A Simple Three-Step Approach for the Automatic Detection of Exaggerated Statements in Health Science News

Jasabanta Patro
ETRO, VUB
jasabanta.patro@vub.be

Sabyasachee Baruah
USC, California
sbaruah@usc.edu

Abstract

There is a huge difference between a scientific journal reporting 'wine consumption might be correlated to cancer', and a media outlet publishing 'wine causes cancer' citing the journal's results. The above example is a typical case of a scientific statement being exaggerated as an outcome of the rising problem of *media manipulation*. Given a pair of statements (say one from the source journal article and the other from the news article covering the results published in the journal), is it possible to ascertain with some confidence whether one is an exaggerated version of the other? This paper presents a surprisingly simple yet rational three-step approach that performs best for this task. We solve the task by breaking it into three sub-tasks as follows – (a) given a statement from a scientific paper or press release, we first extract relation phrases (e.g., 'causes' versus 'might be correlated to') connecting the dependent (e.g., 'cancer') and the independent ('wine') variable, (b) classify the strength of the relationship phrase extracted and (c) compare the strengths of the relation phrases extracted from the statements to identify whether one statement contains an exaggerated version of the other, and to what extent. Through rigorous experiments, we demonstrate that our simple approach by far outperforms baseline models that compare state-of-the-art embedding of the statement pairs through a binary classifier or recast the problem as a textual entailment task, which appears to be a very natural choice in this settings.

1 Introduction

Exaggerations in health news can have tremendous adverse effects on the lifestyle of the common masses who feed themselves mostly on such news instead of the source scientific publication. This problem is challenging as it involves many intrinsic complexities that need to be addressed. First, while encountering a scientific claim, we need to identify the true fact related to the claim from the knowledge base (in most cases the source journal article). For instance, if the press release/ news report states that 'chocolate causes acne' then this claim needs to be compared to the scientific study that actually recruits human subjects and does experiments to study connections between chocolate consumption and acne vulgaris. In fact, the study reported in Fulton-Jr. et al. (1969) proves that chocolate consumption is 'not related' to acne.

In this paper, we propose a very simple three-step method[1] that given a pair of statements, e.g., 'chocolate causes acne' taken from the press release and 'chocolate consumption is not related to acne' taken from the source journal, can identify if the former is an exaggerated version of the latter. Note that the problem that we aim to solve requires the pair of statements to be compared as inputs to produce the desired output (exaggerated or not). However, toward the end of the paper, we outline a simple heuristic that dissolves this constraint for the considered dataset that the pair of statements that we compare from the whole article (press release or source journal) needs to be known to us a priori.

We note that comparing state-of-the-art embedding of the two statements using binary classifiers does not give much advantage; even adapting the problem as a textual entailment task (a natural adaptation since one statement is an exaggerated version of the other) surprisingly does not bring much additional benefits.

This paper therefore puts forward *a simple three step approach, breaking the task down to three rational steps*: (1) *given a statement from a scientific paper or press release, extract relation phrases (e.g., 'causes' versus 'correlated to') connecting*

[1] Code: bit.ly/39crP39

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 3293–3305
April 19 - 23, 2021.

the dependent (*e.g., 'acne'*) *and the independent* (*e.g., 'chocolate'*) *variable,* (2) *classify the strength of the relation phrase and* (3) *compare the strengths of the relation phrases extracted from the statements to identify whether one sentence contains an exaggerated version of the other, and to what extent.*

Our approach is operationalised on the data released by Sumner et al. (2014). The exact statements containing the relation phrases within an article or the source journal are annotated in the dataset allowing us to effectively train our models. The main results of our paper are,

- **Extraction of relation phrases**: We experiment with number of syntax driven and sequence labeling approaches to extract the relation phrases. A adaptive version of recently proposed and highly successful BERT-NER (Devlin et al., 2019) performs best with a F1-score of ~ 0.85 for this task.
- **Strength classification**: Next, given a statement with its relation phrase already labeled, we pass it through standard classifiers to learn the *strength* of phrase[2]. We achieve a micro-F1 and macro-F1 of 0.74 and 0.69 respectively for 6-class classification task.
- **Exaggeration detection**: Now given a pair of statements with their strength levels marked and one taken from the source journal while the other from the news article (or press release) we compute the difference in strengths to output whether the latter is an exaggerated version of the former. We obtain a perfect match of exaggeration levels for 0.62 fraction of cases.
- **Additional contribution**: We also identify a mechanism to spot the exact location of the main claim statement in the whole document so that the exact pair is not needed as inputs for our pipeline to work. Our mechanism seem to work for the considered dataset.

2 Related work

Media manipulation is a set of related techniques in which the manipulator attempts to create an image or argument that favors particular interests (Coxall, 2013). Media manipulation and fake news got huge attention from the research community in the current decade. Hundreds of studies got published in this domain which makes it impossible to cite them all. The surveys of different methods and datasets published in these domain can be found in Parikh and Atrey (2018); Zhou and Zafarani (2018); Sharma et al. (2019); Bondielli and Marcelloni (2019); Oshikawa et al. (2018); Haciyakupoglu et al. (2018); Zhou and Zafarani (2020); Van Eemeren et al. (2009). Previous study (Sumner et al., 2014) suggests that press releases are a major source of exaggeration. Identification of relationship between entities and associated strength in scientific articles and comparing them with that of news reports and press releases is the key to exaggeration detection. In literature, there are many techniques present separately to identify entities along with their relationships (Blake, 2010) and strength associated in a relationship (Light et al., 2004; Vlachos and Craven, 2010). The authors in Lim et al. (2016) propose a framework to identify claims from tweet corpus related to major events. In (Giasemidis et al., 2016) the authors build autonomous message classifier that filters relevant and trustworthy information from Twitter. In (Khoo et al., 2000) the authors develop a knowledge extraction and knowledge discovery system that extracts causal knowledge from textual databases.

Present study: None of the studies mentioned above tackle the exaggeration detection problem. In a recent study Li et al. (2017) analyze the same dataset as we use in this paper and attempt to identify exaggeration. They also assume that the pair of statements are available as inputs for the detection task. However, the biggest drawback of this work is that the authors additionally assume that the relationship phrase in the input statement is also known which makes the task significantly simpler. In fact, the authors also do not use these relationship phrases as important signals but treat each statement as a whole as bag-of-words. We, on the other hand, identify the relation phrases automatically in the first place and use it as an important signal for the next two sub-tasks (strength classification and exaggeration detection) which is the most important contribution of our work. We also compare this work with our method and show that we considerably outperform them.

[2]The strength levels are defined as in (Sumner et al., 2014) and certain coarse-grained revisions of the same.

3 Dataset and Preprocessing

3.1 Dataset

We use the publicly available dataset[3] released by Sumner et al. (2014) for our experiments. This dataset contains detailed annotations of 462 journals, corresponding 462 press releases and 668 news articles, issued in 2011 by the Russel Group of Universities (20 leading UK Research Universities) in health related topics. Every individual press release is a follow-up of a journal paper; we assume this journal paper to be the reference for our analysis as it is followed by Sumner et al. (2014). Every press release in turn, is discussed by some news reports. In the dataset, 230 out of 462 journals and press releases have at least one news article coverage.

For each of the three sources, the dataset has detailed annotations of different types. First for each journal/ news article/ press release, the statement containing the main claim is manually identified by the annotators within the running text. Now in this statement, the stretch of relationship phrase is marked in bold. Third, the strength of this relationship phrase is graduated on a Likert-like scale from '0' to '6'. (see section A in supplementary material).

Note that, as per the annotation guidelines, one article can have only one statement with the main claim. The relationship phrase connects the independent variable (*IV*) with the dependent variable (*DV*) in the statement with the main claim. Note that sometimes either the *IV* or the *DV* or both might not be part of the statement with the main claim (i.e., may be present in other parts of the text in the article and connected to statement with the main claim only implicitly). However, since for our work we need the statement with the main claim and the relationship phrases only the above limitation does not pose a hindrance.

While Sumner et al. (2014) provides the dataset already marked by one of the seven quantization levels for each journal, press release and news report, thus allowing for analysis of the exaggerated content, it is difficult to ascertain the robustness of the results obtained. This is primarily because some of the quantization levels seem to be too close (see the 'Description' of strength categories in supplementary material.) and the data set is unbalanced (see Table **??**). In order to test the robustness of the results that we present in the subsequent sections, we also club the above quantizations into more coarse-grained labels. Essentially, we consider a 4-class and a 2-class quantization in addition to the 6-class (omitting class '0'). For the 4-class we map the above seven quantizations as follows: $1 \rightarrow 1$, $(2, 3) \rightarrow 2$, $(4, 5) \rightarrow 3$ and $6 \rightarrow 4$. For the two class we map as follows: $(1, 2, 3, 4, 5) \rightarrow 1$ and $6 \rightarrow 2$.

Strength	#Statements	Fraction
1	69	0.04
2	321	0.2
3	132	0.08
4	159	0.1
5	108	0.07
6	812	0.51

Table 1: Distribution of statements across the different strength categories.

3.2 Data preprocessing

We discard statements with the main claim in which the relationship phrases are not marked by the annotators; these include the statements from category '0'[4] and a very few statements from the other categories. At the end of this process, we have a set of 1601 statements with annotated relation phrases. The distribution of statements across the different strength categories is shown in Table **??**. The table shows that the categories '6' and '2' are the dominant ones. We divide the 1601 statements (together from journals, press releases and news reports) into training, development and test sets. We keep 1000 statements in the training set, 300 statements in the development set and 301 statements in the test set. Note that from the 301 statements in the test set, we can construct a total of 316 distinct statement pairs each formed from the comparison of a source journal and a corresponding news paper article/press release. While for the first two sub-tasks (i.e., relation phrase labeling and strength classification) we need the individual statements as input, in the last sub-task (i.e., exaggeration identification) we need pairs of statements for final comparison and prediction.

4 Methodology

As discussed earlier, our approach has three steps. They are (i) relation phrase labeling, (ii) strength

[3]Dataset: https://bit.ly/2qc86tk

[4]Note that category '0' refers to cases where there is no relation between the IV and DV and is therefore not useful for further processing.

classification, and (iii) exaggeration identification. All of the three modules are connected in a sequence as shown in Figure 1. The relation phrase labeling module takes a selected statement and labels the relation phrase present in it. Next, the strength classification module takes a relation phrase as input and predicts its strength level. Finally, in the exaggeration identification module, the strengths of source statement and target statement are compared, and it calculates the exaggeration or underplay level of the target statement with respect to the source statement. The individual modules are described in details in the subsequent subsections.

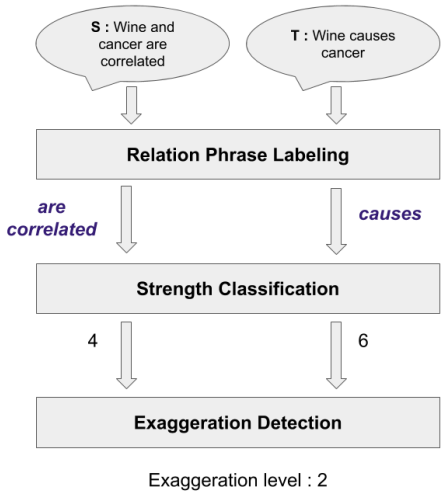

Figure 1: Flow diagram of the methodology. S: source statement, T: target statement.

4.1 Relation phrase labeling

As previously stated, the primary objective of this module is to identify the relation phrase describing the relation between the independent and the dependent variables. This problem is more difficult than the traditional entity-relation extraction problem as multiple entity pairs connected by relation phrases can be present in a statement. Therefore, finding relation phrase that denotes the main claim is a challenging task. We broadly employ two different types of approaches for relation labeling – (i) syntax driven approaches, and (ii) sequence labeling approaches. We describe each of them in next subsections.

4.1.1 Syntax driven approaches

The key idea: In syntax driven approaches, we rely on the fact that the position of the relation phrases in a statement is syntax driven. In particular, we mine the structural patterns in the dependency tree of the statements. In addition, we use heuristics over the state-of-the-art entity relation extraction tool (Angeli et al., 2015) to get relation phrases in the causal statements.

Dependency tree heuristics (DTH): The intuition behind this method is that relation phrases have characteristic syntactic (part-of-speech, lemmas, dependency edges etc.) and semantic features. We intend to leverage these in the dependency tree representation of statements, by collapsing unimportant relations and selecting the rightful node through various heuristics. The phrase corresponding to the selected node is identified as the relation phrase. The details of the steps we follow for the construction of the collapsed dependency tree and the extraction of the relation phrase from the causal statement are presented in supplementary material.

OpenIE heuristics: We use Stanford open information extraction (OpenIE) tool for the identification of entity-relation triplets. Let $t_1, t_2, ...t_n$ be the n triplets obtained for a statement s, where each t_i consists of two entities at its end connected by a relation phrase. We choose t_i having the largest and smallest phrase and consider these as the representative relation phrase for the statement s. The rationale behind choosing the largest phrase is to increase the probability of including the original relation phrase. The smallest phrase is chosen to show that smaller sized phrases always perform worse than the largest phrase.

4.1.2 Sequence labeling approaches

The key idea: The central idea these approaches put forward is use the training data to create a model to label the relation phrase in the input statement. The relation phrases are marked as per traditional BIO[5] encoding format. In addition, the BERT scheme that we shall use has three other labels – X for added morphological inflation, CLS for sentence beginning and SEP for sentence separations. Based on this annotated training data the sequence labeler is tasked to learn the beginning, the stretch and the end of the relationship phrase. Once this is marked for each statement the labeled phrases are passed on to the next phase of the pipeline for strength classification.

LSTM-CRF:We use the LSTM-CRF architecture

[5] https://en.wikipedia.org/wiki/Inside%E2%80%93outside%E2%80%93beginning_(tagging)

Features	Dimension
POS tag	40
POS bigram	40
Wordnet cluster	40

Table 2: Additional features at CRF layer.

similar to Lample et al. (2016) in addition to a set of novel features as described in Table **??** in the CRF layer[6] along with the hidden state vector of the BiLSTM layer. The combination is done by concatenating the feature vectors with the hidden state vector. The input layers to the model are vector representations of the individual words or characters.

LSTM-CNN-CRF: We use the architecture proposed by Ma and Hovy (2016) and adapt for our purpose[7].

LM-LSTM-CRF: Here we use the task-aware neural sequence labeling model proposed by (Liu et al., 2017) and adapt it for our purpose.

BERT-SL: We use the variant BERT (Devlin et al., 2018) technology originally proposed for solving the NER task[8], which is essentially modeled as a sequence labeling problem. We suitably adapt the BERT-NER framework to extract the relation phrases. BERT's model architecture is a multilayer bidirectional transformer encoder based on the original implementation described in (Vaswani et al., 2017)[9]. In particular, we use the BERT base architecture that has 12 transformer layers, hidden vector size of 768 and 12 self-attention heads one corresponding to each transformer module. For fine-tuning, the final hidden representation of each token is fed to a classification layer over the NER label set. For our purpose we train the BERT-NER model using our training data and the label set corresponds to the relationship phrase beginning and end markers instead of the NER labels.

4.1.3 Evaluation

We use the standard token level F1-score and accuracy to compare the different labeling approaches.

[6]In our experiments we have seen that increasing dimension size further improves the results further. POS tags are obtained using NLTK.

[7]All other parameters remaining same, we use LSTM state size of 100, dropout rate of 0.5 and input batch size of 5.

[8]BERT NER: https://github.com/kyzhouhzau/BERT-NER

[9]http://nlp.seas.harvard.edu/2018/04/03/attention.htm

4.2 Strength classification

The primary objective of this module is to predict the strength level of the relationship phrase in the input statement, e.g., in the statement 'wine **causes** cancer' the strength of statement is 6. We feed the standard multi-class classifiers with the annotated relation phrases for each statement in the training set while the training label is the strength of the statement. Since BERT-SL performs best among all relation phrase labeling approaches, we present all our subsequent results for this case only. We obtain the best parameters for each of the classifiers using the validation set. Finally, we report our results on the test set for 6-class, 4-class as well as 2-class scenarios. Unlike prior work (Li et al., 2017) that uses bag of words feature drawn from the whole statement, our model uses the features drawn from the annotated relation phrase, i.e., a part of the statement only. Note that, the results mentioned in the prior work is not directly comparable as the authors assume in their model that the relation phrases are already known, but we extract these phrases in the first place. However, we re-implemented their model (Li et al., 2017) and report the results that we got for our train-test division of the dataset.

Evaluation: We compare the strength classification methods using the micro-F1, which in this case, is equal to the accuracy. This is, as usual, calculated as the fraction of statements that have a correctly classified strength level out of the total number of statements. We also report the macro-F1 obtained for each model for the better understanding of the performance of the models primarily since the classes are unbalanced. This is calculated as the average of F1 scores obtained for each class as predicted by each model for the test set.

4.3 Exaggeration identification

In this module we calculate the strength level difference between a source statement and a target statement. The source statement is the one with the main claim from the journal article, while the target statement is the one with the main claim from corresponding news reports or press release covering the source. We calculate exaggeration (or underplay) levels for each classifier result over each type of strength classification.

Baseline: We present two types of baseline here. In the first, we pass the BERT embedding of the two statements (using respective 'CLS' markers)

to the standard classifiers and attempt to estimate if there is a component of exaggeration. In the second, we model the problem as a textual entailment task which is a natural choice. Given a pair of statements the goal is to predict whether the second statement is an exaggerated version of first. For this purpose, we have used the BERT text classifier module. The final hidden vector of BERT model is passed through a softmax layer for the classification.

Evaluation: We employ *fraction of perfect match* (PM), which in our case, is also equal to the accuracy, as the measure to compare the different exaggeration detection methods. This is calculated as the fraction of source-target statement pairs for which we get correct exaggeration or underplay level as we have in the ground-truth. We also report the mean square error (MSE) for each model which is calculated as the average of the squares of the differences between actual and predicted strength difference for every pair of statements in each model.

5 Results

In this section, we present the results from the three phases – relation phrase labeling, strength classification and exaggeration identification one after the other. Finally we outline a scheme to spot the exact location of main claims in the news articles and press releases that particularly works for this dataset. This enabled us to fully automate the process based on this scheme and measure its performance.

5.1 Relation phrase labeling

We note down the main results in Table **??**. The table shows that BERT-SL outperforms all other approaches by a large margin. The F1-score and the accuracy for this method is 0.85 and 0.95, respectively. Among the others LSTM-CRF augmented with PoS tags and PoS bigrams is the most competitive with a F1-score of 0.73 and accuracy of 0.89.

Success of BERT-SL: Table **??** shows the number of perfect matches between the ground-truth relationship phrase and the relationship phrase extracted using BERT-SL and LSTM-CRF. For all n-gram relationship phrases BERT-SL achieves a much higher number of perfect matches with the ground-truth. The difference in the number of matches obtained from the two methods is particularly large for unigrams and bigrams which covers the bulk of the statements. This is the reason why BERT-SL is able to outperform the other methods by a large margin.

Model	F1-score	Accuracy
DTH	0.58	0.82
OpenIE (large)	0.48	0.76
OpenIE (small)	0.37	0.70
LSTM-CRF (RI (Li et al., 2017))	0.67	0.88
LSTM-CRF (RI) + PoS	0.67	0.87
LSTM-CRF (GloVe) + PoS	0.72	0.88
LSTM-CRF (GloVe) + PoS+ PoS BI	0.73	0.89
LSTM-CRF (GloVe) + PoS tag + PoS BI + WC	0.73	0.88
LSTM-CNN-CRF (GloVe)	0.56	0.83
LM-LSTM-CRF (GloVe)	0.52	0.80
BERT-SL	**0.85**	**0.95**

Table 3: Comparison of the baselines with our approach. RI: Random Initialization, PoS: Part of speech tags, PoS BI: PoS Bigrams, WC: Wordnet clusters

#words	#BERT-SL	#LSTM-CRF
1	**33**	12
2	**34**	28
3	**28**	24
4	**9**	8
5	**6**	2
6	**3**	0

Table 4: Number of perfect matches for the competing models as per the number of words present in the relation phrase.

5.2 Strength classification

We predict the strength of a statement based on its relation phrase. For this purpose we build classifiers that take bag of words vector of the relationship phrase as identified by BERT-SL and predict its strength. The bag of words vector is created from the BERT embeddings as follows. The BERT embeddings of each word in the relation phrase are concatenated to form a single vector which is fed to the classifier. However the number of words in each relation phrase could be different. We fix this length by imagining all relation phrases to have length equal to the length of the largest relation phrase across the 1601 data points. This makes the concatenated vector size same for all the relation phrases; the missing entries in each vector so created are replaced by zero.

We use various classification models such as multinomial naive-bays (MNB), random-forest (RF), stochastic gradient descent (SGD) and XG-

Test type	Classifier	$\mu F1_{6class}$	$MF1_{6class}$	$\mu F1_{4class}$	$MF1_{4class}$	$\mu F1_{2class}$	$MF1_{2class}$
True rel. phs. (Trn:1300, Tst: 301)	MNB	0.68	0.52	0.77	0.62	0.79	0.79
	RF	0.5	0.16	0.51	0.22	0.54	0.48
	SGD	**0.76**	**0.70**	**0.79**	**0.76**	**0.82**	**0.82**
	XGB	0.74	0.69	0.76	0.73	0.76	0.76
BERT-SL rel. phs. (Trn: 1300, Tst: 301)	MNB	0.64	0.45	0.77	0.62	0.76	0.76
	RF	0.51	0.16	0.51	0.22	0.55	0.50
	SGD	**0.74**	**0.69**	**0.79**	**0.76**	**0.79**	**0.79**
	XGB	0.71	0.64	0.76	0.73	0.74	0.74
Baseline (Li et al., 2017) (Trn: 1300, Tst: 301)	BOW (unigram+bigram)	–	–	0.64	0.62	–	–

Table 5: Strength classification results. $\mu F1_{xclass}$: micro-F1 for x (6/4/2) class, $MF1_{xclass}$: macro-F1 for x (6/4/2) class.

Boost (XGB) for this purpose. The training of these classifiers is done on actual relation phrase and strength class pairs as annotated in the dataset. Table **??** shows the micro-F1 and macro-F1 scores of various classifiers across different strength classes (6, 4 and 2). In the table, entries for the true relation phrases represent the obtained micro-F1 and macro-F1 scores corresponding to a test set where the ground-truth annotated relation phrase (instead of what is obtained from BERT-SL) is taken into account. It is interesting to note that for all the three strength classes strengths obtain from BERT labeled relation phrases reaches close to the F1 values as one would have obtained if the true test relationship phrases were supplied to the classifiers at the input. Finally, our model by far outperforms the only known baseline (Li et al., 2017) in both micro and macro-F1 scores. To be fair, we show results for the 4-class which is the only case reported in (Li et al., 2017).

5.3 Exaggeration level identification

We identify the exaggeration level by taking the strength difference between the source statement from the journal and the target statement from the news article/ press release. The number of such source-target statement pairs in our test data is 316. The strength for each statement is obtained from the classifiers reported in the previous section. We present the results of exaggeration detection in Table **??**. Our three step approach attains a perfect match score for 0.62, 0.68 and 0.69 fraction of cases in the 6, 4 and 2 classes respectively (best classifier outputs) which is again close to what one could have obtained if the ground-truth relation phrases were known to the strength classifier in the previous stage.

Baseline: For the baseline we take the entire training set and construct as many possible source-target statement pairs where, by definition, the source statement is from a journal and the target statement is from a corresponding news article/press release. From the 1000 training statements, we could construct a total of 1298 such source-target pairs. If we feed the direct BERT embedding of the source and the target statement to the set of classifiers (RF, SGD and XGB in Table **??**) the results are much worse than our approach. Further if we adapt the BERT text classifier (as used in textual entailment detection module) to solve our problem ('Neural' in Table **??**), there is no benefit obtained. This proves that our proposed concept, a simple yet more interpretable three step sequential approach performs better than all other approaches where the whole sentence is taken into account for the exaggeration detection.

6 Discussion

In this section we present the findings of our error analysis and results correspond to the special sub cases where simple heuristics is taken into account to automate the whole process.

6.1 Error Analysis:

The error cases that we believe could be the possible reasons affecting the overall performance of the system are presented below.

- **Ambiguous modifiers**: Some statements have certain modifiers like 'little', 'any' etc in their predicted relation phrases, which leads to an error in the strength class prediction.
- **Incorrect relation phrase**: In some cases, the predicted relation phrase is incorrect which leads to propagation of errors in next steps.
- **Incorrect relation interpretation**: The system is unable to differentiate between cases like "could be entirely independent" and

Test type	Classifier	PM_{6class}	MSE_{6class}	PM_{4class}	MSE_{4class}	PM_{2class}	MSE_{2class}
True rel. phs. (Tst: 301, Pairs: 316)	MNB	0.53	5.71	**0.68**	1.27	**0.69**	0.37
	RF	0.49	**3.93**	0.51	**1.18**	0.52	0.53
	SGD	0.62	5.01	0.67	1.30	0.68	**0.35**
	XGB	**0.63**	4.65	0.63	1.41	0.65	**0.35**
BERT-SL rel. phs. (Tst: 301, Pairs: 316)	MNB	0.52	4.99	**0.68**	1.27	0.67	0.40
	RF	0.49	**3.88**	0.51	**1.18**	0.53	0.58
	SGD	**0.62**	4.74	0.67	1.30	**0.69**	**0.37**
	XGB	0.6	4.73	0.63	1.41	0.64	0.39
Baseline (Train: 1298, Pairs: 316)	RF	**0.48**	**4.24**	0.49	1.37	**0.61**	0.47
	SGD	0.43	5.13	0.33	1.39	0.49	0.54
	XGB	0.47	4.51	**0.51**	**1.28**	0.59	**0.46**
	Neural	0.44	–	0.50	–	0.56	–

Table 6: Exaggeration level identification results. PM_{xclass}: Perfect match for x (6/4/2) class, MSE_{xclass}: Mean squared error for x (6/4/2) class.

"could be entirely dependent", which leads to the prediction of wrong strength classes.

We present examples of above cases in the supplementary material.

6.2 Spotting the causal statement:

We manually inspect the position of main claims in the documents. We find that in 95% of cases the main claim is present either in the title or in any one of the first three sentences of the document. We repeat the relationship phrase identification experiment using the best model (i.e., BERT-SL) for a test set that is built using all the four sentences (i.e., the title, the first, the second and the third sentences) taken from each press release and news article present in our dataset. Now we select that sentence for which we get the maximum number of words tagged as a relation phrase. Using this heuristic selection criteria we obtain F1-score of **0.8** and an accuracy of **0.9** which is close to what we report in Table **??**. More detail of this experiment are presented in supplementary material.

6.3 Inclusion of IV and DV

Our model can also be used to extract the IV and the DV as well. We also check if this brings additional benefits to the pipeline. The details of this experiment is given in the supplementary material. The F1-score (precision, recall) we obtained for the relation phrase labeling task in this case is 0.72 (0.75, 0.7). This experiment also gives token level accuracy of 0.87. As we observe, the inclusion of the IV and DV information does not improve the labeling performance and therefore we did not use it for the next stages of the pipeline. The use of these additional information actually seems to confuse the labeler more than benefiting it. However, the automatic extraction of the IV and DV can be useful for developing other applications in future.

7 Conclusion

In this work, we proposed a simple yet explainable three step approach that automatically identifies whether a given statement typically from a press release or a news article is exaggerated in comparison to the source statement present in the journal. Our first step adapts the recently proposed BERT technology and models the relationship extraction problem as a sequence labeling task. This beats other sequence labeling and syntax driven approaches by a large margin. The relationship phrase extracted is encoded as a bag of words and fed to standard classifiers to obtain the strength of the phrase. Ground-truth labels of relationship phrases and those obtained from our model achieve similar performance. Finally, once the strength of the relationship phrase is available, a pair of statements can be easily compared. This method of exaggeration identification beats standard baselines that directly feed the two individual BERT embedding of the statement pair into binary classifiers or use the BERT textual entailment framework. In future, one can check if this kind of approaches work for other tasks where whole task can be divided into explainable subtasks.

Acknowledgments

This work is carried out by the authors during their study period at IIT Kharagpur. We thank Dr. Animesh Mukherjee and Dr. Pawan Goyal for their valuable inputs during the development of the work.

References

Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D Manning. 2015. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 344–354.

Catherine Blake. 2010. Beyond genes, proteins, and abstracts: Identifying scientific claims from full-text biomedical articles. *Journal of biomedical informatics*, 43 2:173–89.

Alessandro Bondielli and Francesco Marcelloni. 2019. A survey on fake news and rumour detection techniques. *Information Sciences*, 497:38–55.

Malcolm Coxall. 2013. *Human Manipulation-A Handbook*. Malcolm Coxall-Cornelio Books.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL*. Association for Computational Linguistics.

James E. Fulton-Jr., Gerd Plewig, and Albert M. Kligman. 1969. Effect of chocolate on acne vulgaris. *JAMA*, 210(11):2071–2074.

Georgios Giasemidis, Colin Singleton, Ioannis Agrafiotis, Jason R. C. Nurse, Alan Pilgrim, Chris Willis, and Danica Vukadinovic Greetham. 2016. Determining the veracity of rumours on twitter. In *SocInfo*.

Gulizar Haciyakupoglu, Jennifer Yang Hui, VS Suguna, Dymples Leong, and Muhammad Faizal Bin Abdul Rahman. 2018. Countering fake news: A survey of recent global initiatives.

Christopher S. G. Khoo, Syin Chan, and Yun Niu. 2000. Extracting causal knowledge from a medical database using graphical patterns. In *ACL*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Yingya Li, Jieke Zhang, and Bei Yu. 2017. An nlp analysis of exaggerated claims in science news. In *NLPmJ@EMNLP*.

Marc Light, Xin Ying Qiu, and Padmini Srinivasan. 2004. The language of bioscience: Facts, speculations, and statements in between. In *HLT-NAACL*.

Wee-Yong Lim, Mong-Li Lee, and Wynne Hsu. 2016. Claimfinder: A framework for identifying claims in microblogs. In *#Microposts*.

Liyuan Liu, Jingbo Shang, Frank Xu, Xiang Ren, Huan Gui, Jian Peng, and Jiawei Han. 2017. Empower sequence labeling with task-aware neural language model. *arXiv preprint arXiv:1709.04109*.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1064–1074.

Ray Oshikawa, Jing Qian, and William Yang Wang. 2018. A survey on natural language processing for fake news detection. *arXiv preprint arXiv:1811.00770*.

Shivam B Parikh and Pradeep K Atrey. 2018. Media-rich fake news detection: A survey. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 436–441. IEEE.

Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu. 2019. Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(3):1–42.

Petroc Sumner, Solveiga Vivian-Griffiths, Jacky Boivin, Andy Williams, Christos A Venetis, Aimée Davies, Jack Ogden, Leanne Whelan, Bethan Hughes, Bethan Dalton, et al. 2014. The association between exaggeration in health related science news and academic press releases: retrospective observational study. *Bmj*, 349:g7015.

Frans H Van Eemeren, Bart Garssen, and Bert Meuffels. 2009. *Fallacies and judgments of reasonableness: Empirical research concerning the pragma-dialectical discussion rules*, volume 16. Springer Science & Business Media.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Andreas Vlachos and Mark Craven. 2010. Detecting speculative language using syntactic dependencies and logistic regression. In *CoNLL Shared Task*.

Xinyi Zhou and Reza Zafarani. 2018. Fake news: A survey of research, detection methods, and opportunities. *arXiv preprint arXiv:1812.00315*.

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*.

Strength category	Description	Example
0	No relationship is mentioned	...An international study of 220,000 people has challenged the idea that obese people who have an ”apple shape” - fat around the middle section of the body - are at higher risk of heart attacks and strokes...
1	Explicitly stating there is no relationship	Clinical officers and doctors **did not differ significantly** in key outcomes for Caesarean section significantly..
2	Statement of correlation - IV and DV are associated, but causation cannot be explicitly asserted	the greatest excess risk **associated** with subsequent primary neoplasms at older than 40 years was for digestive and genitourinary neoplasms
3	Ambiguous statement of relationship - It is unclear what the strength of relationship of this statement	...**has linked** eight new DNA variants to the autoimmune disease...
4	Conditional statement of causation - Causal statements show that the IV directly changes the DV. Conditional causal statements carry an element of doubt in them	...one in four patients **may be wrongly** diagnosed with high blood pressure...
5	Statement of “can” - The word “can” is unique as a statement of relationship in that it implies that the IV has the potential to directly change the DV, Therefore it is stronger than any conditional statement of causation.	...An intensive diet intervention soon after diagnosis **can improve** glycaemic control. The addition of an activity intervention conferred no additional benefit...
6	Statement of causation - The strongest are the statements of causation. This statement says that the IV definitely and directly alters the DV.	...capsules containing concentrated phytonnutrients **improved** clinical outcomes...
IV: Independent Variable	DV: Dependent Variable	

Table 7: Description of strength categories with an example.

A Annotated strength levels

The dataset contains manually coded strength levels of main claims from three sources, based on which, authors found that the press release is the main source of exaggeration in health science reports. The main causal claims in journal article, press release and news reports are coded into seven categories with increase in strength of relationship (see Table **??**). Relation phrases in the claim representing the relationship are marked in bold. The relation phrase connects the independent variable (*IV*) with the dependent variable (*DV*).

B Detais of dependency tree heuristics

The steps followed for extraction of relation phrase from causal statement are given as follows:

- We create dependency trees for each statement, with each node corresponding to a word and edges representing the grammatical relation between them. Let D be the set of all possible dependencies. We work with the collapsed version of dependencies, e.g., prepositions are not represented as nodes but collapsed into edges etc. For the statement, ‘gestures improved performance in spatial visualisation problems’, the dependency tree is shown in Figure 2. Here ‘in’ in the statement has been collapsed to the edge *prep_in*.
- We identify a subset $A \subseteq D$ of dependencies that we collapse to merge the connected nodes. The intuition is to collect words that form a coherent phrase inside a single node. The set A has been formed by going through the definition of each grammatical relation from the Stanford dependency manual. We call the nodes of the compact dependency tree as compact nodes. Each compact node is a tree of simple nodes and thus represents a substring. The grammatical relation between two

Order	Heuristic	Definition
1	LONG_VB_NSUBJ_DOBJ	Select compact node with a *verb* root, and *nsubj* and *dobj* out-edges
2	LONG_VB_NSUBJPASS_AGENT	Select compact node with a *verb* root, and *nsubjpass* and *agent* out-edges
3	LONG_VB_NSUBJPASS_PREP_WITH	Select compact node with a *verb* root, and *nsubjpass* and *prep_with* out-edges
4	LONG_VB_NSUBJ_DOBJ_NSUBJPASS	Select compact node with a *verb* root, and having atleast one *nsubj*, *dobj* or *nsubjpass* out-edge
5	LONG_VB	Select compact node whose root is a *verb*
6	LONG_JJ_NSUBJ_XCOMP	Select compact node with an *adjective* root, and *nsubj* and *xcomp* out-edges
7	LONG_JJ	Select compact node with an *adjective* root, and *nsubjpass* and *prep_with* out-edges
8	LONG_NOUN	Select compact node with a *noun* root

Table 8: Dependency tree heuristics.

compact nodes is the grammatical relation between the roots of their corresponding tree.

- The set A consists of the following dependencies - *advmod, amod, appos, aux, auxpass, cop, det, expl, mwe, mark, neg, nn, npadvmod, num, number, pobj, poss, possessive, predet, prt, quantmod* and *vmod*. Note that most relations in A are modifiers. In the previous example (refer to Figure 2) the subtree with nodes 'problems'/NNS, 'spatial'/JJ and 'visualization'/NN are merged together to form a compact node 'spatial visualization problem'/NNS.
- The next step is to select an appropriate compact node. We use heuristic functions, that take a dependency tree as input and output a node (if its condition is met) or *nil*. We apply a sequence of heuristic functions $h1, h2, h3, ...$ to recover the relational phrase (see Table **??** for the list of heuristics). If $h1$ returns a node, we identify its constituent phrase as our relational phrase. Else if $h1$ returns *nil*, we select $h2$. If $h2$ returns a node, we use that else we try $h3$ and so on. The heuristic functions have been hand-coded by observing the statistics of the best compact node (highest normalized lexicalized edit similarity ($NLES$) with the annotated phrase) in the dependency trees of the training set.
  The $NLES$ between two phrases p_1 and p_2 is defined as

$$NLES(p_1, p_2) = 1 - \frac{dist(p_1, p_2)}{min(|p_1|, |p_2|)} \quad (1)$$

where, $|p_1|, |p_2|$ are the number of characters in p_1 and p_2 respectively. $dist$ refers to the standard edit distance where cost of the substitution, the insertion and the deletion are all taken as 1. We also tried jaccard similarity between lexicalized tokens in place of $NLES$ but we same set of heuristic sequence.

The list of heuristic functions is given in Table **??** along with their order of application, which has been found by iterating through all permutations and choosing the one that produces maximum average $NLES$ across all input training sentences. Each heuristic function proceeds according to its definition; if multiple nodes satisfy the heuristic condition, the node with the most number of words is chosen. If none of the nodes satisfy the heuristic condition, the function returns *nil*. For instance, in the previous example (refer to Figure 2), heuristics LONG_VB, LONG_VB_NSUBJ_DOBJ and LONG_NOUN match and return phrases, however, since LONG_VB has the highest priority so the respective phrase 'improved' is returned as the relation phrase. In another example 'Childhood cancer survivors at greater risk in middle age', no heuristics match to its compact tree and hence *nil* is returned as relation phrase.

C Error Analysis

In this section we discuss some of the error cases that we believe could be the possible reasons affecting the overall performance of the system. In future we plan to tackle some of these cases to further improve the system performance. We mention below some of these error cases. The examples of error cases are presented in Table **??**.

Ambiguous modifiers: Some statements have certain modifiers in their predicted relation phrases,

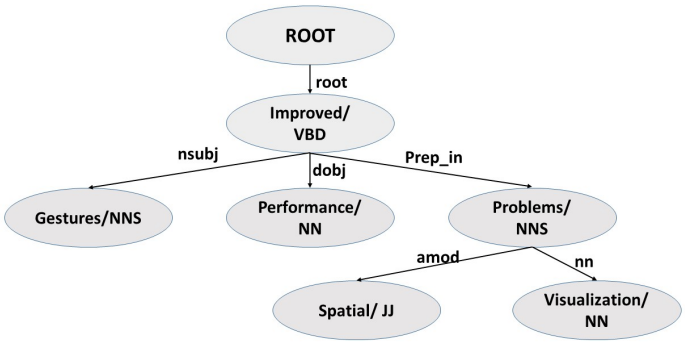

Figure 2: Collapsed dependency tree.

No.	Statement	Ann.Ph.	Pred. Ph.	A_S	P_S
1	a common treatment for a life-threatening heart condition has little significant impact on patient outcomes	has little significant impact on	has little significant impact on	1	6
2	We noted little variation between countries in the rate of maltreatment-related injury admission	little variation	little variation	1	6
3	this study excluded any large increase in the incidence of cases of or deaths from infective endocarditis	excluded any large increase	excluded any large increase	1	6
4	Falling in love sets brain circuits racing in the same way, regardless of sex or sexual orientation	regardless of	sets	1	6
5	breast cancer screening ... does women more harm than good	does	does women more harm than	6	2
6	that the progressive loss of lung function in asthma sufferers could be entirely independent of the effects of inflammation	could be entirely independent of	could be entirely independent	1	4

Table 9: Errors in strength classification. A_S: Actual strength, P_S: Predicted strength

Location	%
Title	26.10
First statement	46.12
Second statement	21.93
Third statement	2.14

Table 10: Location of the statement with the main claim in the press-releases/news articles.

which leads to an error in the strength class prediction. For instance, consider statements 1, 2 and 3 shown in Table **??**. In all these cases the presence of the modifiers such as 'little', 'any' etc. are not separately tackled by the strength classifier.

Incorrect relation phrase: In some cases, the predicted relation phrase is incorrect. For example, consider the statement 4. Here the annotated phrase is 'regardless of' whereas the predicted phrase is 'sets'. Another interesting case is statement 5. Here the predicted phrase fully contains the actual phrase and is also much larger than the actual phrase. The classifier therefore confuses the strength class.

Incorrect relation interpretation: In some cases, like statement 6, the system is unable to differentiate between "could be entirely independent" and "could be entirely dependent", which leads to prediction of class 4 instead of the actual class 1.

D Spotting the causal statement

So far, we have assumed that statement with the main claim is already given to us for labeling and exaggeration identification. However, in order to completely automate the process we need to spot the exact location of these statements so that the previous pipeline of three sub-tasks could be smoothly executed. Manual inspection of the documents indicates that the statement with the main claim is present either in the title or in any one of the first three sentences of the document. In fact, we run an experiment on all the documents in our dataset and observe the location of the main statement. Table **??** shows the percentage of documents in the dataset, in which this statement is present in the title or in the first, second or the third sentence. Surprisingly, in 95% of the documents the causal statement is present in the title or the first three sentences of the document.

We repeat the relationship phrase identification experiment using the best model (i.e., BERT-SL) for a test set that is built using all the four sentences (i.e., the title, the first, the second and the third sentences) taken from each press release and news article present in our dataset. For each press release/news article, all the four sentences are labeled by the sequence labeler. Now we select that sentence for which we get the maximum number of words tagged as a relation phrase. Using this heuristic selection criteria we obtain F1-score of **0.8** and an accuracy of **0.9**. Given this encouraging result, we do not go for any additional algorithmic machinery for separately spotting the statement with the main claim for the dataset. However one can check in future if this works well for other problems and datasets.

E Inclusion of IV and DV

Our model can also be used to extract the IV and the DV as well. We also check if this brings additional benefits to the pipeline. Though the existing annotation of the dataset identifies the entities, i.e., the IVs and DVs associated with each document (journal/press-release/news article), in many cases, they are not explicitly part of the statement with the main claim (i.e., the entities are metaphorically mentioned); therefore, automatic labeling of these variables was impossible without re-annotation. To this purpose, we re-annotated the entire dataset for these two entities. Next we tasked the BERT-SL to identify the IV, DV and relation phrase altogether. The F1-score (precision, recall) we obtained for the labeling task in this case is 0.72 (0.75, 0.7). This experiment also gives token level accuracy of 0.87. As we observe, the inclusion of the IV and DV information does not improve the labeling performance and therefore we did not use it for the next stages of the pipeline. The use of these additional information actually seems to confuse the labeler more than benefiting it. However, the automatic extraction of the IV and DV can be useful for developing other applications in future.