Claim Detection in Biomedical Twitter Posts

Amelie Wührl and **Roman Klinger**
Institut für Maschinelle Sprachverarbeitung, University of Stuttgart, Germany
{amelie.wuehrl,roman.klinger}@ims.uni-stuttgart.de

Abstract

Social media contains unfiltered and unique information, which is potentially of great value, but, in the case of misinformation, can also do great harm. With regards to biomedical topics, false information can be particularly dangerous. Methods of automatic fact-checking and fake news detection address this problem, but have not been applied to the biomedical domain in social media yet. We aim to fill this research gap and annotate a corpus of 1200 tweets for implicit and explicit biomedical claims (the latter also with span annotations for the claim phrase). With this corpus, which we sample to be related to COVID-19, measles, cystic fibrosis, and depression, we develop baseline models which detect tweets that contain a claim automatically. Our analyses reveal that biomedical tweets are densely populated with claims (45 % in a corpus sampled to contain 1200 tweets focused on the domains mentioned above). Baseline classification experiments with embedding-based classifiers and BERT-based transfer learning demonstrate that the detection is challenging, however, shows acceptable performance for the identification of explicit expressions of claims. Implicit claim tweets are more challenging to detect.

1 Introduction

Social media platforms like Twitter contain vast amounts of valuable and novel information, and biomedical aspects are no exception (Correia et al., 2020). Doctors share insights from their everyday life, patients report on their experiences with particular medical conditions and drugs, or they discuss and hypothesize about the potential value of a treatment for a particular disease. This information can be of great value – governmental administrations or pharmaceutical companies can for instance learn about unknown side effects or potentially beneficial off-label use of medications.

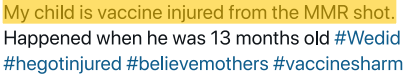

Figure 1: Tweet with a biomedical claim (highlighted).

At the same time, unproven claims or even intentionally spread misinformation might also do great harm. Therefore, contextualizing a social media message and investigating if a statement is debated or can actually be proven with a reference to a reliable resource is important. The task of detecting such claims is essential in argument mining and a prerequisite in further analysis for tasks like fact-checking or hypotheses generation. We show an example of a tweet with a claim in Figure 1.

Claims are widely considered the conclusive and therefore central part of an argument (Lippi and Torroni, 2015; Stab and Gurevych, 2017), consequently making it the most valuable information to extract. Argument mining and claim detection has been explored for texts like legal documents, Wikipedia articles, essays (Moens et al., 2007; Levy et al., 2014; Stab and Gurevych, 2017, i.a.), social media and web content (Goudas et al., 2014; Habernal and Gurevych, 2017; Bosc et al., 2016a; Dusmanu et al., 2017, i.a.). It has also been applied to scientific biomedical publications (Achakulvisut et al., 2019; Mayer et al., 2020, i.a.), but biomedical arguments as they occur on social media, and particularly Twitter, have not been analyzed yet.

With this paper, we fill this gap and explore claim detection for tweets discussing biomedical topics, particularly tweets about COVID-19, the measles, cystic fibrosis, and depression, to allow for drawing conclusions across different fields.

Our contributions to a better understanding of biomedical claims made on Twitter are, (1), to publish the first biomedical Twitter corpus manually labeled with claims (distinguished in explicit and implicit, and with span annotations for explicit claim phrases), and (2), baseline experiments to detect

Proceedings of the BioNLP 2021 workshop, pages 131–142

(implicit and explicit) claim tweets in a classification setting. Further, (3), we find in a cross-corpus study that a generalization across domains is challenging and that biomedical tweets pose a particularly difficult environment for claim detection.

2 Related Work

Detecting biomedical claims on Twitter is a task rooted in both the argument mining field as well as the area of biomedical text mining.

2.1 Argumentation Mining

Argumentation mining covers a variety of different domains, text, and discourse types. This includes online content, for instance Wikipedia (Levy et al., 2014; Roitman et al., 2016; Lippi and Torroni, 2015), but also more interaction-driven platforms, like fora. As an example, Habernal and Gurevych (2017) extract argument structures from blogs and forum posts, including comments. Apart from that, Twitter is generally a popular text source (Bosc et al., 2016a; Dusmanu et al., 2017). Argument mining is also applied to professionally generated content, for instance news (Goudas et al., 2014; Sardianos et al., 2015) and legal or political documents (Moens et al., 2007; Palau and Moens, 2009; Mochales and Moens, 2011; Florou et al., 2013). Another domain of interest are persuasive essays, which we also use in a cross-domain study in this paper (Lippi and Torroni, 2015; Stab and Gurevych, 2017; Eger et al., 2017).

Existing approaches differ with regards to which tasks in the broader argument mining pipeline they address. While some focus on the detection of arguments (Moens et al., 2007; Florou et al., 2013; Levy et al., 2014; Bosc et al., 2016a; Dusmanu et al., 2017; Habernal and Gurevych, 2017), others analyze the relational aspects between argument components (Mochales and Moens, 2011; Stab and Gurevych, 2017; Eger et al., 2017).

While most approaches cater to a specific domain or text genre, Stab et al. (2018) argue that domain-focused, specialized systems do not generalize to broader applications such as argument search in texts. In line with that, Daxenberger et al. (2017) present a comparative study on cross-domain claim detection. They observe that diverse training data leads to a more robust model performance in unknown domains.

2.2 Claim Detection

Claim detection is a central task in argumentation mining. It can be framed as a classification (Does a document/sentence contain a claim?) or as sequence labeling (Which tokens make up the claim?). The setting as classification has been explored, inter alia, as a retrieval task of online comments made by public stakeholders on pending governmental regulations (Kwon et al., 2007), for sentence detection in essays, (Lippi and Torroni, 2015), and for Wikipedia (Roitman et al., 2016; Levy et al., 2017). The setting as a sequence labeling task has been tackled on Wikipedia (Levy et al., 2014), on Twitter, and on news articles (Goudas et al., 2014; Sardianos et al., 2015).

One common characteristic in most work on automatic claim detection is the focus on relatively formal text. Social media, like tweets, can be considered a more challenging text type, which despite this aspect, received considerable attention, also beyond classification or token sequence labeling. Bosc et al. (2016a) detect relations between arguments, Dusmanu et al. (2017) identify factual or opinionated tweets, and Addawood and Bashir (2016) further classify the type of premise which accompanies the claim. Ouertatani et al. (2020) combine aspects of sentiment detection, opinion, and argument mining in a pipeline to analyze argumentative tweets more comprehensively. Ma et al. (2018) specifically focus on the claim detection task in tweets, and present an approach to retrieve Twitter posts that contain argumentative claims about debatable political topics.

To the best of our knowledge, detecting biomedical claims in tweets has not been approached yet. Biomedical argument mining, also for other text types, is generally still limited. The work by Shi and Bei (2019) is one of the few exceptions that target this challenge and propose a pipeline to extract health-related claims from headlines of health-themed news articles. The majority of other argument mining approaches for the biomedical domain focus on research literature (Blake, 2010; Alamri and Stevenson, 2015; Alamri and Stevensony, 2015; Achakulvisut et al., 2019; Mayer et al., 2020).

2.3 Biomedical Text Mining

Biomedical natural language processing (BioNLP) is a field in computational linguistics which also receives substantial attention from the bioinformat-

Query category			
Disease Names	Topical Hashtags	Combinations	Drugs
COVID-19, #COVID-19	#socialdistancing, #chinesevirus	COVID-19 AND cured, COVID-19 AND vaccines	Hydroxychloroquine, Kaletra, Remdesivir
measles, #measles	#vaccineswork, #dontvaccinate	measles AND vaccine, measles AND therapize	M-M-R II, Priorix, ProQuad
cystic fibrosis, #cysticfibrosis	#livesavingdrugs4cf, #orkambinow	cystic fibrosis AND treated, cystic fibrosis AND heal	Orkambi, Trikafta, Tezacaftor
depression, #depression	#depressionisreal, #notjustsad	depression AND cure, depression AND treatment	Alprazolam, Buspirone, Xanax

Table 1: Examples of the four categories of search terms used to retrieve tweets about COVID-19, the measles, cystic fibrosis, and depression via the Twitter API.

ics community. One focus is on the automatic extraction of information from life science articles, including entity recognition, e.g., of diseases, drug names, protein and gene names (Habibi et al., 2017; Giorgi and Bader, 2018; Lee et al., 2019, i.a.) or relations between those (Lamurias et al., 2019; Sousa et al., 2021; Lin et al., 2019, i.a.).

Biomedical text mining methods have also been applied to social media texts and web content (Wegrzyn-Wolska et al., 2011; Yang et al., 2016; Sullivan et al., 2016, i.a.). One focus is on the analysis of Twitter with regards to pharmacovigilance. Other topics include the extraction of adverse drug reactions (Nikfarjam et al., 2015; Cocos et al., 2017), monitoring public health (Paul and Dredze, 2012; Choudhury et al., 2013; Sarker et al., 2016), and detecting personal health mentions (Yin et al., 2015; Karisani and Agichtein, 2018).

A small number of studies looked into the comparison of biomedical information in social media and scientific text: Thorne and Klinger (2018) analyze quantitatively how disease names are referred to across these domains. Seiffe et al. (2020) analyze laypersons' medical vocabulary.

3 Corpus Creation and Analysis

As the basis for our study, we collect a novel Twitter corpus in which we annotate which tweets contain biomedical claims, and (for all explicit claims) which tokens correspond to that claim.

3.1 Data Selection & Acquisition

The data for the corpus was collected in June/July 2020 using Twitter's API[1] which offers a keyword-based retrieval for tweets. Table 1 provides a sample of the search terms we used.[2] For each of the medical topics, we sample English tweets from keywords and phrases from four different query categories. This includes (1) the name of the disease as well as the respective hashtag for each topic, e.g., *depression* and *#depression*, (2) topical hashtags like *#vaccineswork*, (3) combinations of the disease name with words like *cure, treatment* or *therapy* as well as their respective verb forms, and (4) a list of medications, products, and product brand names from the pharmaceutical database DrugBank[3].

When querying the tweets, we exclude retweets by using the API's '-filter:retweets' option. From overall 902,524 collected tweets, we filter out those with URLs since those are likely to be advertisements (Cocos et al., 2017; Ma et al., 2018), and further remove duplicates based on the tweet IDs. From the resulting collection of 127,540 messages we draw a sample of 75 randomly selected tweets per topic (four biomedical topics) and search term category (four categories per topic). The final corpus to be annotated consists of 1200 tweets about four medical issues and their treatments: measles, depression, cystic fibrosis, and COVID-19.

3.2 Annotation

3.2.1 Conceptual Definition

While there are different schemes and models of argumentative structure varying in complexity as well as in their conceptualization of claims, the claim element is widely considered the core component of an argument (Daxenberger et al., 2017).

[1] https://developer.twitter.com/en/docs/twitter-api

[2] The full list of search terms (1771 queries in total) is available in the supplementary material.

[3] https://go.drugbank.com/. At the time of creating the search term list, COVID-19 was not included in DrugBank. Instead, medications which were under investigation at the time of compiling this list as outlined on the WHO website were included for Sars-CoV-2 in this category: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/global-research-on-novel-coronavirus-2019-ncov/solidarity-clinical-trial-for-covid-19-treatments.

Aharoni et al. (2014) suggest a framework in which an argument consists of two main components: a claim and premises. We follow Stab and Gurevych (2017) and define the claim as the argumentative component in which the speaker or writer expresses the central, controversial conclusion of their argument. This claim is presented as if it were true even though objectively it can be true or false (Mochales and Ieven, 2009). The premise which is considered the second part of an argument includes all elements that are used either to substantiate or disprove the claim. Arguments can contain multiple premises to justify the claim. (Refer to Section 3.4 for examples and a detailed analysis of argumentative tweets in the dataset.)

For our corpus, we focus on the claim element and assign all tweets a binary label that indicates whether the document contains a claim. Claims can be either explicitly voiced or the claim property can be inferred from the text in cases in which they are expressed implicitly (Habernal and Gurevych, 2017). We therefore annotate explicitness or implicitness if a tweet is labeled as containing a claim. For explicit cases the claim sequence is additionally marked on the token level. For implicit cases, the claim which can be inferred from the implicit utterance is stated alongside the implicitness annotation.

3.2.2 Guideline Development

We define a preliminary set of annotation guidelines based on previous work (Mochales and Ieven, 2009; Aharoni et al., 2014; Bosc et al., 2016a; Daxenberger et al., 2017; Stab and Gurevych, 2017). To adapt those to our domain and topic, we go through four iterations of refinements. In each iteration, 20 tweets receive annotations by two annotators. Both annotators are female and aged 25–30. Annotator A1 has a background in linguistics and computational linguistics. A2 has a background in mathematics, computer science, and computational linguistics. The results are discussed based on the calculation of Cohen's κ (Cohen, 1960).

After Iteration 1, we did not make any substantial changes, but reinforced a common understanding of the existing guidelines in a joint discussion. After Iteration 2, we clarified the guidelines by adding the notion of an argumentative intention as a prerequisite for a claim: a claim is only to be annotated if the author actually appears to be intentionally argumentative as opposed to just sharing an opinion (Šnajder, 2016; Habernal and Gurevych, 2017). This is illustrated in the following example, which is not to be annotated as a claim, given this additional constraint:

> This popped up on my memories from two years ago, on Instagram, and honestly I'm so much healthier now it's quite unbelievable. A stone heavier, on week 11 of no IVs (back then it was every 9 weeks), and it's all thanks to #Trikafta and determination. I am stronger than I think.

We further clarified the guidelines with regards to the claim being the conclusive element in a Twitter document. This change encouraged the annotators to reflect specifically if the conclusive, main claim is conveyed explicitly or implicitly.

After Iteration 3, we did not introduce any changes, but went through an additional iteration to further establish the understanding of the annotation tasks.

Table 2 shows the results of the agreement of the annotators in each iteration as well as the final κ-score for the corpus. We observe that the agreement substantially increased from Iteration 1 to 4. However, we also observe that obtaining a substantial agreement for the span annotation remains the most challenging task.

	Cohen's κ		
	C/N	E/I/N	Span
Iteration 1	.31	.43	.32
Iteration 2	.34	.24	.12
Iteration 3	.61	.42	.42
Iteration 4	.60	.68	.41
Final corpus	.56	.48	.38

Table 2: Inter-annotator agreement during development of the annotation guidelines and for the final corpus. C/N: Claim/non-claim, E/I/N: Explicit/Implicit/Non-claim, Span: Token-level annotation of the explicit claim expression.

3.2.3 Annotation Procedure

The corpus annotation was carried out by the same annotators that conducted the preliminary annotations. A1 labeled 1000 tweets while A2 annotated 300 instances. From these both sets, 100 tweets were provided to both annotators, to track agreement (which remained stable, see Table 2). Annotating 100 tweets took approx. 3.3 hours. Overall, we observe that the agreement is generally moderate. Separating claim-tweets from non-claim tweets shows an acceptable κ=.56. Including the decision of explicitness/implicitness leads to κ=.48.

Class	# Instances	%	Length
non-claim	663	55.25	30.56
claim (I+E)	537	44.75	39.88
expl. claim	370	30.83	39.89
claim phrase			17.59
impl. claim	167	13.92	39.88
total	1200	100 %	34.73

Table 3: Distribution of the annotated classes and average instance lengths (in tokens).

	incompl.		blended		anecdotal		impl.	
M	8	.16	14	.28	9	.18	14	.28
C	17	.34	15	.30	8	.16	14	.28
CF	12	.24	10	.20	26	.52	18	.36
D	16	.32	9	.18	23	.46	11	.22
total	53	.27	48	.24	66	.33	57	.29

Table 4: Manual analysis of a subsample of 50 tweets/topic. Each column shows raw counts and percentage/topic.

id	Instance
1	*The French have had great success #hydroxycloroquine.*
2	Death is around 1/1000 in measles normally, same for encephalopathy, hospitalisation around 1/5. *With all the attendant costs, the vaccine saves money, not makes it.*
3	Latest: Kimberly isn't worried at all. *She takes #Hydroxychloroquine and feels awesome the next day.* Just think, it's more dangerous to drive a car than to catch corona
4	Lol exactly. It's not toxic to your body idk where he pulled this information out of. *Acid literally cured my depression/anxiety I had for 5 years in just 5 months (3 trips).* It literally reconnects parts of your brain that haven't had that connection in a long time.
5	Hopefully! The MMR toxin loaded vaccine I received many years ago seemed to work very well. More please!
6	Wow! Someone tell people with Cystic fibrosis and Huntington's that they can cure their genetics through Mormonism!

Table 5: Examples of explicit and implicit claim tweets from the corpus. Explicit claims are in italics.

The span-based annotation has limited agreement, with κ=.38 (which is why we do not consider this task further in this paper). These numbers are roughly in line with previous work. Achakulvisut et al. (2019) report an average κ=0.63 for labeling claims in biomedical research papers. According to Habernal and Gurevych (2017), explicit, intentional argumentation is easier to annotate than texts which are less explicit.

Our corpus is available with detailed annotation guidelines at http://www.ims.uni-stuttgart.de/data/bioclaim.

3.3 Corpus Statistics

Table 3 presents corpus statistics. Out of the 1200 documents in the corpus, 537 instances (44.75 %) contain a claim and 663 (55.25 %) do not. From all claim instances, 370 tweets are explicit (68 %). The claims are not equally distributed across topics (not shown in table): 61 % of measle-related tweets contain a claim, 49 % of those related to COVID-19, 40 % of cystic fibrosis tweets, and 29 % for depression.

The longest tweet in the corpus consists of 110 tokens[4], while the two shortest consist only of two tokens[5]. On average, a claim tweet has a length of ≈40 tokens. Both claim tweet types, explicit and implicit, have similar lengths (39.89 and 39.88 tokens respectively). In contrast to that, the average non-claim tweet is shorter, consisting of about 30 tokens. Roughly half of an explicit claim corresponds to the claim phrase.

We generally see that there is a connection between the length of a tweet and its class membership. Out of all tweets with up to 40 tokens, 453 instances are non-claims, while 243 contain a claim. For the instances that consist of 41 and more tokens, only 210 are non-claim tweets, whereas 294 are labeled as claims. The majority of the shorter tweets ($\leq$ 40 tokens) tend to be non-claim instances, while mid-range to longer tweets ($\geq$ 40 tokens) tend to be members of the claim class.

3.4 Qualitative Analysis

To obtain a better understanding of the corpus, we perform a qualitative analysis on a subsample of 50 claim-instances/topic. We manually analyze four claim properties: the tweet exhibits an incomplete argument structure, different argument components blend into each other, the text shows anecdotal evidence, and it describes the claim implicitly. Refer to Table 4 for an overview of the results.

In line with Šnajder (2016), we find that argument structures are often incomplete, e.g., in-

[4]The tweet includes 50 @-mentions followed by a measles-related claim: "Oh yay! I can do this too, since you're going to ignore the thousands of children who died in outbreaks last year from measles... Show me a proven death of a child from vaccines in the last decade. That's the time reference, now? So let's see a death certificate that says it, thx"

[5]"Xanax damage" and "Holy fuck".

stances only contain a stand-alone claim without any premise. This characteristic is most prevalent in the COVID-19-related tweets In Table 5, Ex. 1 is missing a premising element, Ex. 2 presents premise and claim.

Argument components (claim, premise) are not very clear cut and often blend together. Consequently they can be difficult to differentiate, for instance when authors use claim-like elements as a premise. This characteristic is again, most prevalent for COVID-19. In Ex. 3 in Table 5, the last sentence reads like a claim, especially when looked at in isolation, yet it is in fact used by the author to explain their claim.

Premise elements which substantiate and give reason for the claim (Bosc et al., 2016b) traditionally include references to studies or mentions of expert testimony, but occasionally also anecdotal evidence or concrete examples (Aharoni et al., 2014). We find the latter to be very common for our data set. This property is most frequent for cystic fibrosis and depression. Ex. 4 showcases how a personal experience is used to build an argument.

Implicitness in the form of irony, sarcasm or rhetoric questions are common features for these types of claims on Twitter. We observe claims related to cystic fibrosis are most often (in our sample) implicit. Ex. 5 and 6 show instances that use sarcasm or irony. The fact that implicitness is such a common feature in our dataset is in line with the observation that implicitness is a characteristic device not only in spoken language and everyday, informal argumentation (Lumer, 1990), but also in user-generated web content in general (Habernal and Gurevych, 2017).

4 Methods

In the following sections we describe the conceptual design of our experiments and introduce the models that we use to accomplish the claim detection task.

4.1 Classification Tasks

We model the task in a set of different model configurations.

Binary. A trained classifier distinguishes between claim and non-claim.

Multiclass. A trained classifier distinguishes between exlicit claim, implicit claim, and non-claim.

Multiclass Pipeline. A first classifier learns to discriminate between claims and non-claims (as in Binary). Each tweet that is classified as claim is further separated into implicit or explicit with another binary classifier. The secondary classifier is trained on gold data (not on predictions of the first model in the pipeline).

4.2 Model Architecture

For each of the classification tasks (binary/multiclass, steps in the pipeline), we use a set of standard text classification methods which we compare. The first three models (NB, LG, BiLSTM) use 50-dimensional FastText (Bojanowski et al., 2017) embeddings trained on the Common Crawl corpus (600 billion tokens) as input[6].

NB. We use a (Gaussian) naive Bayes with an average vector of the token embeddings as input.

LG. We use a logistic regression classifier with the same features as in NB.

BiLSTM. As a classifier which can consider contextual information and makes use of pretrained embeddings, we use a bidirectional long short-term memory network (Hochreiter and Schmidhuber, 1997) with 75 LSTM units followed by the output layer (sigmoid for binary classification, softmax for multiclass).

BERT. We use the pretrained BERT (Devlin et al., 2019) base model[7] and fine-tune it using the claim tweet corpus.

5 Experiments

5.1 Claim Detection

With the first experiment we explore how reliably we can detect claim tweets in our corpus and how well the two different claim types (*explicit* vs. *implicit claim tweets*) can be distinguished. We use each model mentioned in Section 4.2 in each setting described in Section 4.1. We evaluate each classifier in a binary or (where applicable) in a multi-class setting, to understand if splitting the claim category into its subcomponents improves the claim prediction overall.

[6] https://fasttext.cc/docs/en/english-vectors.html
[7] https://huggingface.co/bert-base-uncased

			NB			LG			LSTM			BERT		
Eval.	Task	Class	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
binary	binary	claim	.67	.65	.66	.66	.74	**.70**	.68	.48	.57	.66	.72	.69
		n-claim	.75	.77	**.76**	.79	.72	**.76**	.69	.84	.75	.78	.72	.75
	multiclass	claim	.66	.65	**.66**	.73	.53	.61	.75	.35	.48	.81	.49	.61
		n-claim	.74	.76	.75	.71	.85	.78	.66	.91	.76	.71	.91	**.80**
multi-class	multiclass	expl	.55	.45	.50	.63	.39	.48	.59	.27	.37	.62	.45	**.52**
		impl	.31	.44	**.36**	.33	.35	.34	.18	.09	.12	.29	.09	.13
		n-claim	.74	.76	.75	.71	.85	.78	.66	.91	.76	.71	.91	**.80**
	pipeline	expl	.56	.45	.50	.52	.55	.53	.50	.37	.43	.54	.65	**.59**
		impl	.31	.44	**.36**	.28	.35	.31	.07	.04	.05	.26	.22	.24
		n-claim	.75	.77	**.76**	.79	.72	**.76**	.69	.84	.75	.78	.72	.75

Table 6: Results for the claim detection experiments, separated into binary and multi-class evaluation. The best F_1 scores for each evaluation setting and class are printed in bold face.

5.1.1 Experimental Setting

From our corpus of 1200 tweets we use 800 instances for training, 200 as validation data to optimize hyperparameters and 200 as test data. We tokenize the documents and substitute all @-mentions by "@username". For the LG models, we use an l2 regularization. For the LSTM models, the hyper-parameters learning rate, dropout, number of epochs, and batch size were determined by a randomized search over a parameter grid and also use l2 regularization. For training, we use Adam (Kingma and Ba, 2015). For the BERT models, we experiment with combinations of the recommended fine-tuning hyper-parameters from Devlin et al. (2019) (batch size, learning rate, epochs), and use those with the best performance on the validation data. An overview of all hyper-parameters is provided in Table 9 in the Appendix. For the Bi-LSTM, we use the Keras API (Chollet et al., 2015) for TensorFlow (Abadi et al., 2015). For the BERT model, we use Simple Transformers (Rajapakse, 2019) and its wrapper for the Hugging Face transformers library (Wolf et al., 2020). Further, we oversample the minority class of implicit claims to achieve a balanced training set (the test set remains with the original distribution). To ensure comparability, we oversample in both the binary and the multi-class setting. For parameters that we do not explicitly mention, we use default values.

5.1.2 Results

Table 6 reports the results for the conducted experiments. The top half lists the results for the binary claim detection setting. The bottom half of the table presents the results for the multi-class claim classification.

For the binary evaluation setting, we observe that casting the problem as a ternary prediction task is not beneficial – the best F_1 score is obtained with the binary LG classifier (.70 F_1 for the class claim in contrast to .61 F_1 for the ternary LG). The BERT and NB approaches are slightly worse (1 pp and 4pp less for binary, respectively), while the LSTM shows substantially lower performance (13pp less).

In the ternary/multi-class evaluation, the scores are overall lower. The LSTM shows the lowest performance. The best result is obtained in the pipeline setting, in which separate classifiers can focus on distinguishing claim/non-claim and explicit/implicit – we see .59 F_1 for the explicit claim class. Implicit claim detection is substantially more challenging across all classification approaches.

We attribute the fact that the more complex models (LSTM, BERT) do not outperform the linear models across the board to the comparably small size of the dataset. This appears especially true for implicit claims in the multi-class setting. Here, those models struggle the most to predict implicit claims, indicating that they were not able to learn sufficiently from the training instances.

5.1.3 Error Analysis

From a manual introspection of the best performing model in the binary setting, we conclude that it is difficult to detect general patterns. We show two cases of false positives and two cases of false negatives in Table 7. The false positive instances show that the model struggles with cases that rely on judging the argumentative intention. Both Ex. 1 and 2 contain potential claims about depression and therapy, but they have not been annotated as such, because the authors' intention is motivational rather than argumentative. In addition, it appears

id	G	P	Text
1	n	c	#DepressionIsReal #MentalHealthAwareness #mentalhealth ruins lives. #depression destroys people. Be there when someone needs you. It could change a life. It may even save one.
2	n	c	The reason I stepped away from twitch and gaming with friends is because iv been slowly healing from a super abusive relationship. Going to therapy and hearing you have ptsd isnt easy. But look how far iv come, lost some depression weight and found some confidence:)plz stay safe
3	c	n	Not sure who knows more about #COVID19, my sister or #DrFauci. My money is on Stephanie.
4	c	n	How does giving the entire world a #COVID19 #vaccine compare to letting everyone actually get #covid? What would you prefer? I'm on team @username #WHO #CDC #math #VaccinesWork #Science

Table 7: Examples of incorrect predictions by the LG model in the binary setting (G:Gold, P:Predictions; n: no claim; c: claim).

that the model struggles to detect implicit claims that are expressed using irony (Ex. 3) or a rhetorical question (Ex. 4).

5.2 Cross-domain Experiment

We see that the models show acceptable performance in a binary classification setting. In the following, we analyze if this observation holds across domains or if information from another out-of-domain corpus can help.

As the binary LG model achieved the best results during the previous experiment, we use this classifier for the cross-domain experiments. We work with paragraphs of persuasive essays (Stab and Gurevych, 2017) as a comparative corpus. The motivation to use this resource is that while they are a distinctly different text type and usually linguistically much more formal than tweets, they are also opinionated documents.[8] We use the resulting essay model for making an in-domain as well as a cross-domain prediction and vice versa for the Twitter model. We further experiment with combining the training portions of both datasets and evaluate its performance for both target domains.

5.2.1 Experimental Setting

The comparative corpus contains persuasive essays with annotated argument structure (Stab and Gurevych, 2017). Eger et al. (2017) used this corpus subsequently and provide the data in CONLL-format, split into paragraphs, and predivided into train, development and test set.[9] We use their version of the corpus. The annotations for the essay corpus distinguish between major claims and claims. However, since there is no such hierarchical differentiation in the Twitter annotations, we consider both types as equivalent. We choose to use paragraphs instead of whole essays as the individual input documents for the classification and assign a claim label to every paragraph that contains a claim. This leaves us with 1587 essay paragraphs as training data, and 199 and 449 paragraphs respectively for validation and testing.

We follow the same setup as for the binary setting in the first experiment.

Train	Test	P	R	F_1
Twitter	Twitter	.66	.74	.70
Essay	Twitter	.51	.69	.59
Twitter+Essay	Twitter	.58	.75	.66
Essay	Essay	.96	1.0	.98
Twitter	Essay	.94	.74	.83
Twitter+Essay	Essay	.95	1.0	.97

Table 8: Results of cross-domain experiments using the binary LG model on the Twitter and the essay corpus. We report precision, recall and F_1 for the claim tweet class.

5.2.2 Results

In Table 8, we summarize the results of the cross-domain experiments with the persuasive essay corpus. We see that the essay model is successful for classifying claim documents (.98 F_1) in the in-domain experiment. Compared to the in-domain setting for tweets all evaluation scores measure substantially higher.

When we compare the two cross-domain experiments, we observe that the performance measures decrease in both settings when we use the out-of-domain model to make predictions (11pp in F_1 for tweets, 15pp for essays). Combining the training portions of both data sets does not lead to an improvement over in-domain experiments. This shows the challenge of building domain-generic models that perform well across different data sets.

6 Discussion and Future Work

In this paper, we presented the first data set for biomedical claim detection in social media. In our

[8] An essay is defined as *"a short piece of writing on a particular subject, often expressing personal views"* (https://dictionary.cambridge.org/dictionary/english/essay).

[9] https://github.com/UKPLab/acl2017-neural_end2end_am/tree/master/data/conll/Paragraph_Level

first experiment, we showed that we can achieve an acceptable performance to detect claims when the distinction between explicit or implicit claims is not considered. In the cross-domain experiment, we see that text formality, which is one of the main distinguishing feature between the two corpora, might be an important factor that influences the level of difficulty in accomplishing the claim detection task.

Our hypothesis in this work was that biomedical information on Twitter exhibits a challenging setting for claim detection. Both our experiments indicate that this is true. Future work needs to investigate what might be reasons for that. We hypothesize that our Twitter dataset contains particular aspects that are specific to the medical domain, but it might also be that other latent variables lead to confounders (e.g., the time span that has been used for crawling). It is important to better understand these properties.

We suggest future work on claim detection models optimize those to work well across domains. To enable such research, this paper contributed a novel resource. This resource could further be improved. One way of addressing the moderate agreement between the annotators could be to include annotators with medical expertise to see if this ultimately facilitates claim annotation. Additionally, a detailed introspection of the topics covered in the tweets for each disease would be interesting for future work since this might shed some light on which topical categories of claims are particularly difficult to label.

The COVID-19 pandemic has sparked recent research with regards to detecting misinformation and fact-checking claims (e.g., Hossain et al. (2020) or Wadden et al. (2020)). Exploring how a claim-detection-based fact-checking approach rooted in argument mining compares to other approaches is up to future research.

Acknowledgments

This research has been conducted as part of the FIBISS project[10] which is funded by the German Research Council (DFG, project number: KL 2869/5-1). We thank Laura Ana Maria Oberländer for her support and the anonymous reviewers for their valuable comments.

[10]Automatic Fact Checking for Biomedical Information in Social Media and Scientific Literature, https://www.ims.uni-stuttgart.de/en/research/projects/fibiss/

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Titipat Achakulvisut, Chandra Bhagavatula, Daniel Acuna, and Konrad Kording. 2019. Claim extraction in biomedical publications using deep discourse model and transfer learning. *arXiv preprint arXiv:1907.00962*.

Aseel Addawood and Masooda Bashir. 2016. "What is your evidence?" A study of controversial topics on social media. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 1–11, Berlin, Germany. Association for Computational Linguistics.

Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. 2014. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proceedings of the First Workshop on Argumentation Mining*, pages 64–68, Baltimore, Maryland. Association for Computational Linguistics.

Abdulaziz Alamri and Mark Stevenson. 2015. Automatic detection of answers to research questions from Medline abstracts. In *Proceedings of BioNLP 15*, pages 141–146, Beijing, China. Association for Computational Linguistics.

Abdulaziz Alamri and Mark Stevensony. 2015. Automatic identification of potentially contradictory claims to support systematic reviews. In *Proceedings of the 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, BIBM '15, page 930–937, USA. IEEE Computer Society.

Catherine Blake. 2010. Beyond genes, proteins, and abstracts: Identifying scientific claims from full-text biomedical articles. *Journal of biomedical informatics*, 43(2):173–189.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Tom Bosc, Elena Cabrio, and Serena Villata. 2016a. DART: a dataset of arguments and their relations

on Twitter. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1258–1263, Portorož, Slovenia. European Language Resources Association (ELRA).

Tom Bosc, Elena Cabrio, and Serena Villata. 2016b. Tweeties squabbling: Positive and negative results in applying argument mining on social media. *COMMA*, 2016:21–32.

François Chollet et al. 2015. Keras. https://keras.io.

Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Social media as a measurement tool of depression in populations. In *In Proceedings of the 5th ACM International Conference on Web Science (Paris, France, May 2-May 4, 2013). WebSci 2013.*

Anne Cocos, Alexander G Fiks, and Aaron J Masino. 2017. Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in twitter posts. *Journal of the American Medical Informatics Association*, 24(4):813–821.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Rion Brattig Correia, Ian B. Wood, Johan Bollen, and Luis M. Rocha. 2020. Mining Social Media Data for Biomedical Signals and Health-Related Behavior. *Annual Review of Biomedical Data Science*, 3(1):433–458. _eprint: https://doi.org/10.1146/annurev-biodatasci-030320-040844.

Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. 2017. What is the essence of a claim? Cross-domain claim identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2055–2066, Copenhagen, Denmark. Association for Computational Linguistics.

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 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mihai Dusmanu, Elena Cabrio, and Serena Villata. 2017. Argument mining on Twitter: Arguments, facts and sources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2317–2322, Copenhagen, Denmark. Association for Computational Linguistics.

Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2017. Neural end-to-end learning for computational argumentation mining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11–22, Vancouver, Canada. Association for Computational Linguistics.

Eirini Florou, Stasinos Konstantopoulos, Antonis Koukourikos, and Pythagoras Karampiperis. 2013. Argument extraction for supporting public policy formulation. In *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 49–54, Sofia, Bulgaria. Association for Computational Linguistics.

John M Giorgi and Gary D Bader. 2018. Transfer learning for biomedical named entity recognition with neural networks. *Bioinformatics*, 34(23):4087–4094.

Theodosis Goudas, Christos Louizos, Georgios Petasis, and Vangelis Karkaletsis. 2014. Argument extraction from news, blogs, and social media. In *Artificial Intelligence: Methods and Applications*, pages 287–299, Cham. Springer International Publishing.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179.

Maryam Habibi, Leon Weber, Mariana Neves, David Luis Wiegandt, and Ulf Leser. 2017. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14):i37–i48.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Tamanna Hossain, Robert L. Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh. 2020. COVIDLies: Detecting COVID-19 Misinformation on Social Media. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Payam Karisani and Eugene Agichtein. 2018. Did you really just have a heart attack? Towards robust detection of personal health mentions in social media. In *Proceedings of the 2018 World Wide Web Conference*, page 137–146, Republic and Canton of Geneva, CHE.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Namhee Kwon, Liang Zhou, Eduard Hovy, and Stuart W. Shulman. 2007. Identifying and classifying subjective claims. In *Proceedings of the 8th Annual International Conference on Digital Government Research: Bridging Disciplines & Domains*, dg.o '07, page 76–81. Digital Government Society of North America.

Andre Lamurias, Diana Sousa, Luka A. Clarke, and Francisco M. Couto. 2019. Bo-lstm: classifying relations via long short-term memory networks along biomedical ontologies. *BMC Bioinformatics*, 20(1):10.

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. *Bioinformatics*.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Ran Levy, Shai Gretz, Benjamin Sznajder, Shay Hummel, Ranit Aharonov, and Noam Slonim. 2017. Unsupervised corpus–wide claim detection. In *Proceedings of the 4th Workshop on Argument Mining*, pages 79–84, Copenhagen, Denmark. Association for Computational Linguistics.

Chen Lin, Timothy Miller, Dmitriy Dligach, Steven Bethard, and Guergana Savova. 2019. A BERT-based universal model for both within- and cross-sentence clinical temporal relation extraction. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 65–71, Minneapolis, Minnesota, USA. Association for Computational Linguistics, Association for Computational Linguistics.

Marco Lippi and Paolo Torroni. 2015. Context-independent claim detection for argument mining. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, page 185–191. AAAI Press.

Christoph Lumer. 1990. *Praktische Argumentationstheorie: theoretische Grundlagen, praktische Begründung und Regeln wichtiger Argumentationsarten*. Hochschulschrift, University of Münster, Braunschweig.

Wenjia Ma, WenHan Chao, Zhunchen Luo, and Xin Jiang. 2018. CRST: a claim retrieval system in Twitter. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 43–47, Santa Fe, New Mexico. Association for Computational Linguistics.

Tobias Mayer, Elena Cabrio, and Serena Villata. 2020. Transformer-based Argument Mining for Healthcare Applications. In *24th European Conference on Artificial Intelligence (ECAI2020)*, Santiago de Compostela, Spain.

Raquel Mochales and Aagje Ieven. 2009. Creating an argumentation corpus: Do theories apply to real arguments? a case study on the legal argumentation of the echr. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, ICAIL '09, page 21–30, New York, NY, USA. Association for Computing Machinery.

Raquel Mochales and Marie-Francine Moens. 2011. Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22.

Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. 2007. Automatic detection of arguments in legal texts. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law*, ICAIL '07, page 225–230, New York, NY, USA. Association for Computing Machinery.

Azadeh Nikfarjam, Abeed Sarker, Karen O'Connor, Rachel Ginn, and Graciela Gonzalez. 2015. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, 22(3):671–681.

Asma Ouertatani, Ghada Gasmi, and Chiraz Latiri. 2020. Parsing argued opinion structure in Twitter content. *Journal of Intelligent Information Systems*, pages 1–27.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: The detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, ICAIL '09, page 98–107, New York, NY, USA. Association for Computing Machinery.

Michael J. Paul and Mark Dredze. 2012. A model for mining public health topics from Twitter. *Health*, 11(1).

Thilina Rajapakse. 2019. Simple transformers. https://github.com/ThilinaRajapakse/simpletransformers.

Haggai Roitman, Shay Hummel, Ella Rabinovich, Benjamin Sznajder, Noam Slonim, and Ehud Aharoni. 2016. On the retrieval of wikipedia articles containing claims on controversial topics. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, page 991–996, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Christos Sardianos, Ioannis Manousos Katakis, Georgios Petasis, and Vangelis Karkaletsis. 2015. Argument extraction from news. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 56–66, Denver, CO. Association for Computational Linguistics.

Abeed Sarker, Karen O'Connor, Rachel Ginn, Matthew Scotch, Karen Smith, Dan Malone, and Graciela

Gonzalez. 2016. Social media mining for toxicovigilance: automatic monitoring of prescription medication abuse from Twitter. *Drug safety*, 39(3):231–240.

Laura Seiffe, Oliver Marten, Michael Mikhailov, Sven Schmeier, Sebastian Möller, and Roland Roller. 2020. From witch's shot to music making bones - resources for medical laymen to technical language and vice versa. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6185–6192, Marseille, France. European Language Resources Association.

Yuan Shi and Yu Bei. 2019. HClaimE: A tool for identifying health claims in health news headlines. *Information Processing & Management*, 56(4):1220–1233.

Jan Šnajder. 2016. Social media argumentation mining: the quest for deliberateness in raucousness. *arXiv preprint arXiv:1701.00168*.

Diana Sousa, Andre Lamurias, and Francisco M. Couto. 2021. *Using Neural Networks for Relation Extraction from Biomedical Literature*, pages 289–305. Springer US, New York, NY.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674, Brussels, Belgium. Association for Computational Linguistics.

Ryan Sullivan, Abeed Sarker, Karen O'Connor, Amanda Goodin, Mark Karlsrud, and Graciela Gonzalez. 2016. Finding potentially unsafe nutritional supplements from user reviews with topic modeling. In *Biocomputing 2016*, pages 528–539, Kohala Coast, Hawaii, USA.

Camilo Thorne and Roman Klinger. 2018. On the semantic similarity of disease mentions in medline and twitter. In *Natural Language Processing and Information Systems: 23rd International Conference on Applications of Natural Language to Information Systems, NLDB 2018, Paris, France, June 13-15, 2018, Proceedings*, Cham. Springer International Publishing.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or Fiction: Verifying Scientific Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Parameter	LSTM2	LSTM3
Embedding	fastText	fastText
Emb. Dim.	50	50
# LSTM units	75	75
Training epochs	60	70
Training batch size	10	30
Loss function	Binary CE	Categ. CE
Optimizer	Adam	Adam
Learning rate	1e-3	1e-3
L2 regularization	1e-3	1e-3
dropout	0.5	0.6

(a) Overview of architectural choices and hyper-parameter settings for the binary (LSTM2) and multi-class (LSTM3) LSTM-based models used in our experiments.

Parameter	BERT2	BERT3
Training epochs	4	4
Training batch size	16	16
Learning rate	2e-5	3e-5

(b) Overview of fine-tuning hyper-parameters for the binary (BERT2) and multi-class (BERT3) models used in our experiments.

Table 9: Overview of model hyper-parameters.

Katarzyna Wegrzyn-Wolska, Lamine Bougueroua, and Grzegorz Dziczkowski. 2011. Social media analysis for e-health and medical purposes. In *2011 International Conference on Computational Aspects of Social Networks (CASoN)*, pages 278–283.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Fu-Chen Yang, Anthony J.T. Lee, and Sz-Chen Kuo. 2016. Mining health social media with sentiment analysis. *Journal of Medical Systems*, 40(11):236.

Zhijun Yin, Daniel Fabbri, S Trent Rosenbloom, and Bradley Malin. 2015. A scalable framework to detect personal health mentions on Twitter. *Journal of Medical Internet Research*, 17(6):e138.