

We further generate augmented data by two data augmentation methods to obtain a bigger dataset for pre-training. A comparison of the results will be given in Section 7.

5 EXPERIMENT SETUPS

5.1 Base Models

As mentioned above, the first stage is to encode the input sentence into a vector. We experiment three types of encoders each having three settings of the original encoder, the encoder trained with transfer learning and the encoder trained on SCL.

(1) **CNN-1D.** Similar to regular CNN used in feature extraction from 2-dimensional images, 1-dimensional CNN has been used for extracting features from word sequences, e.g., [38]. This method works by sliding a window with a fix-width over a sequence and convolving features of tokens covered by the window [39]. An average pooling was used to aggregate features from individual tokens. Similar to a 2D-CNN, the 1D CNN can be used for extracting patterns from local 1D patches (aka sub-sequences)

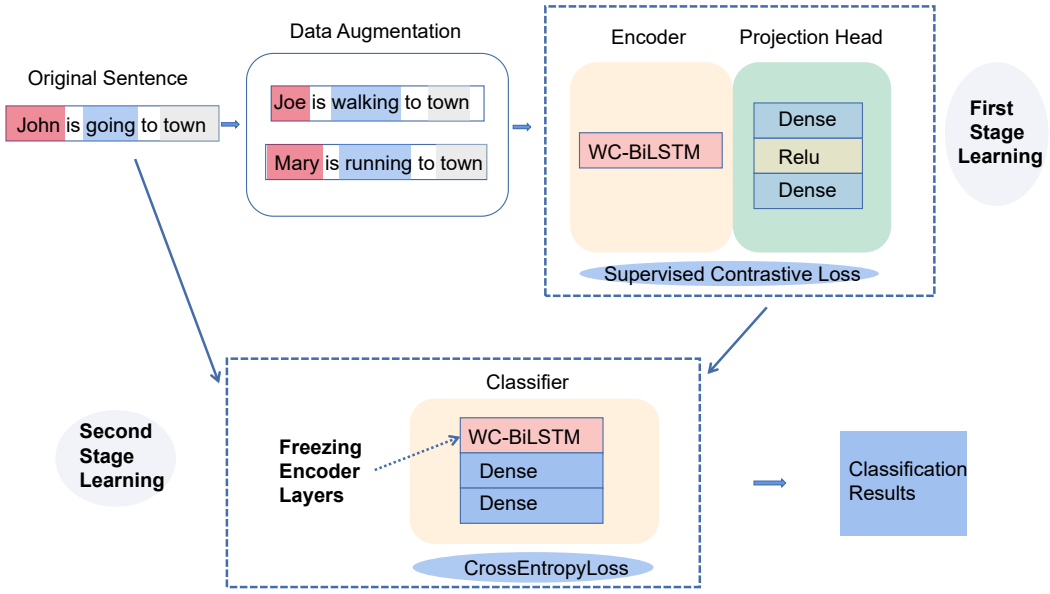

Figure 4: Architecture of our proposed framework: ClaimDistiller. The encoder can be customized.

from sequences. After each word-level token is converted to initial vectors, 1D convolutional layers with the convolutional kernels of size w were used to extract the patterns (Figure 6). These layers can recognize patterns in an input sequence. We used a 2-layer 1D CNN, which is flattened at the end before the presentation is fed to a dense fully-connected layer for classification.

(2) **USE-dense.** We adopted the pre-trained Universal Sentence Encoder (USE) [40] to encode claim text into dense 512-dimensional vectors. The initial embeddings produced by USE were fine-tuned on the SciCE corpus, after which the sentences were encoded to dense feature vectors used by the fully-connected layer for classification.

(3) **WC-BiLSTM (Word and Character embedding Bidirectional Long Short-Term Memory).** One drawback of applying pre-trained word embedding is that unseen words have to be encoded as a default vector in the prediction time. The representations of these words could only be inferred by surrounding words. Word prefixes and suffixes often contain semantic information. Therefore, we combine pre-trained Word2Vec embedding [41] with character embedding [42] to encode unseen words. The combined embedding is fed to bidirectional long short-term memory (BiLSTM) layers to extract patterns from claim sentences (Figure 7). Finally, the representations were passed to a fully-connected layer for classification.

5.2 Experiments

To evaluate the robustness of the proposed framework, we investigate the base models in different training frameworks: only the base model, transfer learning, and supervised contrastive learning. Figure 5 shows a comparison of the three different training frameworks. 'Network' in this figure can be any of the base models. The training frameworks are as follows:

(1) **Trained from Scratch.** In this setting, the neural classifier is trained directly using the SciCE corpus with only the base models, namely CNN-1D, USE-Dense, and WC-BiLSTM, described in the previous subsection.

(2) **Transfer Learning.** In this setting, the neural classifier is firstly pre-trained using the PubMed-RCT corpus and then fine-tuned on the SciCE corpus. During the fine-tuning stage, we freeze the weights of all layers except the fully-connected classification layer. Then we replaced that fully-connected layer with a new layer with classes in the target dataset.

(3) **Supervised Contrastive Learning.** As discussed in Section 4, in supervised contrastive learning the neural network is firstly pre-trained with augmented training data from the SciCE corpus and then fine-tuned on the original SciCE data. Note that in this setting, only SciCE is used, which is a dataset much smaller than the PubMed-RCT dataset.

As a result, we have in total 9 experiment specifications: 3 different frameworks for each base model. In addition, we include the following two experiments from previous academic papers as baselines:

(1) **Heuristic Method.** This baseline is adopted from Sateli & Witte [31]. This method used gazetteering, deictic phrases and handcrafted rules to match against the text. The sentence containing the deictic phrase must be a statement in form of a factual implication, and have a comparative voice or asserts a property of the author's contribution, such as novelty or performance.

(2) **CRF-based Transfer Learning.** This baseline is adopted from Achakulvisut et al. [14], in which transfer learning was applied

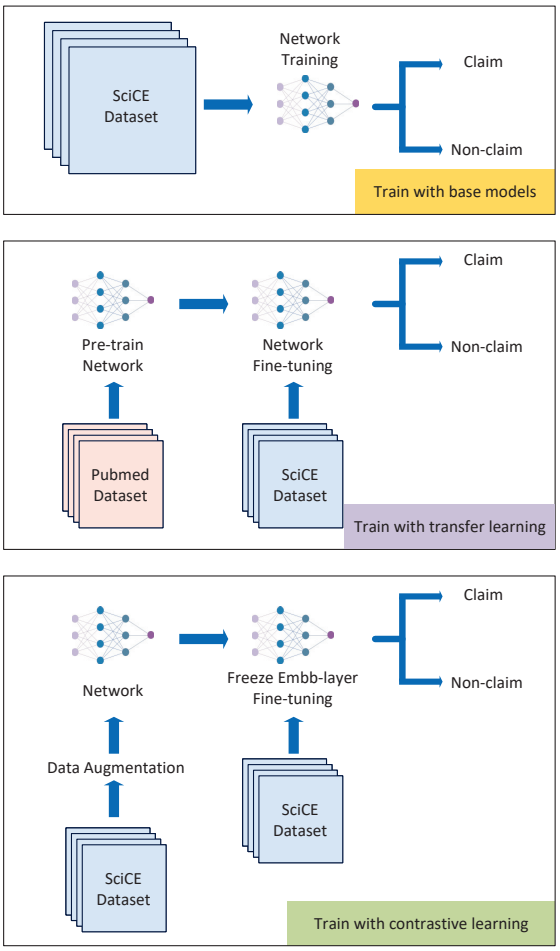

Figure 5: Comparing training from scratch, transfer learning and supervised contrastive learning .

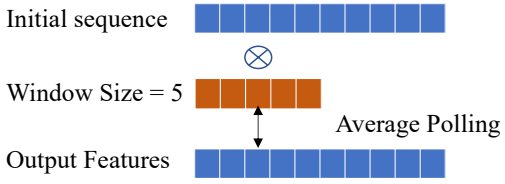

Figure 6: CNN-1D Working flows.

on a conditional random field (CRF) model. This is the state-of-the-art to our best knowledge. This method treats claim extraction as a sequence tagging task and uses CRF to capture the dependencies of the label of the current sentence to the features and labels of neighbor sentences.

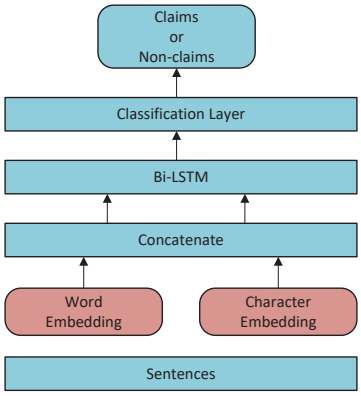

Figure 7: The architecture of the WC-BiLSTM base model.

6 EVALUATION

6.1 Evaluation Metrics

The proposed methods and baselines are evaluated using the standard precision, recall, and F1 scores, defined below.

$$P = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FP}}}, \quad R = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FN}}}, \quad F_1 = \frac{2PR}{P + R} \tag{2}$$

In Eq.(2), P and R stand for precision and recall, respectively. N_{TP} is the number of predicted claims that are true. N_{FP} is the number of predicted claims that are false. N_{FN} is the number of predicted non-claims that are false. F_1 is the harmonic mean of P and R.

In addition, we also compare the training time. The training time was measured as the time elapsed between when the program started taking inputs (including pre-training) and when the model stopped training after certain numbers of epochs.

6.2 Experiment Details

All the experiments were performed on a single computer with a 4 physical core CPU, 16GB RAM, and Solid State Disks and an Nvidia V100 GPU.

When working on the CNN-1D model, the window size w = 5. Because the convolution is performed on the word level, we truncated sentences longer than 120 words and padded sentences shorter than 120 words.

When training the WC-BiLSTM model, the learning rate was set to 0.001, the batch size was set to 256, and the dropout rate was 0.5. Each encoder model and its variants were trained for a maximum of 50 epochs before which the loss function of the validation data reached the minimum. Early stopping was applied to avoid overfitting. We found that at this stage, the loss functions have asymptotically converged to the minimum.

6.3 Results

The results of the experiments are shown in Table 2. The first column shows the evaluation results of models trained on **SciCE**. The second column shows the training time. The third column shows the evaluation results of models trained on **SciARK**. The

Table 2: A comparison of models on the scientific claim extraction task. Base models are trained from scratch. The model with the best performance is highlighted in bold.

Model Names	SciCE (2933) [2]			Time [3]	SciARK (3558) [2]			Time [3]	Data Size [4]
	P %	R %	F1 %	(sec)	P %	R %	F1 %	(sec)	(training)
Baseline Models									
Rule-based [31][1]	31.50	32.20	31.90	–	–	–	–	–	–
Transfer CRF [14][1]	86.60	72.70	79.00	–	–	–	–	–	–
Base Models									
CNN-1D	83.43	83.73	83.57	12	75.42	86.84	80.72	9	SciCE: 5823
USE-Dense	82.42	83.73	83.06	15	85.74	87.99	86.85	11	SciARK: 4768
WC-BiLSTM	83.35	84.65	83.99	150	86.66	87.83	87.24	38	
Transfer Learning									
CNN-1D-transfer	84.45	85.74	85.09	4502	85.57	86.38	85.97	1388	pre: 2 million
USE-Dense-transfer	85.81	86.71	86.24	12735	87.05	88.32	87.68	2878	tune: 5823 (SciCE)
WC-BiLSTM-transfer	84.87	84.59	84.73	49324	87.61	88.70	88.15	16245	tune: 4768 (SciARK)
Contrastive Learning									
CNN-1D-contrastive	85.80	86.49	86.14	108	84.86	86.05	85.45	48	SciCE: 5823
USE-Dense-contrastive	86.84	87.28	87.06	11823	89.06	89.74	89.40	3489	SciARK: 4768
ClaimDistiller [5]	**87.08**	**87.83**	**87.45**	15001	88.93	90.02	**89.47**	7201	

[1] Quoted from reference because they used the same test data.
[2] Testing data size is in the parentheses. Measured by number of sentences.
[3] Training time including both pre-training and fine-tuning.
[4] Measured by number of sentences in training dataset.
[5] WC-BiLSTM-contrastive.

training time on **SciARK** is shown in column 4. In column 5 we compared the training data size for all scenarios.

As seen in Table 2, Deep learning based models achieving much better performance than rule-based models suggests that the semantic features of scientific claims are complicated and are better represented by neural models.

In general, transfer learning based models achieve better performance than the corresponding original encoders by Δ F1=0.74–3.18. The efficacy of transfer learning comes from source data used for pre-training. The discourse information in the **PubMed-RCT** corpus used here is relevant and helps improve the performance.

The comparison of transfer learning and contrastive learning is performed on two datasets: **SciCE** and **SciARK**. Contrastive learning achieves better performance than transfer learning consistently across all models. With **SciCE**, SCL beats transfer learning by Δ F1=0.82 – 2.72% for the SciCE dataset and Δ F1=1.32–1.72% for the SciARK dataset. The only exception is that CNN-1D-contrastive underperformed CNN-1D-transfer by 0.52%. Therefore, SCL in general achieves a comparable or better performance than transfer learning.

Contrastive-learning-based model ClaimDistiller has the best performance across all metrics compared with other models, achieving F1=87.45%, precision=87.08%, and recall=87.83%. With **SciARK**, ClaimDistiller has the best performance with F1=88.93%, precision=90.02%, and recall=89.47%.

The training time needed for each model varies. In general, transfer learning needs significantly more time for training than supervised contrastive learning. In the last column, we see a clear comparison of training data size for contrastive learning and transfer learning. Comparing the training data size, contrastive learning uses less than 6000 sentences while transfer learning uses 2 million sentences for pre-training in order to achieve the performance reported in Table 1.

7 DISCUSSION

7.1 Data Augmentation Analysis

As discussed in Section 4, we tried several methods of text augmentation. Here we show the experimental results obtained with the best model WC-BiLSTM-contrastive model in Table 3. The results show that various types of text augmentation methods have marginal effect on the classification performance of the SCL base model, with the range of F1 going from 86.11% to 87.45%. Wordnet synonym replacement achieves the best performance while random deletion is the worst. We choose to use the best one "Wordnet synonym Replacement" as the data augmentation method.

7.2 Error Analysis

In this section, we perform error analysis focusing on the best model: WC-BiLSTM-contrastive. Out of the 375 abstracts in testing set of **SciCE**, this model correctly predicts **all** the claims and non-claims in 125 abstracts. As shown in Figure 8, in the remaining 250 abstracts, the majority of them have 1–2 wrongly predicted

Table 3: A comparison of performances with different data augmentation methods.

DA Methods	P %	R %	F1 %
EDA Random Deletion	85.67	86.56	86.11
EDA Replacement	86.08	86.94	86.50
RTT	86.28	86.90	86.59
Original Data	86.59	87.38	86.98
EDA Random Insertion	86.85	87.62	87.23
WordNet	**87.08**	**87.83**	**87.45**
WordNet + RTT	86.42	87.11	86.76
WordNet + Insertion	86.88	87.55	87.21
WordNet + Deletion	86.25	86.36	86.30
WordNet + Replacement	86.51	87.35	86.92

sentences, with the maximum prediction errors of 4 in a single abstract. As shown in Figure 9, the error rates are all below 0.5 and with an average of 0.13.

We demonstrate two examples containing typical errors in the prediction results for case studies (Figure 10). The ground truth claims are highlighted in blue. Green labels mean the sentences are non-claims and red labels mean sentences are claims. Labels with red frames indicate wrong predictions. In the first example, the model is able to identify all the claims, but it mistakenly recognizes two sentences as claims. In the second example, there should be two claims but the model only identified one of them.

Example 1 is challenging because the two false positives look like claims but when they are read together with the first sentence, it is clear that the second sentence (starting with "The discussion emphasizes") describes what the authors have done in the paper and the third sentence (starting with "A fundamental need") describes a background, which is the motivation of the research. Example 2 contains a false negative. It is not straightforward to determine why the sentence starting with "Results indicated that" was misclassified to a non-claim because the leading pattern clearly indicates the sentence conveys key findings. The error analyses indicate that although the recurrent model attempted to incorporate context information, it may still miss the nuances of semantics. Fine-tuning the hyperparameters may help, but a more sophisticated and robust model is needed to capture the nuances. One method is to combine latent and rule-based features. Another possible method is to leverage the "knowledge" encoded in large language models (LLMs), e.g., GPT 3, or using the LLM-adapter method to train an adapter for this task.

7.3 Domain Adaptability

The SciCE corpus is in the biomedical domain. To test whether the model performs well in a different domain, we applied the best model (WC-BiLSTM-contrastive) to classify sentences in a random selection of 30 abstracts in computer science papers. Out of the 195 sentences in this dataset, 60 sentences were predicted as claims. By visually examining these predicted claims, 50 of them are consistent with the definition of claims [14]. Examples of the successfully predicted claims are given in Figure 11. This post-hoc evaluation result indicates that the model's precision for computer science

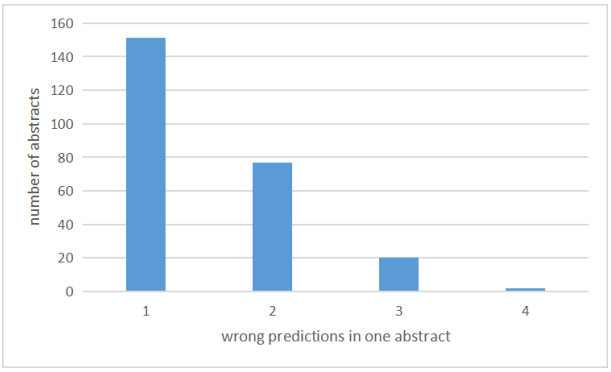

Figure 8: The distribution of the number of wrong predictions per abstract in the testing dataset of SciCE. In the abstracts where there are wrong predictions, the majority of them have 1 or 2 wrongly predicted sentences.

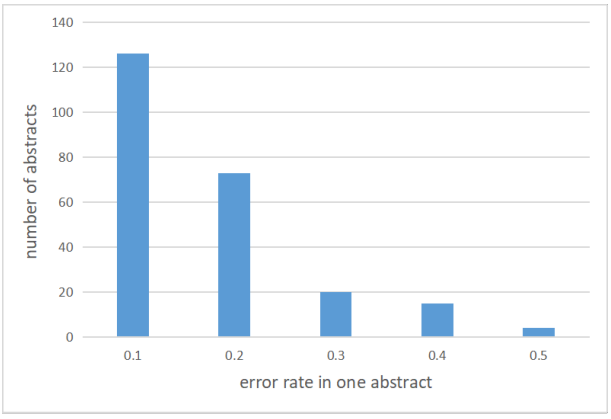

Figure 9: Distribution of error rate in the testing dataset of SciCE. The error rate is defined as wrongly-predicted sentences divided by the total number of sentences in an abstract. In most cases, the error rate is below 0.2.

abstracts is roughly consistent with biomedical domains with ($P \approx 83.3\%$), implying that claims in these two different domains are usually written with similar language patterns. We also observed that the model tends to omit claims, indicating that a more robust domain adaptation may be needed to improve the recall.

7.4 Visualization on SciCE Data

To further qualitatively demonstrate the effect of supervised contrastive learning, we project the 128-dimensional vectors output by the WC-BiLSTM base model into a 2-dimensional feature space using tSNE [43], and then compare it with results in supervised contrastive learning. Figure 12 shows that the model with supervised contrastive learning grouped the same class altogether, making them more separated in the feature space.

Prediction results on abstract 1:

Grounded in a socio-ecological framework, we describe salient health care system and policy factors that influence engagement in human immunodeficiency virus (HIV) clinical care. Non-claim

The discussion emphasizes successful programs and models of service delivery and highlights the limitations of current, fragmented health care system components in supporting effective, efficient, and sustained patient engagement across a continuum of care. Claim

A fundamental need exists for improved synergies between funding and service agencies that provide HIV testing, prevention, treatment, and supportive services. Claim

We propose a feedback loop whereby actionable, patient-level surveillance of HIV testing and engagement in care activities inform educational outreach and resource allocation to support integrated \"testing and linkage to care plus\" service delivery. Claim

Ongoing surveillance of programmatic performance in achieving defined benchmarks for linkage of patients who have newly diagnosed HIV infection and retention of those patients in care is imperative to iteratively inform further educational efforts, resource allocation, and refinement of service delivery. Claim

Prediction results on abstract 2:

"Bovine viral diarrhea virus (BVDV) is an emerging pathogen in alpacas and many questions still persist regarding disease mechanisms and control strategies." Non-claim

"The purpose of this study was to evaluate a commercial BVDV vaccine for safety and efficacy in alpacas." Non-claim

"Five nonpregnant alpacas were vaccinated with a modified-live BVDV vaccine and challenged 25 days post-immunization by nasal and ocular inoculation with a BVDV Type 1b strain isolated from a confirmed BVDV persistently infected alpaca." Non-claim

"Two nonpregnant alpacas served as non-vaccinated controls and were similarly challenged." Non-claim

"Results indicated that BVDV virus could not be detected from the vaccinated alpacas but was detected in the unvaccinated alpacas." Non-claim

"Results suggest that administration of modified-live BVDV vaccine protected the alpacas in this study from experimental challenge and no adverse effects from the vaccine were observed." Claim

Figure 10: Two examples of errors in the prediction results in the test set of SciCE. The ground truth claims are highlighted in blue. Green labels mean the sentences are non-claims and red labels mean sentences are claims. Labels with red frames indicate wrong predictions.

We demonstrate that scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even becoming competitive with prior state-of-the-art fine-tuning approaches.

These results highlight the importance of previously overlooked design choices, and raise questions about the source of recently reported improvements.

Moreover, DETR can be easily generalized to produce panoptic segmentation in a unified manner.

We show that it significantly outperforms competitive baselines.

Figure 11: Successful prediction results from papers in Computer Science domain.

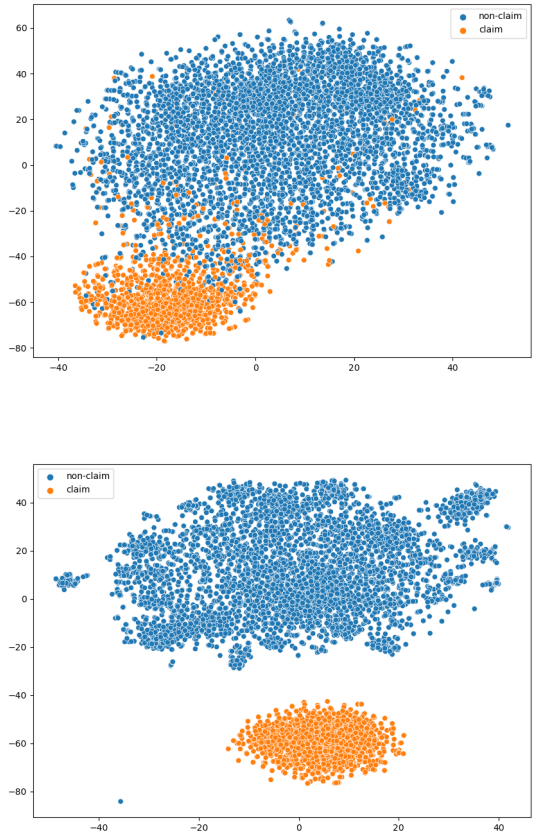

Figure 12: The t-SNR plots showing the effects of supervised contrastive learning. The upper panel shows the two classes without supervised contrastive learning. The lower panel shows the two classes with supervised contrastive learning. Orange dots represent claims and blue dots represent non-claims.

8 CONCLUSION

To automatically obtain scientific findings from the ever increasing volume of scientific papers, an effective and efficient claim-extracting tool is becoming increasingly important for information aggregation, summarization, and retrieval of scientific papers. One bottleneck of this task is the limitation of annotated training data. The challenge is how to efficiently use existing limited data. We propose the **ClaimDistiller** framework, which uses supervised contrastive learning on top of existing text encoders to boost the performance of classification. We showcased the efficacy of this mechanism on two benchmark datasets. Our result establish a new state-of-the-art on the SciCE dataset, outperforming the existing method by 7%, which used transfer learning on a BiLSTM-CRF architecture. We demonstrated that the SCL achieved comparable or higher F1 scores compared with transfer learning methods with significantly less training data and time. Future research will explore hybrid methods and LLMs to capture nuances of context. .

REFERENCES

[1] Rosina Weber. Applying artificial intelligence in the science & technology cycle. *Inf. Serv. Use*, 39(4):303–318, 2019.
[2] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Dan S Weld. S2orc: The semantic scholar open research corpus. *arXiv preprint arXiv:1911.02782*, 2019.
[3] Mario Lipinski, Kevin Yao, Corinna Breitinger, Joeran Beel, and Bela Gipp. Evaluation of header metadata extraction approaches and tools for scientific pdf

documents. In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '13, pages 385–386, New York, NY, USA, 2013. ACM.

[4] Noah Siegel, Nicholas Lourie, Russell Power, and Waleed Ammar. Extracting scientific figures with distantly supervised neural networks. In Jiangping Chen, Marcos André Gonçalves, Jeff M. Allen, Edward A. Fox, Min-Yen Kan, and Vivien Petras, editors, *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries, JCDL 2018, Fort Worth, TX, USA, June 03-07, 2018*, pages 223–232. ACM, 2018.

[5] Florin Bulgarov and Cornelia Caragea. A comparison of supervised keyphrase extraction models. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pages 13–14, New York, NY, USA, 2015. ACM.

[6] Yihong Gong and Xin Liu. Generic text summarization using relevance measure and latent semantic analysis. In W. Bruce Croft, David J. Harper, Donald H. Kraft, and Justin Zobel, editors, *SIGIR 2001: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, September 9-13, 2001, New Orleans, Louisiana, USA*, pages 19–25. ACM, 2001.

[7] Ye Liu, Jian-Guo Zhang, Yao Wan, Congying Xia, Lifang He, and Philip S. Yu. HETFORMER: heterogeneous transformer with sparse attention for long-text extractive summarization. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 146–154. Association for Computational Linguistics, 2021.

[8] Tianwen Jiang, Tong Zhao, Bing Qin, Ting Liu, Nitesh V Chawla, and Meng Jiang. The role of" condition" a novel scientific knowledge graph representation and construction model. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1634–1642, 2019.

[9] Matthew Hutson. Artificial intelligence faces reproducibility crisis. *Science*, 359(6377):725–726, 2018.

[10] Colin F. Camerer, Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A. Nosek, Thomas Pfeiffer, Adam Altmejd, Nick Buttrick, Taizan Chan, Yiling Chen, Eskil Forsell, Anup Gampa, Emma Heikensten, Lily Hummer, Taisuke Imai, Siri Isaksson, Dylan Manfredi, Julia Rose, Eric-Jan Wagenmakers, and Hang Wu. Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour*, 2(9):637–644, 2018.

[11] Nazanin Alipourfard, Beatrix Arendt, Daniel M Benjamin, Noam Benkler, Michael M Bishop, Mark Burstein, Martin Bush, James Caverlee, Yiling Chen, Chae Clark, and et al. Systematizing confidence in open research and evidence (score), May 2021.

[12] Maria Liakata, Simone Teufel, Advaith Siddharthan, and Colin Batchelor. Corpora for the conceptualisation and zoning of scientific papers. 2010.

[13] Beatriz Fisas, Horacio Saggion, and Francesco Ronzano. On the discoursive structure of computer graphics research papers. In *Proceedings of the 9th linguistic annotation workshop*, pages 42–51, 2015.

[14] Titipat Achakulvisut, Chandra Bhagavatula, Daniel E. Acuna, and Konrad P. Körding. Claim extraction in biomedical publications using deep discourse model and transfer learning. *CoRR*, abs/1907.00962, 2019.

[15] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2021.

[16] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 806–813, 2014.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[18] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 491–507. Springer, 2020.

[19] Philip Bachman, R Devon Hjelm, and William Buchwalter. *Learning Representations by Maximizing Mutual Information across Views*. Curran Associates Inc., Red Hook, NY, USA, 2019.

[20] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 2020.

[21] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2020.

[22] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

[23] Araly Barrera and Rakesh M. Verma. Combining syntax and semantics for automatic extractive single-document summarization. In Alexander F. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing - 13th International Conference, CICLing 2012, New Delhi, India, March 11-17, 2012, Proceedings, Part II*, volume 7182 of *Lecture Notes in Computer Science*, pages 366–377. Springer, 2012.

[24] Nouf Ibrahim Altmami and Mohamed El Bachir Menai. Automatic summarization of scientific articles: A survey. *Journal of King Saud University - Computer and Information Sciences*, 34(4):1011–1028, 2022.

[25] John Lawrence and Chris Reed. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818, 2020.

[26] Christos Sardianos, Ioannis Manousos Katakis, Georgios Petasis, and Vangelis Karkaletsis. Argument extraction from news. In *Proceedings of the 2nd Workshop on Argumentation Mining, ArgMining@HLT-NAACL 2015, June 4, 2015, Denver, Colorado, USA*, pages 56–66. The Association for Computational Linguistics, 2015.

[27] Mihai Dusmanu, Elena Cabrio, and Serena Villata. Argument mining on twitter: Arguments, facts and sources. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2317–2322. Association for Computational Linguistics, 2017.

[28] Christian Stab, Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. Argumentation mining in persuasive essays and scientific articles from the discourse structure perspective. In Elena Cabrio, Serena Villata, and Adam Z. Wyner, editors, *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing, Forlì-Cesena, Italy, July 21-25, 2014*, volume 1341 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2014.

[29] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and verification. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 809–819. Association for Computational Linguistics, 2018.

[30] Tom Jansen and Tobias Kuhn. Extracting core claims from scientific articles. In Tibor Bosse and Bert Bredeweg, editors, *BNAIC 2016: Artificial Intelligence - 28th Benelux Conference on Artificial Intelligence, Amsterdam, The Netherlands, November 10-11, 2016, Revised Selected Papers*, volume 765 of *Communications in Computer and Information Science*, pages 32–46. Springer, 2016.

[31] Bahar Sateli and René Witte. Semantic representation of scientific literature: bringing claims, contributions and named entities onto the linked open data cloud. *PeerJ Computer Science*, 1:e37, December 2015.

[32] Shi Yuan and Bei Yu. Hclaime: A tool for identifying health claims in health news headlines. *Inf. Process. Manag.*, 56(4):1220–1233, 2019.

[33] Mostafa Dehghani, Aliaksei Severyn, Sascha Rothe, and Jaap Kamps. Learning to learn from weak supervision by full supervision, 2017.

[34] Franck Dernoncourt and Ji Young Lee. Pubmed 200k RCT: a dataset for sequential sentence classification in medical abstracts. In Greg Kondrak and Taro Watanabe, editors, *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017, Volume 2: Short Papers*, pages 308–313. Asian Federation of Natural Language Processing, 2017.

[35] Aris Fergadis, Dimitris Pappas, Antonia Karamolegkou, and Harris Papageorgiou. Argumentation mining in scientific literature for sustainable development. In *Proceedings of the 8th Workshop on Argument Mining*, pages 100–111, 2021.

[36] Vukosi Marivate and Tshephisho Sefara. Improving short text classification through global augmentation methods. In *Machine Learning and Knowledge Extraction: 4th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2020, Dublin, Ireland, August 25–28, 2020, Proceedings 4*, pages 385–399. Springer, 2020.

[37] Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*, 2019.

[38] Edward Kim, Kevin Huang, Alex Tomala, Sara Matthews, Emma Strubell, Adam Saunders, Andrew McCallum, and Elsa Olivetti. Machine-learned and codified synthesis parameters of oxide materials. *Scientific Data*, 4(1):170127, 2017.

[39] Serkan Kiranyaz, Onur Avci, Osama Abdeljaber, Turker Ince, Moncef Gabbouj, and Daniel J. Inman. 1d convolutional neural networks and applications: A survey. *Mechanical Systems and Signal Processing*, 151:107398, 2021.

[40] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. Universal sentence encoder for english. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*, pages 169–174, 2018.

[41] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

[42] Ziang Xie, Anand Avati, Naveen Arivazhagan, Dan Jurafsky, and Andrew Y. Ng. Neural language correction with character-based attention. *CoRR*, abs/1603.09727, 2016.

[43] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.