DeSpin: a prototype system for detecting spin in biomedical publications

Anna Koroleva
Zurich University of Applied Sciences (ZHAW),
Waedenswil, Switzerland
Swiss Institute of Bioinformatics (SIB),
Lausanne, Switzerland
aakorolyova@gmail.com

Sanjay Kamath
Total,
France
sanjay@lri.fr

Patrick M.M. Bossuyt
Academic Medical Center,
University of Amsterdam,
Amsterdam, Netherlands
p.m.bossuyt@amsterdamumc.nl

Patrick Paroubek
LIMSI, CNRS, Université Paris-Saclay,
Orsay, France
pap@limsi.fr

Abstract

Improving the quality of medical research reporting is crucial to reduce avoidable waste in research and to improve the quality of health care. Despite various initiatives aiming at improving research reporting – guidelines, checklists, authoring aids, peer review procedures, etc. – overinterpretation of research results, also known as distorted reporting or spin, is still a serious issue in research reporting.

In this paper, we propose a Natural Language Processing (NLP) system for detecting several types of spin in biomedical articles reporting randomized controlled trials (RCTs). We use a combination of rule-based and machine learning approaches to extract important information on trial design and to detect potential spin.

The proposed spin detection system includes algorithms for text structure analysis, sentence classification, entity and relation extraction, semantic similarity assessment. Our algorithms achieved operational performance for the these tasks, F-measure ranging from 79,42 to 97.86% for different tasks. The most difficult task is extracting reported outcomes.

Our tool is intended to be used as a semi-automated aid tool for assisting both authors and peer reviewers to detect potential spin. The tool incorporates a simple interface that allows to run the algorithms and visualize their output. It can also be used for manual annotation and correction of the errors in the outputs.

The proposed tool is the first tool for spin detection. The tool and the annotated dataset are freely available.

1 Background

It is widely acknowledged nowadays that that the quality of reporting of research results in the clinical domain is suboptimal. As a consequence, research findings can often not be replicated, and billions of euros may be wasted yearly (Ioannidis, 2005).

Numerous initiatives aim at improving the quality of research reporting. Guidelines and checklists have been developed for every type of clinical research. Still, the quality of reporting remains low: authors fail to choose and follow a correct guideline/checklist (Samaan et al., 2013). Automated tools, such as Penelope[1], are introduced to facilitate the use of guidelines/checklists. It was proved that authoring aids improve the completeness of reporting (Barnes et al., 2015).

Enhancing the quality of peer reviewing is another step to improve research reporting. Peer reviewing requires assessing a large number of information items. Nowadays, Natural Language Processing (NLP) is applied to facilitate laborious manual tasks such as indexing of medical literature (Huang et al., 2011) and systematic review process (Ananiadou et al., 2009). Similarly, the peer reviewing process can be partially automated with the help of NLP.

Our project tackles a specific issue of research reporting that, to our knowledge, has not been addressed by the NLP community: spin, also referred to as overinterpretation of research results. In the context of clinical trials assessing a new (experi-

At the time of reported work, Anna Koroleva was a PhD student at LIMSI-CNRS in Orsay, France and at the Academic Medical Center, University of Amsterdam in Amsterdam, the Netherlands. Sanjay Kamath was a PhD student at LIMSI-CNRS and LRI Univ. Paris-Sud in Orsay, France.

[1] https://www.penelope.ai/

Proceedings of the BioNLP 2020 workshop, pages 49–59

mental) intervention, spin consists in exaggerating the beneficial effects of the studied intervention (Boutron et al., 2010).

Spin is common in articles reporting randomized controlled trials (RCTs) - clinical trials comparing health interventions, to which participants are allocated randomly to avoid biases - with non-significant primary outcome. Abstracts are more prone to spin than full texts. Spin is found in a high percentage of abstracts of articles in surgical research (40%) (Fleming, 2016), cardiovascular diseases (57%) (Khan et al., 2019), cancer (47%) (Vera-Badillo et al., 2016), obesity (46.7%) (Austin et al., 2018), otolaryngology (70%) (Cooper et al., 2018), anaesthesiology (32,2%) (Kinder et al., 2018), and wound care (71%) (Lockyer et al., 2013). Although the problem of spin has started to attract attention in the medical community in the recent years, the shown prevalence of spin proves that it often remains unnoticed by editors and peer reviewers.

Abstracts are often the only part of the article available to readers, and spin in abstracts of RCTs poses a serious threat to the quality of health care by causing overestimation of the intervention by clinicians (Boutron et al., 2014), which may lead to the use of an ineffective of unsafe intervention in clinical practice. Besides, spin in research articles is linked to spin in press releases and health news (Haneef et al., 2015; Yavchitz et al., 2012), which has the negative impact of raising false expectations regarding the intervention among the public.

The importance of the problem of spin motivated our work. We aimed at developing NLP algorithms to aid authors and readers in detecting spin. We focused on randomized controlled trials (RCTs) as they are the most important source of evidence for Evidence-based medicine, and spin in RCTs has high negative impact.

Our work lies within the scope of the Methods in Research on Research (MiRoR) project[2], an international project devoted to improving the planning, conduct, reporting and peer reviewing of health care research. For the design and development of our toolkit, we benefited from advice from the MiRoR consortium members.

In this paper, we introduce a prototype of a system, called DeSpin (Detector of Spin), that automatically detects potential spin in abstracts of RCTs and relevant supporting information. This prototype comprises a set of spin-detecting algorithms and a simple interface to run the algorithms and display their output.

This paper is organized as follows: first, we provide an overview of some existing semi-automated aid systems for authors, reviewers and readers of biomedical articles. Second, we introduce in more detail the notion of spin, the types of spin that we address, and the information that is required to assess an article for spin. After that, we describe our current algorithms, methods employed and provide their evaluation. Finally, we discuss the potential future development of the prototype.

2 Related work

Although there has been no attempt to automate spin detection in biomedical articles, a number of works addressed developing automated aid tools to assist authors and readers of scientific articles in performing various other tasks. Some of these tools were tested and were shown to reduce the workload and improve the performance of human experts on the corresponding task.

2.1 Authoring aid tools

Barnes et al. (2015) assessed the impact of a writing aid tool based on the CONSORT statement (Schulz et al., 2010) on the completeness of reporting of RCTs. The tools was developed for six domains of the Methods section (trial design, randomization, blinding, participants, interventions, and outcomes) and consisted of reminders of the corresponding CONSORT item(s), bullet points enumerating the key elements to report, and good reporting examples. The tool was assessed in an RCT in which the participants were asked to write a Methods section of an article based on a trial protocol, either using the aid tool ('intervention' group) or without using the tool ('control' group). The results of 41 participants showed that the mean global score for reporting completeness was higher with the use of the tool than without it.

2.2 Aid tools for readers and reviewers

Kiritchenko et al. (2010) developed a system called ExaCT to automatically extract 21 key characteristics of clinical trial design, such as treatment names, eligibility criteria, outcomes, etc. ExaCT consists of an information extraction algorithm that looks for text fragments corresponding to the target information elements, a web-based user interface

[2] http://miror-ejd.eu/

through which human experts can view and correct the suggested fragments.

The National Library of Medicine's Medical Text Indexer (MTI) is a system providing automatic recommendations based on the Medical Subject Headings (MeSH) terms for indexing medical articles (Mork et al., 2013). MTI is used to assist human indexers, catalogers, and NLM's History of Medicine Division in their work. Its use by indexers was shown to grow over years (used to index 15.75% of the articles 2002 vs 62.44% in 2014) and to improve the performance (precision, recall and F-measure) of indexers (Mork et al., 2017).

Marshall et al. (2015) addressed the task of automating assessment of risk of bias in clinical trials. Bias is phenomenon related to spin: it is a systematic error or a deviation from the truth in the results or conclusions that can cause an under- or overestimation of the effect of the examined treatment (Higgins and Green, 2008). The authors developed a system called RobotReviewer that used machine learning to assess an article for the risk of different types of bias and to extract text fragments that support these judgements. These works showed that automated risk of bias assessment can be achieve reasonable performance, and the extraction of supporting text fragments reached similar quality to that of human experts. Marshall et al. (2017) further developed RobotReviewer, adding functionality for extracting the PICO (Population, Interventions/Comparators, Outcomes) elements from articles and detecting study design (RCT), for the purpose of automated evidence synthesis. Soboczenski et al. (2019) assessed RobotReviewer in a user study involving 41 participants, evaluating time spent for bias assessment, text fragment suggestions by machine learning, and usability of the tool. Semi-automation in this study was shown to be quicker than manual assessment; 91% of the automated risk of bias judgments and 62% of supporting text suggestions were accepted by the human reviewers.

The cited works demonstrate that semi-automated aid tools can prove useful for both authors and readers/reviewers of medical articles and has a potential to improve the quality of the articles and facilitate the analysis of the texts.

3 Spin: definition and types

We adopt the definition and classification of spin introduced by Boutron et al. (2010) and Lazarus et al. (2015), who divided instances of spin into several types and subtypes.

We addressed the following types of spin:

1. Outcome switching – unjustified change of the pre-defined trial outcomes, leading to reporting only the favourable outcomes that support the hypothesis of the researchers (Goldacre et al., 2019). Outcome switching is one of the most common types of spin. It can consist in omitting the primary outcome in the results / conclusions of the abstract, or in the focus on significant secondary outcomes, e.g.:

   *The primary end point of this trial was **overall survival**. <...> This trial showed a significantly increased **R0 resection rate** although it failed to demonstrate a **survival** benefit.*

   In this example, the primary outcome ("overall survival"), the results for which were not favourable, is mentioned in the conclusion, but it is not reported in the first place and occurs within a concessive clause (starting by "although"). This way of reporting puts the focus on the other, favourable, outcome ("R0 resection rate").

2. Interpreting non-significant outcome as a proof of equivalence of the treatments, e.g.:

   *The median PFS was 10.3 months in the XELIRI and 9.3 months in the FOLFIRI arm (**p = 0.78**). Conclusion: The XELIRI regimen showed **similar PFS** compared to the FOLFIRI regimen.*

   The results for the outcome "median PFS" are not significant, which is often erroneously interpreted as a proof of similarity of the treatments. However, a non-significant result means that the null hypothesis of a difference could not be rejected, which is not equivalent to a demonstration of similarity of the treatments. This would require the rejection of the null hypothesis of a difference, or a substantial difference, in outcomes between treatments.

3. Focus on within-group comparisons, e.g.:

   ***Both groups showed robust improvement** in both symptoms and functioning.*

   The goal of randomized controlled trials is to compare two treatments with regard to some outcomes. If the superiority of the experimental treatment over the control treatment was

not shown, within-group comparisons (reporting the changes within a group of patients receiving a treatment, instead of comparing patients receiving different treatments) can be used to persuade the reader of beneficial effects of the experimental treatment.

Two concepts are vital for spin detection and play a key role in our algorithms:

1. The primary outcome of a trial – the most important variable monitored during the trial to assess how the studied treatment impacts it. Primary outcomes are recorded in trial registries (open online databases storing the information about registered clinical trials), and should be defined in the text of clinical articles, e.g.:

   *The primary end point was **a difference of $>$ 20% in the microvascular flow index of small vessels among groups**.*

2. Statistical significance of the primary outcome. Statistical hypothesis testing is used to check for a significant difference in outcomes between two patient groups, one receiving the experimental treatment and the other receiving the control treatment. Statistical significance is often reported as a P-value compared to predefined threshold, usually set to 0.05. Spin most often occurs when the results for the primary outcome are not significant (Boutron et al., 2010; Fleming, 2016; Khan et al., 2019; Vera-Badillo et al., 2016; Austin et al., 2018; Cooper et al., 2018; Kinder et al., 2018; Lockyer et al., 2013), although trials with significant effect on the primary outcome may also be prone to spin (Beijers et al., 2017).

   Trial results are commonly reported as an effect on the (primary) outcome[3], along with the p-value.

   ***Microcirculatory flow indices of small and medium vessels** were **significantly** higher in the levosimendan group as compared to the control group (**$p < 0.05$**).*

   Statistical significance levels of trial outcomes are vital for spin detection, as spin is commonly related to non-significant results for the primary outcome, or to selective reporting of significant outcomes only.

[3]It is important to distinguish between the notions of outcome, effect and result in this context: an outcome is a measure/variable monitored during a clinical trial; effect refers to the change in an outcome observed during a trial; trial results refer to the set of effects for all measured outcomes.

4 Algorithms

Spin is a complex notion and thus detecting spin cannot be seen as a binary classification problem. We believe that the most viable approach to spin detection is to assess each (sub)type of spin separately. We aimed at developing algorithms to extract and analyse pieces of information relevant to the addressed types of spin. The extracted information and its analysis, provided by our tool, can help human experts in making the conclusion on presence or absence of spin of the given (sub)type.

Detection of spin and related information is a complex task which cannot by fully automated. Our system is designed as a semi-automated tool that finds potential instances of the addressed types of spin and extracts the supporting information that can help the user to make the final decision on the presence of spin. In this section, we present the algorithms currently included in the system, according to the types of spin that they are used to detect.

As we aim at detecting spin in the Results and Conclusions sections of articles' abstracts, we first need an algorithm analyzing the given article to detect its abstract and the Results and Conclusions sections within the abstract. We will not mention this algorithm in the list of algorithms for each spin type to avoid repetition. If we talk about extracting some information from the abstract, it implies that the text structure analysis algorithm was applied.

4.1 Outcome switching

We focus on the switching (change/omission) of the primary outcome. Primary outcome switching can occur at several points:

- the primary outcome(s) recorded in the trial registry can differ from the primary outcome(s) declared in the article;
- the primary outcome(s) declared in the abstract can differ from the primary outcome(s) declared in the body of the article;
- the primary outcome(s) recorded in the trial registry can be omitted when reporting the results for the outcomes in the abstract;
- the primary outcome(s) recorded in the article can be omitted when reporting the results for the outcomes in the abstract.

Primary outcome switching detection involves the following algorithms:

1. Identification of primary outcomes in trial registries and in the article's text.
2. Identification of reported outcomes from sentences reporting the results, e.g. (reported outcomes are in bold):

   *The results of this study showed that **symptom Scores** in massage group were improved significantly compared with control group, and the rate of **dyspnea**, **cough** and **wheeze** in the experimental group than the control group were reduced by approximately 45%, 56% and 52%.*

3. Assessment of semantic similarity of pairs of outcomes extracted by the above algorithms to check for missing outcomes. We perform the assessment for the following sets of outcomes:
   - The primary outcome extracted from the registry is compared to the primary outcome(s) declared in the article;
   - The primary outcome extracted from the abstract is compared to the primary outcome(s) declared in the body of the article;
   - The primary outcome extracted from the article is compared to the outcomes reported in the abstract;
   - The primary outcome extracted from the registry is compared to the outcomes reported in the abstract.

   These assessments allow to detect switching of the primary outcome at all the possible stages. If the primary outcome in the registry and in the article, or in the abstract and body of the article differ, we conclude that there is potential outcome switching, which is reported to the user. Similarly, if the primary outcome (from the article or from the registry) is missing from the list of the reported outcomes, we suspect selective reporting of outcomes, and the system reports it to the user.

   In the example on the page 3, the system should extract "overall survival" as the primary outcome, and "R0 resection rate" and "survival" as reported outcomes. The similarity between "overall survival" and "R0 resection rate" is low, while the similarity between "overall survival" and "survival" is high, thus, we conclude that the primary outcome "overall survival" is reported as "survival".

   As semantic similarity often depends on the context, the conclusions of the system are presented to the user, who can check them to make the conclusions on correctness of the analysis.

4. Assessing the discourse prominence of the reported primary outcome (detected by the previous algorithms) by checking if it is reported the first place among all the outcomes; if it is reported in a concessive clause.

   In the example above, the system will detect that the primary outcome "survival" is reported within a concessive clause (starting by "although") and will flag the sentence as potentially focusing on secondary outcomes.

4.2 Interpreting non-significant outcome as a proof of equivalence of the treatments

As we stated above, conclusions on the similarity/equivalence of the studies treatments are justified only if the trial was of non-inferiority or equivalence type. Thus, we employ two algorithms to detect this type of spin:

1. Identification of statements of similarity between treatments, e.g.:

   *Both products caused **similar** leukocyte counts diminution and had **similar** safety profiles.*

2. Identifying the markers of non-inferiority or equivalence trial design, e.g.:

   *ONCEMRK is a phase 3, multicenter, double-blind, **noninferiority** trial comparing raltegravir 1200mg QD with raltegravir 400mg BID in treatment-naive HIV-1–infected adults.*

If there is a statement of similarity of treatments while no markers of non-inferiority / equivalence design are found, we conclude the presence of spin and report it to the user.

4.3 Focus on within-group comparisons

Any statement in the results and conclusions of the abstract that presents a comparison of two states of a patient group without comparing it to another group is a within-group comparison. This type of spin is detected by a single algorithm that identifies

within-group comparisons that are further reported to the user:

Young Mania Rating Scale total scores ***improved with ritanserin***.

4.4 Other algorithms

We support extraction of some information that is not directly involved in the detection of spin, but that can help user in spin assessment and that can be used in the future when new spin types are added. The algorithms include:

1. Extraction of measures of statistical significance, both numerical and verbal (in bold):

   Study group patients had a ***significant*** *lower reintubation rate than did controls; six patients (17%) versus 19 patients (48%),* ***P<0.05****; respectively.*

2. Extraction of the relation between the reported outcomes and their statistical significance, extracted at the previous stages. For the example above, we extract pairs ("reintubation rate", "significant") and ("reintubation rate", "P<0.05").

   These algorithms, in combination with the assessment of semantic similarity of extracted outcomes, allows to identify the significance level for the primary outcome.

5 Methods

In this section, we briefly outline the methods used in our algorithms, the datasets used for evaluation, and the current performance of the algorithms. Our approach is based on some previous works for the related tasks. As the details on development of the algorithms, annotating the data and testing different approaches are described in detail in the corresponding articles, we limit ourselves here to only a brief description of the best-performing method that we selected for each task.

The methods we employ can be divided into two groups: machine learning, including deep learning, used for the core tasks for which we have sufficient training data, and rule-based methods, used for the simpler tasks or for tasks where we do not have enough data for machine learning.

5.1 Rule-based methods

We developed rules for the following tasks:

- To find the abstract, we use regular expressions rules that are evaluated on the set of 3938 PubMed Central (PMC)[4] articles in XML format with a specific tag for the abstract, used as the gold standard. To evaluate our algorithm, we applied it to the raw texts extracted from the XML files and compared the extracted abstracts to those obtained using the XML tag.

- To extract outcomes from trial registries, we use regular expressions to extract the trial registration number from the article; using it, we find on the web, download and parse the registry entry corresponding to the trial.

- To extract significance levels, we use rules based on regular expressions and token, lemma and pos-tag information.

- To assess the discourse prominence of an outcome, to detect statements of similarity between treatments, within-group comparisons and markers of non-inferiority design, we employ rules based on token, lemma and pos-tag information.

  We annotated abstracts of 180 articles (2402 sentences) for similarity statements and within-group comparisons (Koroleva, 2020). The proportion of these types of statements in our corpus is low: we identified only 72 similarity statements and 127 within-group comparisons. The evaluation of statements of similarity between treatments and within-group comparisons was performed with two settings: 1) using the whole text of abstracts; 2) using only the Results and Conclusions sections of the abstract, which raised the precision, as expected (Table 1).

5.2 Machine learning methods

For the core tasks of our system, we either used an existing annotated corpus or annotated our own corpora. Our corpora were annotated by a single annotator (AK), consulted by consulted our medical advisors from the MiRoR network (Isabelle Boutron, Patrick Bossuyt and Liz Wager).

We tested several approaches for each task, including rule-based and machine-learning approaches (see details below). Overall, we found that the best performance on our tasks was shown

[4] https://www.ncbi.nlm.nih.gov/pmc/

Algorithm	Method	Annotated dataset	Precision	Recall	F1
Primary outcomes extraction	Deep learning	2,000 sentences / 1,694 outcomes	86.99	90.07	88.42
Reported outcomes extraction	Deep learning	1,940 sentences / 2,251 outcomes	81.17	78.09	79.42
Outcome similarity assessment	Deep learning	3,043 pairs of outcomes	88.93	90.76	89.75
Similarity statements extraction	Rules	180 abstracts / 2402 sentences whole abstract results and conclusions	 77.8 85.1	 87.5 87.5	 82.4 86.3
Within-group comparisons	Rules	180 abstracts / 2402 sentences whole abstract results and conclusions	 53.2 71.9	 90.6 90.6	 67.1 80.1
Abstract extraction	Rules	3938 abstracts	94.7	94	94.3
Text structure analysis: sections of abstract	Deep learning	PubMed200k	97.82	95.81	96.8
Extraction of significance levels	Rules	664 sentences / 1,188 significance level markers	99.18	96.58	97.86
Outcome - significance level relation extraction	Deep learning	2,678 pairs of outcomes and significance level markers	94.3	94	94

Table 1: Overview of algorithms, methods, results and annotated datasets

by a deep learning approach that was recently proved to be highly successful in many NLP applications. It employs language representations pretrained on large unannotated data and fine-tuned on a relatively small amount of annotated data for a specific downstream task. The language representations that we tested include: BERT (Bidirectional Encoder Representations from Transformers) models (Devlin et al., 2018), trained on a general-domain corpus of 3.3B words; BioBERT model (Lee et al., 2019), trained on the BERT corpus and a biomedical corpus of 18B words; and SciBERT models (Beltagy et al., 2019), trained on the BERT corpus and a scientific corpus of 3.1B words. For each task, we chose the best-performing model.

Details about the annotated datasets that we used and the tested approaches can be found below. The best results for each task are summarised in Table 1.

5.2.1 Identification of sections in the abstract

For identifying sections within the abstract (in particular, Results and Conclusions), we used the PubMed 200k dataset introduced in Dernoncourt and Lee (2017). This dataset contains approximately 200,000 abstracts of RCTs with 2.3 million sentences. Each sentence is annotated with one of the following classes, corresponding to the sections of the abstract: background, objective, method, result, or conclusion. We used the train-dev-test split provided by the developers of the dataset.

We compared a rule-based approach and BERT, SciBERT and BioBERT models, fine-tuned for the sentence classification task on the PubMed 200k dataset. The best performance was shown by the fine-tuned BioBERT model.

5.2.2 Outcome extraction

The outcome extraction task includes two subtasks: extracting primary and reported outcomes. For each subtask, we annotated a separate corpus. For primary outcome extraction, we annotated a corpus of 2,000 sentences, coming from 1,672 articles. The sentences were selected randomly, from both abstracts and full texts, without restriction to a particular medical domain. A total of 1,694 primary outcomes was annotated (Koroleva, 2019a). For reported outcome extraction, we annotated reported outcomes in the abstracts of articles for which we annotated the primary outcomes. The corpus contains 1,940 sentences from 402 articles, with a total of 2,251 reported outcomes (Koroleva, 2019a).

We compared a rule-based system and several machine learning algorithms for primary and reported outcome extraction. Details about the annotated datasets and the methods that we tested can be found in Koroleva et al. (EasyChair, 2020). We selected the best performing approach to be included in our tool

For primary outcomes extraction, the best performance was demonstrated by the BioBERT model

fine-tuned for named entity recognition task on our corpus of 2,000 sentences annotated for primary outcomes. For reported outcomes extraction, the best performance was achieved by the SciBERT model fine-tuned for named entity recognition task on our corpus of 1,940 sentences annotated with reported outcomes.

5.2.3 Assessment of semantic similarity of outcomes

To annotate semantic similarity between outcomes, we used pairs of sentences from our corpora of outcomes: the first sentence in each pair comes from the corpus of primary outcomes, the second sentence comes from the corpus of reported outcomes, and both sentences are from the same article. We assigned a binary label of similarity (similar/dissimilar) to each pair of outcomes in each sentence pair. The corpus contains 3,043 pairs of outcomes (Koroleva, 2019b).

We tested several semantic similarity measures (string-based, lexical, vector-based) and the BERT, SciBERT and BioBERT models, fine-tuned for sentence pair classification task on the corpus of outcome pairs. Details on the corpus annotation and on the methods tested can be found in Koroleva et al. (2019). The best performance was shown by the fine-tuned BioBERT model.

5.2.4 Extraction of the relation between reported outcomes and statistical significance levels

To annotate the relation between reported outcomes and statistical significance levels, we selected sentences containing markers of statistical significance from the corpus annotated with reported outcomes. We annotated the pairs of outcomes and significance levels with a binary label ("positive": the significance level is related to the outcome; "negative": the significance level is not related to the outcome). The final corpus contains 663 sentences with 2,552 annotated relations (Koroleva, 2019c).

We tested several machine learning algorithms and the BERT, SciBERT and BioBERT model fine-tuned for the relation extraction task on the annotated corpus. The details on the corpus and the method can be found in Koroleva and Paroubek (2019). The best result for this task was achieved by the fine-tuned BioBERT model.

6 Interface

Our prototype system allows the user to load a text (with or without annotations), run algorithms, visualize their output, correct, add or remove annotations. The expected input is an article reporting an RCT in the text format, including the abstract.

Figure 1 shows the interface with an example of a processed text.

The main items of the drop-down menu on the top of the page are **Annotations**, allowing to visualize and manage the annotations, and **Algorithms**, allowing to run the described algorithms to detect potential spin and the related information. The text fragments identified by the algorithms can be highlighted in the text. When running the algorithms, a report is generated that contains the extracted information and its analysis by the tool (e.g. a mismatch between the outcomes in the text and in the trial registry; absence of the declared primary outcome among the reported outcomes in the abstract). The report is saved into the Metadata section of Annotations menu, which can be accessed through the interface, and can be exported to a file via the **Generate report** item of the Algorithms menu. Human experts can use this report to check the extracted information and the analysis performed by the tool, and to make a final decision on the presence/absence of a given type of spin.

7 Results and conclusions

The current functionality, methods in use, annotated datasets and the best achieved results are outlined in Table 1. Performance is assessed per-token for outcome and significance level extraction and per-unit for other tasks.

In this paper, we presented a first prototype tool for assisting authors and reviewers to detect spin and related information in abstracts of articles reporting RCTs. The employed algorithms show operational performance in complex semantic tasks, even with relatively low volume of available annotated data. We envisage two possible applications of our system: as an authoring aid or as peer-reviewing tool. The authoring aid version can be further developed into an educational tool, explaining the notion of spin and its types to the user.

Possible directions for future work include: improving the implementation and interface (adding prompts for interaction with the user; facilitating installation process), algorithms (improving current performance, adding detection of new spin

o fine tuned BERT models outperform others in all settings. Best fine-tuned model is BioBERT.

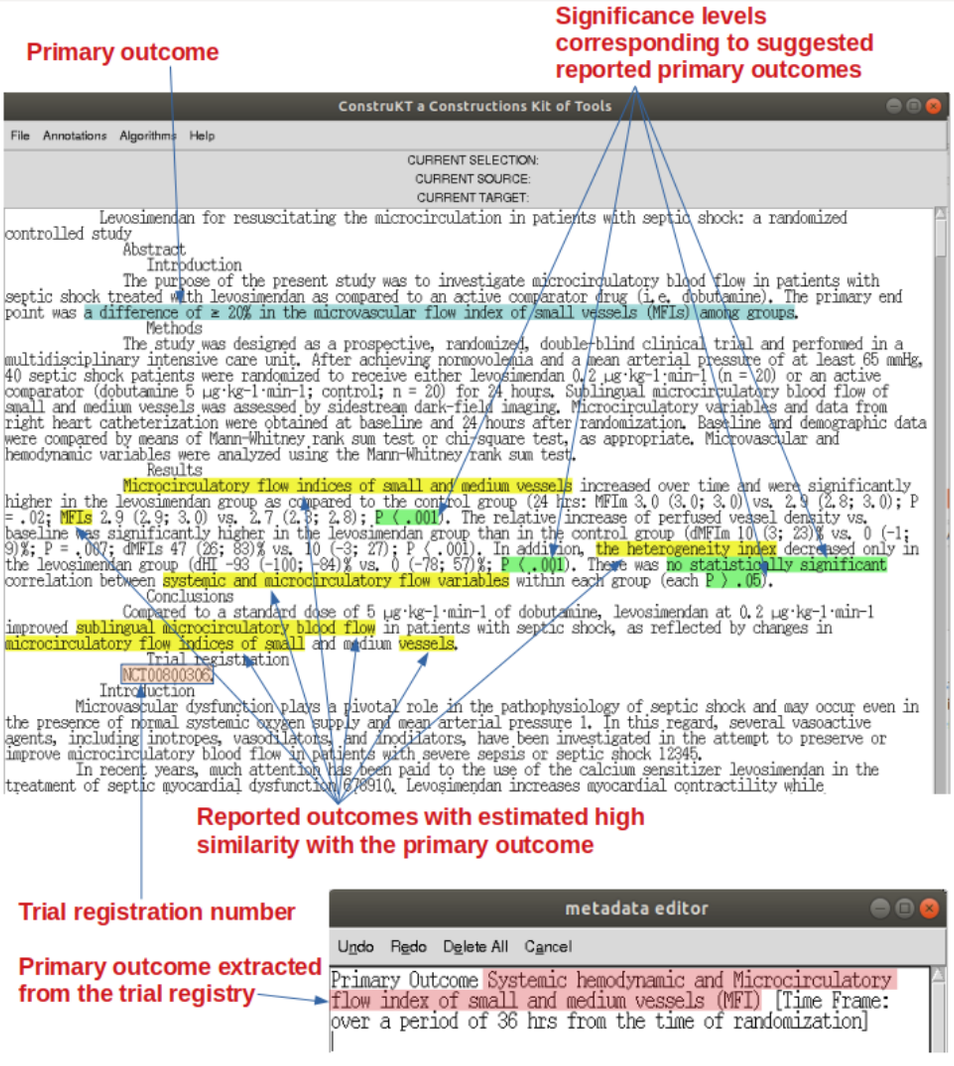

Figure 1: Example of a processed text

types), application (promoting the tool among the target audience; encouraging users to submit their manually annotated data, to be used to improve the algorithms), and optimization (parallel processing of multiple input text files). Our system can be easily incorporated into other text processing tools.

Another interesting yet challenging direction for the future work is detecting spin/distorted reporting in texts belonging to scientific domains other than biomedicine. First of all, a qualitative study of spin is needed to define and classify spin in each scientific domain (similar to the work of Boutron et al. (2010) and Lazarus et al. (2015) for clinical trials). To our best knowledge, there have been no attempts to conduct such a study for non-biomedical texts. It is therefore difficult to hypothesise whether spin-detection algorithms developed for texts reporting clinical trials could be applicable for other domains. It appears that the definition and the types of spin are domain-specific (e.g. outcome-related types of spin, prevalent in the biomedical domain, would not be relevant in domains that do not use the notion of outcome). Hence, we suppose that spin-detection algorithms are domain-specific as well and cannot be applied to other domains.

8 Availability

The proposed prototype tool and associated models are available at: https://github.com/aakorolyova/DeSpin.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 676207.

References

Sophia Ananiadou, Brian Rea, Naoaki Okazaki, Rob Procter, and James Thomas. 2009. Supporting systematic reviews using text mining. *Social Science Computer Review - SOC SCI COMPUT REV*, 27:509–523.

Jennifer Austin, Christopher Smith, Kavita Natarajan, Mousumi Som, Cole Wayant, and Matt Vassar. 2018. Evaluation of spin within abstracts in obesity randomized clinical trials: A cross-sectional review: Spin in obesity clinical trials. *Clinical Obesity*, 9:e12292.

Caroline Barnes, Isabelle Boutron, Bruno Giraudeau, Raphael Porcher, Douglas Altman, and Philippe Ravaud. 2015. Impact of an online writing aid tool for writing a randomized trial report: The cobweb (consort-based web tool) randomized controlled trial. *BMC medicine*, 13:221.

Lian Beijers, Bertus F. Jeronimus, Erick H. Turner, Peter de Jonge, and Annelieke M. Roest. 2017. Spin in rcts of anxiety medication with a positive primary outcome: a comparison of concerns expressed by the us fda and in the published literature. *BMJ Open*, 7(3).

Iz Beltagy, Arman Cohan, and Kyle Lo. 2019. Scibert: Pretrained contextualized embeddings for scientific text. *arXiv preprint arXiv:1903.10676*.

Isabelle Boutron, Douglas Altman, Sally Hopewell, Francisco Vera-Badillo, Ian Tannock, and Philippe Ravaud. 2014. Impact of spin in the abstracts of articles reporting results of randomized controlled trials in the field of cancer: the SPIIN randomized controlled trial. *Journal of Clinical Oncology*.

Isabelle Boutron, Susan Dutton, Philippe Ravaud, and Douglas Altman. 2010. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *JAMA*.

Craig M. Cooper, Harrison M. Gray, Andrew E. Ross, Tom A. Hamilton, Jaye B. Downs, Cole Wayant, and Matt Vassar. 2018. Evaluation of spin in the abstracts of otolaryngology randomized controlled trials: Spin found in majority of clinical trials. *The Laryngoscope*.

Franck Dernoncourt and Ji Young Lee. 2017. PubMed 200k RCT: a dataset for sequential sentence classification in medical abstracts. In *8th IJCNLP (Volume 2: Short Papers)*, pages 308–313, Taipei, Taiwan. Asian Federation of NLP.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Padhraig S. Fleming. 2016. Evidence of spin in clinical trials in the surgical literature. *Ann Transl Med.*, 4,19(385).

Ben Goldacre, Henry Drysdale, Aaron Dale, Ioan Milosevic, Eirion Slade, Philip Hartley, Cicely Marston, Anna Powell-Smith, Carl Heneghan, and Kamal R. Mahtani. 2019. Compare: a prospective cohort study correcting and monitoring 58 misreported trials in real time. *Trials*, 20(1):118.

Romana Haneef, Clement Lazarus, Philippe Ravaud, Amelie Yavchitz, and Isabelle Boutron. 2015. Interpretation of results of studies evaluating an intervention highlighted in google health news: a cross-sectional study of news. *PLoS ONE*.

Julian P. Higgins and Sally Green, editors. 2008. *Cochrane handbook for systematic reviews of interventions*. Wiley & Sons Ltd., West Sussex.

Minlie Huang, Aurélie Névéol, and Zhiyong Lu. 2011. Recommending mesh terms for annotating biomedical articles. *Journal of the American Medical Informatics Association : JAMIA*, 18:660–7.

John Ioannidis. 2005. Why most published research findings are false. *PLoS medicine*, 2:e124.

Muhammad Khan, Noman Lateef, Tariq Siddiqi, Karim Abdur Rehman, Saed Alnaimat, Safi Khan, Haris Riaz, M Hassan Murad, John Mandrola, Rami Doukky, and Richard Krasuski. 2019. Level and prevalence of spin in published cardiovascular randomized clinical trial reports with statistically nonsignificant primary outcomes: A systematic review. *JAMA Network Open*, 2:e192622.

N.C. Kinder, M.D. Weaver, Cole Wayant, and Matt Vassar. 2018. Presence of 'spin' in the abstracts and titles of anaesthesiology randomised controlled trials. *British Journal of Anaesthesia*, 122.

Svetlana Kiritchenko, Berry de Bruijn, Simona Carini, Joel D. Martin, and Ida Sim. 2010. Exact: automatic extraction of clinical trial characteristics from journal publications. *BMC Med Inform Decis Mak.*

Anna Koroleva. 2019a. MiRoR11 - P2 - Annotated corpus for primary and reported outcomes extraction.

Anna Koroleva. 2019b. MiRoR11 - P2 - Annotated corpus for semantic similarity of clinical trial outcomes.

Anna Koroleva. 2019c. MiRoR11 - P2 - Annotated corpus for the relation between reported outcomes and their significance levels.

Anna Koroleva. 2020. MiRoR11 - P2 - Annotated dataset for spin-related types of statements (statements of similarity and within-group comparisons).

Anna Koroleva, Sanjay Kamath, and Patrick Paroubek. 2019. Measuring semantic similarity of clinical trial outcomes using deep pre-trained language representations. *Journal of Biomedical Informatics: X*, 4:100058.

Anna Koroleva, Sanjay Kamath, and Patrick Paroubek. EasyChair, 2020. Extracting outcomes from articles reporting randomized controlled trials using pre-trained deep language representations. EasyChair Preprint no. 2940.

Anna Koroleva and Patrick Paroubek. 2019. Extracting relations between outcomes and significance levels in randomized controlled trials (RCTs) publications. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 359–369, Florence, Italy. Association for Computational Linguistics.

Clément Lazarus, Romana Haneef, Philippe Ravaud, and Isabelle Boutron. 2015. Classification and prevalence of spin in abstracts of non-randomized studies evaluating an intervention. *BMC Med Res Methodol.*

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*.

Suzanne Lockyer, Robert Willard Hodgson, Jo C. Dumville, and Nicky Cullum. 2013. "Spin" in wound care research: the reporting and interpretation of randomized controlled trials with statistically non-significant primary outcome results or unspecified primary outcomes. In *Trials*.

Iain Marshall, Joël Kuiper, Edward Banner, and Byron C. Wallace. 2017. Automating biomedical evidence synthesis: RobotReviewer. In *Proceedings of ACL 2017, System Demonstrations*, pages 7–12, Vancouver, Canada. Association for Computational Linguistics.

Iain James Marshall, Joël Kuiper, and Byron C. Wallace. 2015. Robotreviewer: Evaluation of a system for automatically assessing bias in clinical trials. *Journal of the American Medical Informatics Association : JAMIA*, 23.

James Mork, Alan Aronson, and Dina Demner-Fushman. 2017. 12 years on – is the NLM medical text indexer still useful and relevant? *Journal of Biomedical Semantics*, 8(1).

James G. Mork, Antonio Jimeno-Yepes, and Alan R. Aronson. 2013. The NLM medical text indexer system for indexing biomedical literature. *CEUR Workshop Proceedings*, 1094.

Zainab Samaan, Lawrence Mbuagbaw, Daisy Kosa, Victoria Borg Debono, Rejane Dillenburg, Shiyuan Zhang, Vincent Fruci, Brittany Dennis, Monica Bawor, and Lehana Thabane. 2013. A systematic scoping review of adherence to reporting guidelines in health care literature. *Journal of multidisciplinary healthcare*, 6:169–88.

Kenneth F. Schulz, Douglas G. Altman, and David Moher. 2010. Consort 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ*, 340.

Frank Soboczenski, Thomas Trikalinos, Joël Kuiper, Randolph G. Bias, Byron Wallace, and Iain J. Marshall. 2019. Machine learning to help researchers evaluate biases in clinical trials: A prospective, randomized user study. *BMC Medical Informatics and Decision Making*, 19.

Francisco E. Vera-Badillo, Marc Napoleone, Monika K. Krzyzanowska, Shabbir M.H. Alibhai, An-Wen Chan, Alberto Ocana, Bostjan Seruga, Arnoud J. Templeton, Eitan Amir, and Ian F. Tannock. 2016. Bias in reporting of randomised clinical trials in oncology. *European Journal of Cancer*, 61:29 – 35.

Amélie Yavchitz, Isabelle Boutron, Aida Bafeta, Ibrahim Marroun, Pierre Charles, Jean J. C. Mantz, and Philippe Ravaud. 2012. Misrepresentation of randomized controlled trials in press releases and news coverage: a cohort study. *PLoS Med*.