Detecting Health Advice in Medical Research Literature

Yingya Li
School of Information Studies
Syracuse University
yli48@syr.edu

Jun Wang
Independent Researcher
Syracuse, NY
junwang4@gmail.com

Bei Yu
School of Information Studies
Syracuse University
byu@syr.edu

Abstract

Health and medical researchers often give clinical and policy recommendations to inform health practice and public health policy. However, no current health information system supports the direct retrieval of health advice. This study fills the gap by developing and validating an NLP-based prediction model for identifying health advice in research publications. We annotated a corpus of 6,000 sentences extracted from structured abstracts in PubMed publications as "strong advice", "weak advice", or "no advice", and developed a BERT-based model that can predict, with a macro-averaged F1-score of 0.93, whether a sentence gives strong advice, weak advice, or not. The prediction model generalized well to sentences in both unstructured abstracts and discussion sections, where health advice normally appears. We also conducted a case study that applied this prediction model to retrieve specific health advice on COVID-19 treatments from LitCovid, a large COVID research literature portal, demonstrating the usefulness of retrieving health advice sentences as an advanced research literature navigation function for health researchers and the general public.

1 Introduction

Clinical practice and public health policies need to be guided by evidence presented in peer-reviewed medical literature, where researchers present their findings and discuss implications (Schaafsma et al., 2005). Sometimes, researchers would even give clinical and policy recommendations, defined as "health advice" in this study. However, whether to give health advice in medical research papers is a controversial issue. The opponents are concerned about the quality of health advice given in individual research papers in that a single paper lacks sufficient information for all evidence in real practice and there is limited manuscript space for full review of alternative choices (Cummings, 2007). For example, some medical experts warned that a large proportion of health advice inferred from over-interpreted observational results was not fully supported by the study presented (Wilson and Chestnutt, 2016; Banerjee and Prasad, 2020), although such health advice is frequently found in medical publications, even in highly influential journals (Prasad et al., 2013).

On the other hand, proponents of "actionable research" would like to encourage more efficient and effective transmission of science evidence into practice (Green et al., 2009). Giving health advice may also benefit the general public. Some journal editors suggested "keeping the needs of less scientifically inclined readers in mind" (Pless, 2007). After all, if researchers themselves do not discuss the practical value of their findings, press officers and journalists might misinterpret the results and give exaggerated health advice in press releases and news articles (Sumner et al., 2014; Haneef et al., 2015).

Both arguments for and against giving health advice in individual studies indicate a strong need for identifying and accessing health advice, for either practical use or quality evaluation purpose. However, navigating the large volume of medical literature is a daunting task (Straus and Haynes, 2009; Fry and Attawet, 2018), and outdated information system has been a barrier to accessing health advice (Green et al., 2009). The fast growth of medical and health literature further exacerbates the challenge (Williamson and Minter, 2019). For example, the most recent COVID-19 outbreak has brought an explosion of scientific papers about the disease. The strong need for understanding the fast-growing scientific evidence has led to the creation of specialized data hubs and search platforms (e.g. Chen et al., 2020; Wang et al., 2020; Hope et al., 2020). Nevertheless, they have not been able to provide functions to support the direct retrieval of health advice.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6018–6029
November 7–11, 2021. ©2021 Association for Computational Linguistics

A new information service that allows for direct access to health advice would be able to reduce the information barrier. The core function would be a prediction model that can automatically identify health advice in science literature. However, past studies of health advice were limited to small-scale manual analyses (e.g. Prasad et al., 2013; Sumner et al., 2014; Wilson and Chestnutt, 2016), although they were instrumental in defining health advice semantically and establishing snapshot views based on small data sets. Due to the significant time and labor cost, machine learning and NLP-based methods are needed for fast and iterative analyses of health advice in a large volume of research publications, news articles, and online posts, not to mention tracking the diffusion of health advice across domains and over time. Recent advance in natural language processing has resulted in a body of research on information extraction from medical literature (e.g. Patel et al., 2018; Zhou and Li, 2020; Resnik et al., 2020), however, automated detection of health advice has not been well explored.

This study aims to develop a computational approach to automatically detect and categorize health advice. We developed and validated an NLP-based prediction model to automatically find health advice in PubMed literature. A total of 6,000 sentences were extracted from research papers, manually annotated, and then used to fine-tune a BERT-based prediction model to predict whether a sentence contains "strong advice", "weak advice", or "no advice". To demonstrate the potential use scenario of the prediction model, this model was then applied to retrieving health advice regarding the use of Hydroxychloroquine (HCQ) as a treatment option from LitCovid, a large COVID research literature database curated by NIH. HCQ was considered a promising treatment option at the beginning of COVID but was found to be ineffective in later clinical trials. More specifically, we seek empirical answers to the following research questions: 1) to what extent can NLP models identify health advice in medical research literature? 2) what health advice has been made regarding the use of HCQ for COVID-19 treatment?

2 Related Work

2.1 Definition of Health Advice

Our task of detecting health advice consists of identifying statements that give advice and categorize advice by its level of commitment. From the language perspective, advice is a form of imperative utterance that can convey a speaker's wishes or suggestions of an action (Condoravdi and Lauer, 2012). Imperative language is also considered as part of an illocutionary act which is one of the essential units of human linguistic communication (Austin, 1975; Searle, 1976). Level of commitment indicates how strong the advice is. It is normally realized by language indicators, such as hedges (Lakoff, 1972; Myers, 1989; Hyland, 1998a,b), modalities (Hyland, 1994, 1995, 1996), and evidentials (Anderson, 1986; Mushin, 2001).

Prasad et al. (2013) defined health advice as recommendations related to any activities that might be performed by members of a health care team. They gave binary labels to research articles as either providing health advice or not. Read et al. (2016) annotated recommendations in clinical practice guidelines based on their strength. They categorized advice into "strong", "moderate", and "weak" to indicate its importance and level of confidence of the advice giver. Sumner et al. (2014) annotated health advice at sentence level, and further distinguished health advice as either "explicit" or "implicit" type. "Explicit advice" is linguistically characterized by a direct recommendation for changes. In comparison, "implicit advice" hints for changes without a direct recommendation, and thus may use different linguistic cues. Furthermore, "explicit advice" indicates a higher level of commitment than "implicit advice" since straightforward recommendations are made for behavioral change.

Despite the different naming conventions, the concept of "explicit/implicit" advice seems to be well aligned with the "strong/weak" classification, with both distinguishing the levels of commitment. In fact, Sumner et al. (2014) defined a piece of "exaggerated health advice" if an original research paper expressed it as implicit, but a news article paraphrased it as explicit.

Drawing on the past research, we categorized sentences as "no advice", "weak advice", or "strong advice" to capture both advice occurrence and level of commitment.

2.2 Suggestion Mining in the NLP Field

The task of identifying health advice is closely related to the NLP problem of suggestion mining. Prior studies defined suggestion mining as a sentence-level classification task with the purpose to detect wishes, advice, or recommendations

from opinionated text (e.g. Goldberg et al., 2009; Ramanand et al., 2010; Brun and Hagege, 2013; Negi, 2016; Negi et al., 2019). To this end, different kinds of opinionated text such as customer reviews (e.g. Goldberg et al., 2009; Ramanand et al., 2010; Brun and Hagege, 2013; Negi, 2016; Negi et al., 2019), discussion forum posts (Goldberg et al., 2009; Wicaksono and Myaeng, 2012, 2013), and tweets (Dong et al., 2013) were built to train computational models. To date, available datasets are mostly for online customer reviews and social media posts. Corpora on advice in scientific literature are lacking, especially in the health domain.

To automatically extract suggestions, both rule-based and machine learning approaches were proposed. Earlier work of suggestion mining applied a rule-based approach to identify sentences with suggestions. This type of studies often applied domain-specific and hand-craft linguistic rules to extract advice-related statements (e.g. Ramanand et al., 2010; Brun and Hagege, 2013). Meanwhile, machine learning approaches such as CRF (Wicaksono and Myaeng, 2013), Factorization Machines (Dong et al., 2013), and SVMs (Negi and Buitelaar, 2015) were utilized and compared to identify suggestions.

Recently, deep learning approaches were also used to identity sentences with suggestions. For example, in the suggestion mining task of SemEval-2019 (Negi et al., 2019), CNN-based (e.g. Park et al., 2019; Yue et al., 2019) and LSTM-based models (Cabanski, 2019) were developed to extract suggestions in online reviews and forums. Additionally, pre-trained language models such as BERT (Devlin et al., 2019) were used to detect suggestions (e.g. Liu et al., 2019; Park et al., 2019).

Overall suggestion mining still remains an emerging research area in comparison to other NLP tasks. Health advice has not been computationally modeled as a language construct. Therefore, more work is needed to examine the feasibility of applying NLP techniques to detect health advice in science communication. In this work, we investigated both traditional and deep learning methods for predicting health advice with a new, human-annotated corpus.

3 Corpus Construction

We chose PubMed as the data source, since PubMed is the largest health literature database. Besides abstracts, it provides rich metadata that can help select research papers with different types of study designs. PubMed Central, an open-access subset of PubMed, provides full text content of a portion of PubMed-indexed publications.

The Evidence-Based Medicine (EBM) Pyramid specifies that different study designs lead to different levels of evidence toward medical decision making (Murad et al., 2016). To ensure our health advice prediction model is effective in identifying health advice across study designs, we sampled research papers from both randomized controlled trials (RCTs) and observational studies by the MeSH terms in PubMed. An RCT would randomly assign individuals into experiment and control groups in order to compare the effect of treatments or interventions (Kabisch et al., 2011). In an observational study, individuals are observed or certain outcomes are measured; however, no interventions and treatments are carried out by researchers to affect the outcome (Mann, 2003). Observational studies are widely applied in fields of epidemiology, social sciences, and psychology, when RCTs are not always possible or ethical to conduct (Song and Chung, 2010). Within the observational studies, we further sampled from four common subtypes, namely cross-sectional, case-control, retrospective, and prospective studies, in increasing order of evidence strength.

Health advice normally appears in either abstracts or conclusion/discussion sections. Since sentences containing health advice account for a very small portion of all sentences, to avoid annotating a large number of non-advice sentences, we annotated a sample of sentences from the conclusion subsections in structured abstracts. Note that the abstracts in PubMed are either "unstructured" or "structured". An "IMRaD" structured abstract consists of several subsections: introduction/background, method, result, and conclusion/discussion. In fact structured abstracts have now become the predominant mode of abstracts in major medical journals (Nakayama et al., 2005). The conclusion/discussion subsection is usually a few sentences long, and thus is the most balanced source for both health advice and non-advice sentences.

In the end, a total of 6,000 sentences were sampled from conclusion/discussion subsections, including 3,000 from observational studies and 3,000 from RCTs. Based on the three-category coding schema, each sentence was assigned to one of the

Label	Description	Example Sentence
Strong Advice	The statement makes a straightforward recommendation for health-related behavior and practice. The recommendation could lead to actionable practice and policy changes. It can target patients, health and medical professionals or the public.	1. Nurses should assess patient decision-making styles to ensure maximum patient involvement in the decision-making process based on personal desires regardless of age. (PMID: 26679453) 2. A carefully integrated diabetic retinopathy screening service is needed, particularly in remote areas, to improve adherence rates. (PMID: 28490306)
Weak Advice	The statement hints that either behavior or health-related practice needs changing. Or the statement suggests that there are certain options and alternative approaches for the existing clinical and medical practice.	3. Adolescents with high risk factors, especially those with menstrual disorders and hyperandrogenism, may need careful clinical screening. (PMID: 23089573) 4. A TyG threshold of 8.5 was highly sensitive for detecting NAFLD subjects and may be suitable as a diagnostic criterion for NAFLD in Chinese adults. (PMID: 28103934)
No Advice	The statement just describes study background, results, findings, limitations or suggestions for future studies etc., and there is no suggestion for behavioral or clinical practice.	5. Former smokers are at risk for hypertension, probably because of the higher prevalence of overweight and obese subjects in this group. (PMID: 11821702) 6. The results of the study show that in the course of HIV infection overweight/obesity affected men and women admitted with normal weight, although a greater proportion of women progressed to obesity. (PMID: 20694301)

Table 1: Health advice annotation schema and sentence examples.

three category labels "no advice", "weak advice", or "strong advice". Table 1 shows the category definitions and examples.

Later in this paper we will demonstrate that a prediction model based on this annotated data set is generalizable to distinguishing advice and non-advice sentences in unstructured abstracts and discussion sections. For this evaluation we further annotated all sentences in 100 unstructured abstracts and 100 discussion sections, which will be described in the model evaluation section.

To test the validity of the proposed schema, a sample of 100 sentences were randomly selected for inter-coder agreement evaluation. Two annotators each labelled the 100 sentences and highlighted the linguistic cues for health advice. The overall Cohen's Kappa agreement (Cohen, 1960) was 0.86, indicating a near-perfect inter-coder agreement (McHugh, 2012). Disagreed cases were later resolved by the two annotators through discussion.

Three annotators with academic backgrounds in clinical psychology, linguistics, and information science were then trained to annotate the entire training corpus. All ambiguous cases were highlighted during the annotation and brought to all team members for group discussion to reach an agreement on the annotations.

In most cases, an advice sentence has one advice type only. Occasionally, a sentence includes both "weak advice" and "strong advice". We treated such cases as mixed examples and excluded them from the training corpus. The final corpus includes 5,982 sentences. Since the majority of conclusion sentences do not contain advice, the category distribution in Table 2 shows a skewed distribution with "no advice" as the largest category.

4 Prediction Model Development and Evaluation

Similar to tasks in suggestion mining, we framed automated detection of health advice as a sentence-level text classification task. We trained and evaluated three machine learning approaches: Linear SVM, BERT (Devlin et al., 2019), and BioBERT (Lee et al., 2020) to identify sentences containing health advice.

4.1 Prediction Model

LinearSVM: We chose the SVM algorithm with different vectorization methods to train the advice-type classifier using the Scikit-learn python package (Pedregosa et al., 2011). The penalty value C in LinearSVM was set to 1. A comparison of different word vector representation methods showed that the tf-idf vectorization performed similarly to the count vectorization, and adding bigrams also improved the SVM model's performance.

BERT: BERT is a recent method for pre-training language representations, and it has achieved state-of-the-art results in a number of NLP tasks (Devlin et al., 2019; Fan et al., 2020). As for suggestion mining, the BERT-based model also outperformed

	RCTs	Cross-Sectional	Case-Control	Retrospective	Prospective	Total	Percentage
None	1227	582	588	587	591	3575	59.8%
Weak	1037	82	85	144	134	1482	24.8%
Strong	652	92	45	84	52	925	15.5%
Total	2916	756	718	815	777	5982	

Table 2: Distribution of advice type in annotated corpus.

the other machine learning approaches developed in the SemEval-2019 task (Negi et al., 2019). Our model settings are: cased BERT-base, 3 epochs, learning rate of 2e-5, and max sequence length of 128.

BioBERT: Compared to BERT, BioBERT is further pre-trained on a large-scale biomedical dataset. It outperformed the original BERT model on biomedical named entity recognition, biomedical relation extraction, and biomedical question answering (Lee et al., 2020). In this study, we used the same BERT parameter settings as described above, except with the utilization of the BioBERT pre-trained model rather than the cased BERT-base one. We hypothesized that BioBERT would perform the best, followed by BERT and SVM.

4.2 Model Evaluation

To compare the performance of the three models, we mainly used macro-averaged precision, recall, and F1-score as evaluation measures. Since our ultimate goal is to retrieve health advice, we also reported individual precision, recall and F1 scores for each advice category.

Table 3 shows the model performance with a stratified 5-fold cross validation on the annotated corpus. Consistent with our hypothesis, BioBERT performed the best by all measures, achieving a macro-F1 score of 0.933. BERT's performance was slightly lower than BioBERT with a score of 0.918, indicating a modest benefit of domain-specific pretraining. Since both models outperformed the baseline SVM model (0.833) with wide margin, it is evident that the transformer-based method is a better choice for this task.

Table 4 shows that BioBERT performed well on all kinds of advice and study designs, ranging from 0.907 to 0.943 in macro-F1 score. This indicates a low risk of prediction bias against any category.

4.3 Error Analysis

Error analysis of misclassified cases showed that most of the prediction errors were caused by confusion between "no advice" and "weak advice".

	Advice Type	Precision	Recall	F1
SVM	None	0.868	0.927	0.897
	Weak	0.845	0.771	0.806
	Strong	0.852	0.748	0.797
	macro avg	0.855	0.815	0.833
BERT	None	0.949	0.943	0.946
	Weak	0.890	0.904	0.897
	Strong	0.910	0.912	0.911
	macro avg	0.917	0.920	0.918
BioBERT	None	0.963	0.951	0.957
	Weak	0.908	0.922	0.915
	Strong	0.917	0.941	0.928
	macro avg	0.929	0.938	0.933

Table 3: Model performance of detecting different types of health advice.

A further examination of these errors showed that some "no advice" sentences contained confounding cues like "the importance of" or "is suitable for" which provide implications for further study but not health behavior changes (see example 1). Sometimes a "no advice" sentence would use common advice cues such as "usefulness" and "applications" to describe study limitations instead of weak advice (see example 2), or the statement gives a vague recommendation without specifying the actions that should be taken (see example 3).

There was also some confusion between "no advice" and "strong advice". Some "no advice" sentences would use strong advice cues (e.g., "is necessary") or modal verbs (e.g., "should be") to describe research background or implications for follow-up studies (see examples 4 and 5), and thus confuse the prediction model.

Examples of prediction errors:

1. *"Therefore, this FFQ is suitable for the investigation of nutrient-disease associations in future."*
2. *"Its usefulness for this application is questionable."*
3. *"Our findings could inform health policy, guide prevention strategies, and justify the design and implementation of targeted interventions."*

	RCTs	Cross-Sectional	Case-Control	Retrospective	Prospective	*macro avg*
None	0.919	0.971	0.983	0.973	0.966	0.957
Weak	0.924	0.842	0.922	0.905	0.885	0.915
Strong	0.934	0.937	0.925	0.927	0.868	0.928
macro avg	0.926	0.917	0.943	0.935	0.907	0.933

Table 4: Performance of BioBERT on each study design (macro-F1).

4. *"Knowledge of molecular factors is necessary."*
5. *"Further investigations should address the rationale for the early detection and control of glucose fluctuation in the era of universal statin use for CAD patients."*

4.4 Extending Prediction Model to Unstructured Abstracts and Discussion Sections

We trained the BERT-based model using sentences extracted from conclusion subsections of structured abstracts. However, unstructured abstracts and full-text content, especially the discussion and conclusion sections, may also include health advice. To evaluate the models' generalizability to sentences in unstructured abstracts and full-text content, we randomly sampled 100 research papers that have unstructured abstracts and full-text access in PubMed Central (20 from each type of the five study designs). A total of 934 sentences from the abstracts and 3,932 sentences from the discussion/conclusion sections—which will be referred to as *discussion sections* for brevity—were manually annotated as "strong advice", "weak advice", or "no advice".

Directly applying the BioBERT prediction model. We applied BioBERT, the best-performing model, to detect health advice in each of the above sentences. The result is presented in Tables 5 and 6. The result shows lower precision scores in both unstructured abstracts and discussion sections, but the recalls are comparable to that in structured abstracts (the training corpus). This means the prediction model is equally effective at retrieving health advice in unstructured abstracts and discussion sections, but more non-advice sentences were included in the result as "false positive" predictions.

Error analysis shows that the false positive predictions were mainly caused by non-advice sentences that describe study background, motivation and prior study implications. These non-advice sentences are highly similar to advice sentences linguistically. This error pattern is actually the same as the errors in the training data. The main reason for the increased error rate is that these confusing sentences appear more often in unstructured abstracts and discussion sections. The human annotations of this sample showed that "no advice" accounted for 95.3% of the 934 unstructured abstract sentences and 92.4% of the 3,932 discussion sentences, compared to 59.8% for the conclusion subsections in structured abstracts.

Unstructured Abstracts				
Directly applying the fine-tuned BioBERT model				
Advice	Precision	Recall	F1	Cases
None	0.998	0.962	0.979	890
Weak	0.519	0.964	0.675	28
Strong	0.625	0.938	0.750	16
macro avg	0.714	0.955	0.801	934
After applying a simple filtering rule				
Advice	Precision	Recall	F1	
None	0.997	0.990	0.993	
Weak	0.765	0.929	0.839	
Strong	0.938	0.938	0.938	
macro avg	0.900	0.952	0.923	

Table 5: Performance improves on the unstructured abstracts when we apply a simple filtering rule to post-process prediction results.

Improving performance on unstructured abstracts. For unstructured abstracts, since health advice only occurs after a result description, which is near the end, a simple improvement is to assume all sentences in the first half are non-advice. Using this location-based filtering technique, the prediction model's precision improves to 0.900 (Table 5), comparable to that in the training data.

Improving performance on discussion sections with data and feature augmentation. Compared to that in unstructured abstracts, the distribution of health advice in discussion sections is more varied. As Fig. 1 shows, although health advice, especially strong advice, tends to occur in the second half of discussion sections, 29.3% of 297 advice sentences occurs in the first half, indicating that even an op-

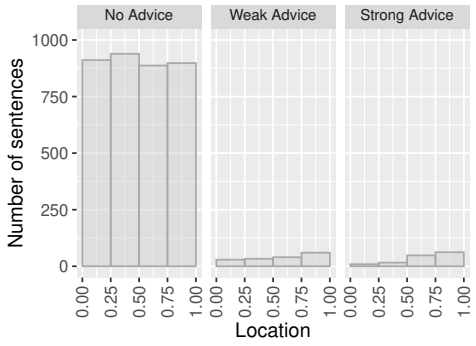

Figure 1: Distribution of health advice in discussion sections (calculated by number of sentences).

timal location filter would miss nearly a third of health advice sentences.

To improve the BioBERT model's precision performance on the discussion sentences, we proposed to use two techniques: (1) augment the training data, and (2) add language-style features.

First, augment the training data. Since we have annotated 3,932 discussion sentences—a considerable number of annotations, we can further fine-tune the BioBERT prediction model by utilizing them. Specifically, for each fold in the 5-fold cross-validation evaluation, we added 80% of discussion sentences to the original 6,000 training sentences, and then test the newly fine-tuned model performance on the remaining 20%.

Second, add language-style features. We observed that our model has not captured certain language-style markers that can effectively distinguish advice and non-advice sentences in discussion sections. One most common language-style marker is whether a sentence uses past tense. Advice sentences do not use past tense because of its imperative mood. In comparison, the following non-advice sentence used past tense, despite using advice-like cues "to ensure": *"We took great care to ensure adequate training of the neuropsychological evaluators at each site, and we monitored quality of test administration, scoring, and data entry on an ongoing basis."*

Another common marker is whether a sentence cites other studies. Advice-like sentences that contain citations are often citing advice from other studies, while our goal is to identify advice given by the authors in the current study rather than advice given in prior studies. For example, *"NMDA receptor antagonists such as ketamine or magnesium have been suggested for postoperative pain management [22,23]."*

Discussion Sections					
Directly applying the fine-tuned BioBERT model					
	Advice	Precision	Recall	F1	Cases
	None	0.997	0.950	0.973	3635
	Weak	0.537	0.988	0.696	162
	Strong	0.696	0.881	0.778	135
	macro avg	0.743	0.940	0.815	3932
After fine-tuning models with data and feature augmentation					
	Advice	Precision	Recall	F1	
Data augmentation only	None	0.991	0.977	0.984	
	Weak	0.708	0.883	0.786	
	Strong	0.793	0.852	0.821	
	macro avg	0.831	0.904	0.864	
+ Data source	None	0.987	0.987	0.987	
	Weak	0.781	0.815	0.798	
	Strong	0.875	0.830	0.852	
	macro avg	0.881	0.877	0.879	
+ Data source + Has citation	None	0.989	0.986	0.988	
	Weak	0.806	0.846	0.825	
	Strong	0.833	0.852	0.842	
	macro avg	0.876	0.895	0.885	
+ Data source + Has citation + Past tense	None	0.991	0.990	0.990	
	Weak	0.827	0.883	0.854	
	Strong	0.892	0.859	0.875	
	macro avg	0.903	0.911	0.907	

Table 6: Performance improves on the discussion sections when we fine-tune BioBERT models with data and feature augmentation.

To add the language-style markers into our model, we augmented the BERT input (a single sentence) with three "binary" features: (1) data source: whether a sentence is from a structured abstract or a discussion section, (2) citation: whether a sentence contains a citation, and (3) past tense: whether a sentence uses past tense.

When integrating the above features into BERT models, we used the special BERT mark [SEP] to concatenate the features with the original sentence, in the following form:

data source [SEP] citation [SEP] past tense [SEP] sentence

For example, a sentence from a discussion section that uses past tense but does not cite other studies will be represented as:

discussion [SEP] No [SEP] Yes [SEP] sentence

All sentences from structured abstracts are represented as:

structured abstract [SEP] [SEP] [SEP] sentence

With the above setting, we then ran a 5-fold cross-validation evaluation on the new method with augmented training data and added language-style features. The result in Table 6 shows that the augmented training data resulted in a significant im-

provement in macro-F1 score, from 0.815 to 0.864. The added language-style features further improved the F1 score to 0.907.

5 Identifying Health Advice on COVID-19 Treatment: A Case Study on Hydroxychloroquine (HCQ)

In this section we apply the health advice prediction model to research articles in the LitCovid corpus, in order to find health advice on specific treatment options for COVID-19. This is a case study to demonstrate the model's usefulness for retrieving health advice for a specific medical topic, especially when used in combination with existing health information services like LitCovid.

LitCovid is a corpus curated by NIH that includes all COVID-19 related research publications in PubMed. LitCovid has organized the research papers by topics, such as "transmission", "diagnosis", "prevention", and "treatment". LitCovid further tagged all chemicals that were studied, and assigned normalized terms for chemicals with multiple names, such as the MeSH Unique ID D006886 for "Hydroxychloroquine" and its 40+ alternative names like "HCQ" and "(hydroxy)chloroquine sulfate". We downloaded the LitCovid corpus on 04/30/2021 that includes 126,000 research papers. Using the MeSH ID of HCQ, we retrieved 3,400 HCQ-related papers with 10,000 sentences tagged with HCQ in abstracts and discussion sections.

These sentences were then sent to our prediction model to identify HCQ-related health advice. In the prediction result we found 605 strong advice sentences and 815 weak ones. The advice ranges from recommendation to use (see example 1 below), advice on doses and usage (example 2), cautions and warnings on treating patients with certain conditions (example 3), and objection to use (example 4).

Examples of health advice regarding HCQ:

1. *"We therefore recommend that COVID-19 patients be treated with hydroxychloroquine and azithromycin to cure their infection and to limit the transmission of the virus to other people in order to curb the spread of COVID-19 in the world."*
2. *In order to meet predefined HCQ exposure target, HCQ dose may need to be reduced in young children, elderly subjects with organ impairment and/or coadministration with a strong CYP2C8/CYP2D6/CYP3A4 inhibitor, and be increased in pregnant women."*
3. *"Additionally, hypoglycemia must be looked for in patients with diabetes especially with concurrent use of chloroquine/HCQ and lopinavir/ritonavir."*
4. *"Taken together, HCQ should not be used in prophylaxis against COVID-19."*

Summarizing the health advice by opinions and themes is beyond the scope of this study. However it indicates a future direction to build advanced information navigation tools to better assist researchers in retrieving key information from a large volume of literature, especially during public health crises like COVID, when literature is fast growing, evidence might be conflicting, and actionable health advice is much needed. For example, when more metadata like publication dates and study designs are available, health advice could be sorted by chronological order and the strength of study designs, to illustrate when a new piece of advice is given and how reliable it is. The above example 1 was given by Gautret et al. (2020), a widely cited and reported study for using HCQ in treatment of COVID-19. The study was also criticized for lack of randomization in the study design. Other NLP tools such as claim and stance classification (e.g. Walker et al., 2012; Ferreira and Vlachos, 2016; Li et al., 2017; Yu et al., 2019; Kilicoglu et al., 2019) may further aggregate the health advice by supporting HCQ use or not. None of these functions are available in current health information services like LitCovid, but could be built based on our health advice detection model and thus benefit health researchers and practitioners in the future.

6 Conclusion

In this study, we developed a high-performing NLP model that can detect weak and strong health advice from abstracts and discussion sections in medical research publications. We further developed a case study, in which we applied this model to retrieve health advice regarding the use of HCQ for COVID-19 treatment from the LitCovid data hub. The case study demonstrated that this health advice prediction model can be combined with existing health information service systems to provide more convenient navigation of a large volume of health literature. If further combined with other NLP tools, such as claim and stance classification,

the health advice service would be able to compare and summarize the evidence strength of recommendations for or against certain policies or treatments. The prediction model may be further extended to detect exaggerated health advice in science communication by comparing advice given in research papers against its counterparts in press releases, news articles, and social media posts (Yu et al., 2020). In future work, we will extend health advice identification to news and social media.

Our annotated corpus and code are available at https://github.com/junwang4/detecting-health-advice.

7 Ethics Statement

We would like to address the following ethics issues relevant to this study.

- This NLP model is designed to identify sentences that provide health advice in medical literature. However, this model cannot verify whether a piece of health advice is valid or not.
- As discussed in the introduction section and the case study in Section 5, health advice given by individual research papers may lack sufficient evidence or be outdated, and thus requires further verification by health professionals before being recommended for clinical use.
- Researchers often write for professional audiences, and thus may have provided health advice intended for health professionals instead of the general public. Furthermore, the interpretation of health advice may also require more context than a sentence alone. Therefore, average users are urged to discuss with their doctors whether to follow a piece of health advice found by this NLP model.
- For the same reason, when incorporating this NLP model in real-world applications, the application developers should provide a function to flag or remove inaccurate or outdated health advice upon requests from authors and health experts.
- Although this NLP model achieves a high prediction accuracy, false positive and false negative predictions may still occur. While the false positive predictions (non-advice sentences in the result) may just be a nuisance, the false negative predictions (missed health advice) may cause misunderstandings if the model is used for the purpose of retrieving all health advice. Users should be trained to understand that the model does not provide a perfect recall.

Acknowledgement

This research is supported by the US National Science Foundation under grant 1952353, the Microsoft Investigator Fellowship program, and the Syracuse University CUSE Grant. We thank Dr. Aesoon Park from Department of Psychology for discussions on concept definitions in the annotation schema. Thanks to Fatima Dobani, Aatish Suman, and Raj Desai for their help with annotation, and Albert Wang for proofreading the manuscript. Special thanks to the reviewers for insightful feedback on writing the ethics statement.

References

Lloyd B. Anderson. 1986. Evidentials, paths of change, and mental maps: Typologically regular asymmetries. In Wallace L. Chafe and Johanna Nichols, editors, *Evidentiality: The linguistic coding of epistemology*, pages 273–312. Norwood, NJ: Ablex Publishing Corporation.

John Langshaw Austin. 1975. *How to do things with words*. Harvard University Press.

Rahul Banerjee and Vinay Prasad. 2020. Are observational, real-world studies suitable to make cancer treatment recommendations? *JAMA network open*, 3(7):e2012119.

Caroline Brun and Caroline Hagege. 2013. Suggestion mining: Detecting suggestions for improvement in users' comments. *Research in Computer Science*, 70:199–209.

Tobias Cabanski. 2019. DS at SemEval-2019 Task 9: From Suggestion Mining with neural networks to adversarial cross-domain classification. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1192–1198.

Qingyu Chen, Alexis Allot, and Zhiyong Lu. 2020. Keep up with the latest coronavirus research. *Nature*, 579(7798):193–193.

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

Cleo Condoravdi and Sven Lauer. 2012. Imperatives: Meaning and illocutionary force. *Empirical issues in syntax and semantics*, 9:37–58.

Peter Cummings. 2007. Policy recommendations in the discussion section of a research article. *Injury Prevention*, 13(1):4–5.

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.

Li Dong, Furu Wei, Yajuan Duan, Xiaohua Liu, Ming Zhou, and Ke Xu. 2013. The automated acquisition of suggestions from tweets. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, pages 239–245.

Brandon Fan, Weiguo Fan, Carly Smith, and Harold Skip Garner. 2020. Adverse drug event detection and extraction from open data: A deep learning approach. *Information Processing & Management*, 57(1):102131.

William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 1163–1168.

Margaret Fry and Jutharat Attawet. 2018. Nursing and midwifery use, perceptions and barriers to evidence-based practice: a cross-sectional survey. *International journal of evidence-based healthcare*, 16(1):47–54.

Philippe Gautret, Jean-Christophe Lagier, Philippe Parola, Van Thuan Hoang, Line Meddeb, Morgane Mailhe, Barbara Doudier, Johan Courjon, Valérie Giordanengo, Vera Esteves Vieira, Hervé Tissot Dupont, Stéphane Honoré, Philippe Colson, Eric Chabrière, Bernard La Scola, Jean-Marc Rolain, Philippe Brouqui, and Didier Raoult. 2020. Hydroxychloroquine and azithromycin as a treatment of COVID-19: results of an open-label non-randomized clinical trial. *International Journal of Antimicrobial Agents*, 56(1):105949.

Andrew B Goldberg, Nathanael Fillmore, David Andrzejewski, Zhiting Xu, Bryan Gibson, and Xiaojin Zhu. 2009. May all your wishes come true: A study of wishes and how to recognize them. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 263–271.

Lawrence W Green, Russell E Glasgow, David Atkins, and Kurt Stange. 2009. Making evidence from research more relevant, useful, and actionable in policy, program planning, and practice slips "twixt cup and lip". *American journal of preventive medicine*, 37(6):S187–S191.

Romana Haneef, Clement Lazarus, Philippe Ravaud, Amélie Yavchitz, and Isabelle Boutron. 2015. Interpretation of results of studies evaluating an intervention highlighted in google health news: a cross-sectional study of news. *PLoS ONE*, 10(10):e0140889.

Tom Hope, Jason Portenoy, Kishore Vasan, Jonathan Borchardt, Eric Horvitz, Daniel Weld, Marti Hearst, and Jevin West. 2020. SciSight: Combining faceted navigation and research group detection for COVID-19 exploratory scientific search. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 135–143, Online. Association for Computational Linguistics.

Ken Hyland. 1994. Hedging in academic writing and eaf textbooks. *English for specific purposes*, 13(3):239–256.

Ken Hyland. 1995. The author in the text: Hedging scientific writing. *Hong Kong papers in linguistics and language teaching*, 18:33–42.

Ken Hyland. 1996. Writing without conviction? hedging in science research articles. *Applied linguistics*, 17(4):433–454.

Ken Hyland. 1998a. Boosting, hedging and the negotiation of academic knowledge. *Text & Talk*, 18(3):349–382.

Ken Hyland. 1998b. *Hedging in Scientific Research Articles*, volume 54. John Benjamins Publishing.

Maria Kabisch, Christian Ruckes, Monika Seibert-Grafe, and Maria Blettner. 2011. Randomized Controlled Trials. *Dtsch Arztebl International*, 108(39):663–668.

Halil Kilicoglu, Zeshan Peng, Shabnam Tafreshi, Tung Tran, Graciela Rosemblat, and Jodi Schneider. 2019. Confirm or refute?: A comparative study on citation sentiment classification in clinical research publications. *Journal of biomedical informatics*, 91:103123.

Robin Lakoff. 1972. Language in context. *Language*, 48(4):907–927.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Yingya Li, Jieke Zhang, and Bei Yu. 2017. An NLP analysis of exaggerated claims in science news. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 106–111, Copenhagen, Denmark. Association for Computational Linguistics.

Jiaxiang Liu, Shuohuan Wang, and Yu Sun. 2019. Olenet at semeval-2019 task 9: Bert based multi-perspective models for suggestion mining. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1231–1236.

CJ Mann. 2003. Observational research methods. research design ii: cohort, cross sectional, and case-control studies. *Emergency medicine journal*, 20(1):54–60.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica*, 22(3):276–282.

M Hassan Murad, Noor Asi, Mouaz Alsawas, and Fares Alahdab. 2016. New evidence pyramid. *BMJ Evidence-Based Medicine*, 21(4):125–127.

Ilana Mushin. 2001. *Evidentiality and Epistemological Stance: Narrative Retelling*. John Benjamins Publishing.

Greg Myers. 1989. The pragmatics of politeness in scientific articles. *Applied linguistics*, 10(1):1–35.

Takeo Nakayama, Nobuko Hirai, Shigeaki Yamazaki, and Mariko Naito. 2005. Adoption of structured abstracts by general medical journals and format for a structured abstract. *Journal of the Medical Library Association*, 93(2):237.

Sapna Negi. 2016. Suggestion mining from opinionated text. In *Proceedings of the ACL 2016 Student Research Workshop*, pages 119–125, Berlin, Germany. Association for Computational Linguistics.

Sapna Negi and Paul Buitelaar. 2015. Towards the extraction of customer-to-customer suggestions from reviews. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2159–2167, Lisbon, Portugal. Association for Computational Linguistics.

Sapna Negi, Tobias Daudert, and Paul Buitelaar. 2019. SemEval-2019 task 9: Suggestion mining from online reviews and forums. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 877–887, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Cheoneum Park, Juae Kim, Hyeon-gu Lee, Reinald Kim Amplayo, Harksoo Kim, Jungyun Seo, and Changki Lee. 2019. ThisIsCompetition at SemEval-2019 task 9: BERT is unstable for out-of-domain samples. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1254–1261, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Roma Patel, Yinfei Yang, Iain Marshall, Ani Nenkova, and Byron Wallace. 2018. Syntactic patterns improve information extraction for medical search. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 371–377, New Orleans, Louisiana. Association for Computational Linguistics.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *The Journal of machine Learning research*, 12:2825–2830.

Ivan Barry Pless. 2007. Unfinished business. *Injury Prevention*, 13(3):145–146.

Vinay Prasad, Joel Jorgenson, John PA Ioannidis, and Adam Cifu. 2013. Observational studies often make clinical practice recommendations: an empirical evaluation of authors' attitudes. *Journal of clinical epidemiology*, 66(4):361–366.

J. Ramanand, Krishna Bhavsar, and Niranjan Pedanekar. 2010. Wishful thinking - finding suggestions and 'buy' wishes from product reviews. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 54–61, Los Angeles, CA. Association for Computational Linguistics.

Jonathon Read, Erik Velldal, Marc Cavazza, and Gersende Georg. 2016. A corpus of clinical practice guidelines annotated with the importance of recommendations. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1724–1731, Portorož, Slovenia. European Language Resources Association (ELRA).

Philip Resnik, Katherine E. Goodman, and Mike Moran. 2020. Developing a curated topic model for COVID-19 medical research literature. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Frederieke Schaafsma, Jos Verbeek, Carel Hulshof, and Frank Van Dijk. 2005. Caution required when relying on a colleague's advice; a comparison between professional advice and evidence from the literature. *BMC health services research*, 5(1):1–5.

John R Searle. 1976. A classification of illocutionary acts. *Language in society*, 5(1):1–23.

Jae W Song and Kevin C Chung. 2010. Observational studies: cohort and case-control studies. *Plastic and reconstructive surgery*, 126(6):2234–2242.

Sharon Straus and R Bryan Haynes. 2009. Managing evidence-based knowledge: the need for reliable, relevant and readable resources. *CMAJ*, 180(9):942–945.

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

Marilyn Walker, Pranav Anand, Rob Abbott, and Ricky Grant. 2012. Stance classification using dialogic properties of persuasion. In *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 592–596.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Nancy Xin Ru Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The COVID-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

Alfan Farizki Wicaksono and Sung-Hyon Myaeng. 2012. Mining advices from weblogs. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 2347–2350.

Alfan Farizki Wicaksono and Sung-Hyon Myaeng. 2013. Automatic extraction of advice-revealing sentences for advice mining from online forums. In *Proceedings of the seventh international conference on Knowledge capture*, pages 97–104.

Peace Ossom Williamson and Christian IJ Minter. 2019. Exploring PubMed as a reliable resource for scholarly communications services. *Journal of the Medical Library Association: JMLA*, 107(1):16.

MK Wilson and IG Chestnutt. 2016. Prevalence of recommendations made within dental research articles using uncontrolled intervention or observational study designs. *Journal of Evidence Based Dental Practice*, 16(1):1–6.

Bei Yu, Yingya Li, and Jun Wang. 2019. Detecting causal language use in science findings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4664–4674, Hong Kong, China. Association for Computational Linguistics.

Bei Yu, Jun Wang, Lu Guo, and Yingya Li. 2020. Measuring correlation-to-causation exaggeration in press releases. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4860–4872, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ping Yue, Jin Wang, and Xuejie Zhang. 2019. YNU-HPCC at SemEval-2019 task 9: Using a BERT and CNN-BiLSTM-GRU model for suggestion mining. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1277–1281, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Sijia Zhou and Xin Li. 2020. Feature engineering vs. deep learning for paper section identification: Toward applications in chinese medical literature. *Information Processing & Management*, 57(3):102206.