Knowledge discovery from patient forums: gaining novel medical insights from patient experiences

Dirkson, A.R.

4

CONVERSATION-AWARE FILTERING OF RELEVANT MESSAGES

Edited from: **Anne Dirkson**, Suzan Verberne & Wessel Kraaij (2020). *Conversation-aware Filtering from Online Patient Forums.* Proceedings of the Fifth Social Media Mining for Health Applications (SMM4H) Workshop @ COLING.

In this chapter, we explore the benefit of exploiting conversational context for filtering posts relevant to a specific medical topic, such as adverse drug events. The filtering of relevant posts from a larger corpus is a commonly used first step towards knowledge extraction from social media.

Previous approaches to NLP tasks on online patient forums have been limited to single posts as units, thereby neglecting the overarching conversational structure. Here, we experiment with two approaches to add conversational context to the state-of-the-art BERT model: a sequential CRF layer and manually engineered features.

Although neither approach can outperform the F_1 score of the baseline, we find that adding a sequential layer improves precision for all target classes, whereas adding a non-sequential layer with manually engineered features leads to a higher recall for two out of three target classes. Thus, depending on the end goal, conversation-aware modeling may be beneficial for identifying relevant messages. We hope our findings encourage other researchers in this domain to move beyond studying messages in isolation towards more discourse-based data collection and classification.

4.1. INTRODUCTION

In the past decade, social media has emerged as a source of valuable knowledge in the health domain [116], for instance during the COVID-19 pandemic [159, 269]. In order to use social media to answer a medical question, it is necessary to identify posts on the forum that are relevant to the question at hand e.g., posts mentioning adverse drug events (ADEs) [183], personal experiences [87], medication abuse [267] or medical misinformation [158]. This filtering step is often the first step of the analysis pipeline. In this chapter, we will refer to this specific type of filtering as relevance classification.

Previous automatic methods for medical relevance classification generally consider posts as units without context, thereby ignoring any information that can be gained from the conversational context. One example of such an approach is the recent shared task on ADE relevance classification [337]. Yet, including the conversational context may prove beneficial to relevance classification, as responses in a thread often relate to previous responses. For example, responses to a question or comment about a specific side effect are likely to also concern this side effect. To test this hypothesis, we investigate how positive labels are distributed across and within conversational threads.

At present, only one study into medical relevance classification has included some engineered features to capture aspects of the conversational structure [158]. However, as this study includes only two discourse-based features, the effect of including manually engineered features that capture conversational structure is still largely unknown for relevance classification tasks.

Furthermore, including the relation between posts on a discourse level may also be able to improve classifier performance. Each post serves a conversational function in a dialogue, e.g., a question, explanation or statement [14]. These functions are called *dialogue acts* [288]. We have not found any study that included dialogue acts as features for medical relevance classification.

As an alternative to using manually engineered features, conversational threads can also be modeled with a sequential model. This has proven beneficial in other fields such as rumor classification in social media discussions [354]. As of yet, the use of sequential models for medical relevance classification has also not been explored.

We address the following research questions in this chapter:

RQ1 To what extent can the addition of a sequential model on top of state-of-the-art non-sequential models improve medical relevance classification of social media data?

RQ2 To what extent can the addition of manually engineered features for conversational structure and discourse improve medical relevance classification?

We use two different data sets for answering our questions. In our current research, we are particularly interested in discovering ADEs in online discussions. We have collected and annotated a data set about this topic. Since this data set is new, no other results have been published for it. We therefore use one other data set for evaluating our methods: the medical misinformation data set by Kinsora et al. [158]. We use a BERT-based model as baseline. BERT models constitute the current state of the art for most NLP tasks [84] including ADE relevance classification [337].

In the following section, we will elaborate on related work. Hereafter, we describe our methodology and data in Section 4.3 and 4.4 respectively. Finally, we present and discuss our results in Section 4.5 and 4.6.

4.2. RELATED WORK

The use of conversational structure for improving the performance of classifiers of social media posts is prevalent in the field of rumor classification [354] and related fields like disagreement detection [254]. Conversational structure has previously been exploited through (a) manually engineered features or (b) sequential classifiers.

The most commonly employed engineered features to model the conversational structure are the similarity to the previous message and to the thread in general [354]. In addition to these features, the current state-of-the-art model on a leading shared task for rumor stance classification (RumourEval-2019) uses the label of the previous message and the distance to the start of the thread [181]. In the health domain, the only study that employs manually engineered features for conversational structure is Kinsora et al. [158]. Specifically, they use the running count of positive labels and the distance to the previous positive label. In this study, we will employ the above features as well as expand upon them with additional discourse-related features.

Other studies have used sequential classifiers to model the discursive nature of social media, although according to Zubiaga et al. [354] this is “still in its infancy” (p. 276). Their comparison of various classifiers for rumor stance classification revealed that sequential classifiers outperform non-sequential classifiers overall. This is probably due to their ability to leverage information about sequential structure and preceding labels. Furthermore, Zubiaga et al. [354] found that sequential classifiers did not benefit from contextual features representing thread context (e.g., similarity to the source tweet) whereas non-sequential classifiers did. They speculate that sequential classifiers take the surrounding context into account implicitly. To see if this also holds true for relevance classification in medical social media, we will compare the addition of conversation-aware features to both sequential and non-sequential models.

4.3. METHODS

4.3.1. MODELS

CRF As a sequential model we use Conditional Random Fields (CRF). We train the models using the implementation in sklearn-crfsuite. L1 and L2 regularization parameters were tuned for each fold.

Linear SVM As a non-sequential counterpart, we use the sklearn implementation of Linear Support Vector Machines. The hyper-parameter C is tuned per fold with a grid of 10^{-3} to 10^{3} in steps of ×10.

DistilBERT As BERT model, we opt for DistilBERT (distilbert-base-uncased), which is a lighter, more computationally efficient variant of BERT [260]. We use the Huggingface implementation [339] with the wrapper ktrain [195] to train our models. The initialization seed is set to 1. We use the default learning rate of 5×10^{-5} and tune the number of epochs (3 or 4) per fold.

Ensemble models To investigate the benefit of adding a sequential model on top of the DistilBERT model, we experiment with a blending-based ensemble method: we input the raw confidence scores from DistilBERT for each label as features in a CRF model (i.e., CRF + BERTpred). We create an equivalent non-sequential baseline by using the same approach with an SVM (i.e., SVM + BERTpred).

4.3.2. FEATURE ANALYSIS

To explore the benefit of manually engineered features that capture thread context, we use step-wise greedy forward feature selection using the features in Table 4.1. For each step-wise iteration, we select the best feature to add to the model until the F_1 score no longer improves. We use 10-fold cross-validation in which for each fold features are selected on the development data (10%) and tested on a held-out test set (10%). For a fair comparison, we keep the folds and hyper-parameters the same as for the respective base model. Since the label distribution features could leak information, we omit these gold annotated features for evaluation. Instead, we perform an initial run without these features and use the resulting predictions to calculate them for the final evaluation.

4.3.3. MODEL COMPARISON

We used 10-fold cross-validation in all experiments. Instead of splitting per message, we split on whole discussion threads to ensure possible dependencies between posts do not bias the outcome. Statistical comparisons of model performance are done using Wilcoxon signed rank tests across the 10 folds. To avoid the multiple testing problem, we only compare the three best models, namely those with the highest F_1 score, precision, and recall, with the BERT baseline.

4.4. DATA

Data collection At present, there is only one publicly available medical relevance classification data set that includes the conversational structure: the Medical Misinformation Data set [158]. It is based on MedHelp data and annotated for the presence of misinformation. We collected a second data set from a Facebook group of Gastrointestinal Stromal Tumor (GIST) patients. We selected 527 discussions based on their likelihood to contain an ADE: We selected the threads that contained (1) at least one drug name according to a match with RxNorm [314] and (2) a high percentage of posts in which authors shared experiences. The latter criterion was included since sharing that you had an ADE is an example of experience sharing. To estimate this, we used the classifier described in Chapter 3. According to our classifier, at least 80% of the posts within each selected thread is a personal experience. Due to privacy issues and ownership of the data by the GIST International patient organization, we are not able to share this data set at present. See Table 4.2 for more details on the data sets.

[1] We opt for USE instead of BERT embeddings, as cosine similarity cannot be applied directly to BERT embeddings

Feature type	Name	Description	Explanation (if applicable)
Local	+Emb	Sentence Vectors	We use Universal Sentence Encoder (USE) [57] to encode sentences into 512 dimensional vectors based on pre-trained embeddings so their cosine similarity (normalized between 0 and 1) approximates their semantic similarity.[1]
	+BERTpred	distilBERT predictions	The raw confidence scores for each label
Relational	+PrevSim	Similarity to previous message	Similarity is calculated using the USE sentence vectors
	+ThreadSim	Thread similarity	Similarity to USE sentence vector of all other posts in the thread combined into one vector
Positional	+Dist	Absolute distance from start of thread	
Label distribution	+PrevLbl	Label of previous post	We use the true labels for training and the predicted labels for testing for all label distribution features.
	+CountPos	Absolute running count of preceding positive labels in thread	
	+CountNeg	Absolute running count of preceding negative labels in thread	
	+RelPos	Percentage of preceding positive labels	
	+DistPos	Distance from previous positive label	
	+DistNeg	Distance from previous negative label	
Discourse	+DA	Dialogue act of post	Dialogue acts are calculated using the Dialogue Act tagger as trained by Tortoreto et al. [299]
	+PrevDA	Dialogue act of previous post	

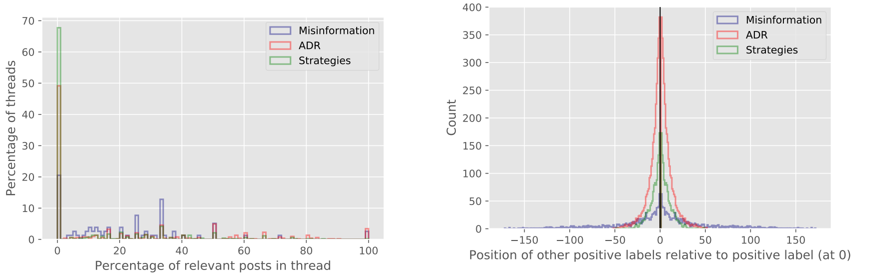
Table 4.1: Manually engineered features to model conversational structure

Data annotation Following a pilot annotation round, the data was annotated by the first author and three patients for the presence of ADEs and coping strategies for dealing with ADEs (hereafter also called: Strategies) using an annotation guideline.[2] The pairwise inter-annotator agreement was substantial for ADEs (mean κ =0.71) and moderate for Coping Strategies (mean κ =0.54).

[2] Available at: `https://github.com/AnneDirkson/ConversationAwareFiltering`

Data set	Target	#Posts	#Discussions	Median length	% Positive
Medical Misinformation Dataset [158]	Misinformation	1,566	78	8.0	15.0 %
ADE Discussions (In-house)	Adverse Drug Event (ADE) & Coping Strategies	4,195	527	6	22.9 % & 12.3%

Table 4.2: Statistics on the data sets. The ADE Discussions data set has two target classes.



(a) Distribution of target posts *across* threads

(b) The relative position of target posts to each other *within* threads

Figure 4.1: Distribution of the target class (i.e., positively labeled posts)

4.5. RESULTS

4.5.1. DISTRIBUTION OF THE TARGET CLASS IN THE DISCUSSION THREADS

As visualized in Figure 4.1a, the target class is not distributed equally across the discussion threads for any of the data sets; There appear to be many threads with few or no target posts. According to z-tests, the distribution is significantly different from normal. An inspection of the relative position of target posts *within* discussion threads reveals that the target posts also cluster together (see Figure 4.1b). The probability that the post after a target post is also a target post is 27% for Misinformation and 40% and 34% for ADEs and Coping Strategies respectively. These probabilities are higher than is to be expected based on the percentage of positively labeled posts (see Table 4.2). Thus, it appears that the conversational structure is indeed related to the probability of a post being relevant and consequently incorporating conversational structure or discourse may be able to improve the performance of relevance classifiers.

4.5.2. MODEL COMPARISON

The results of model evaluation are presented in Table 4.3. It appears that neither the addition of a sequential layer nor manual features can improve upon the F_1 score of the BERT model. Misinformation detection appears to be the exception, as any additional layer, sequential or not, outperforms the BERT baseline model. The highest overall

	Misinformation		
	F_1	P	R
BERT	0.366 ± 0.155	0.386 ± 0.154	0.396 ± 0.235
SVM+Emb	**0.478 ± 0.083**	0.492 ± 0.109	0.482 ± 0.111
+ *Features*	0.392 ± 0.089	0.457 ± 0.169	0.405 ± 0.156
CRF+Emb	0.424 ± 0.155	**0.565 ± 0.148**	0.352 ± 0.162
+*Features*	0.457 ± 0.137	0.557 ± 0.155	0.420 ± 0.167
SVM + BERTpred	0.443 ± 0.078	0.449 ± 0.082	0.479 ± 0.151
+*Features*	0.454 ± 0.070	0.449 ± 0.081	**0.492 ± 0.140**
CRF + BERTpred	0.434 ± 0.079	0.453 ± 0.100	0.447 ± 0.138
+*Features*	0.428 ± 0.078	0.435 ± 0.092	0.446 ± 0.126

	ADEs		
	F_1	P	R
BERT	**0.714 ± 0.034**	0.715 ± 0.038	0.718 ± 0.062
SVM+Emb	0.640 ± 0.054	0.673 ± 0.055	0.613 ± 0.069
+*Features*	0.610 ± 0.068	0.621 ± 0.087	0.624 ± 0.128
CRF+Emb	0.654 ± 0.059	0.710 ± 0.036	0.611 ± 0.086
+*Features*	0.638 ± 0.067	0.695 ± 0.037	0.601 ± 0.110
SVM + BERTpred	**0.714 ± 0.035**	0.724 ± 0.043	0.707 ± 0.056
+*Features*	0.677 ± 0.121	0.673 ± 0.164	**0.738 ± 0.103**
CRF+ BERTpred	**0.714 ± 0.038**	**0.728* ± 0.040**	0.704 ± 0.062
+*Features*	0.713 ± 0.039	0.726 ± 0.040	0.705 ± 0.060

	Strategies		
	F_1	P	R
BERT	**0.581 ± 0.060**	0.622 ± 0.087	**0.563± 0.111**
SVM+Emb	0.517 ± 0.101	**0.660 ± 0.111**	0.434 ± 0.111
+*Features*	0.502 ± 0.108	0.603 ± 0.137	0.453 ± 0.128
CRF+Emb	0.441 ± 0.134	0.597 ± 0.120	0.373 ± 0.151
+*Features*	0.512 ± 0.106	0.609 ± 0.110	0.462 ± 0.143
SVM+Bertpred	0.578 ± 0.059	0.632 ± 0.091	0.545 ± 0.089
+*Features*	0.561 ± 0.095	0.601 ± 0.146	0.552 ± 0.087
CRF + BERTpred	**0.581 ± 0.065**	0.629 ± 0.087	0.558 ± 0.115
+*Features*	0.573 ± 0.058	0.635 ± 0.090	0.539 ± 0.100

Table 4.3: Evaluation results of mean model performance over 10 folds. Features are selected through step-wise greedy feature selection. **<0.01 *<0.05

performance is attained by an SVM model based on USE sentence vectors (+Emb), which were specifically designed for representing whole sentences. Perhaps sentence vectors perform better than BERT embeddings when the BERT model performs poorly ($F_1 = 0.366$). Additional research will be necessary to substantiate this.

Despite a lack of improvement in the F_1 score for the detection of ADEs and Strategies, an additional layer does seem to offer flexibility in tailoring the model towards a higher recall or precision. On the one hand, recall can be improved for two target classes by adding a non-sequential SVM layer with manual features to the BERT model. On the other hand, precision can be improved through the addition of a sequential CRF layer on top of BERT predictions for all target classes. Adding manually engineered features in addition to the sequential layer only improves the precision further for the detection of coping strategies. Our findings are thereby in line with Zubiaga et al. [354]. They speculated that sequential classifiers may take the surrounding context into account implicitly and therefore do not benefit from features representing thread context.

The only significant increase according to Wilcoxon signed rank tests is the precision for ADE detection. This may be related to the high variance between folds. Further research is necessary to validate these results and advance our understanding of how conversation-aware modeling can be best be used for relevance classification. We believe that this first study shows that this is a promising direction.

4.5.3. Analysis of selected features

The variation in which features are selected per fold is large. Manual inspection of the selected features shows that features relating to the distribution of labels in the thread are chosen most often, especially the running count of negative and positive labels in the thread (CountNeg, CountPos), and the label of the previous post (PrevLbl) (see Table 4.1). Features of this type may therefore be the most promising for future work. The number of features that is chosen is more consistent; On average, 1 or 2 of the 11 features are chosen.

To further explore why certain features are chosen, we compute the correlations between the target label and the manually engineered features and between the BERT predictions and the manually engineered features (see Figure 4.2). We find, firstly, that features relating to the label distribution indeed appear to correlate most strongly with the ground truth labels. Secondly, the correlation between these features and the BERT predictions is often equal to or stronger than the respective correlation to the ground truth. This might indicate that this variance is already captured by the BERT model and therefore manually engineered features have little to add to the baseline model.

4.6. Discussion

We find that the distribution of target posts across discussion threads is skewed and that within a conversational thread posts cluster together. Thus, our hypothesis that the probability of a target post occurring is related to the conversational structure appears valid.

In answer to **RQ1**, we find that adding a sequential CRF layer on top of a BERT model improves precision slightly, although only significantly so for ADE detection. In answer to **RQ2**, we find that the addition of manually engineered features representing

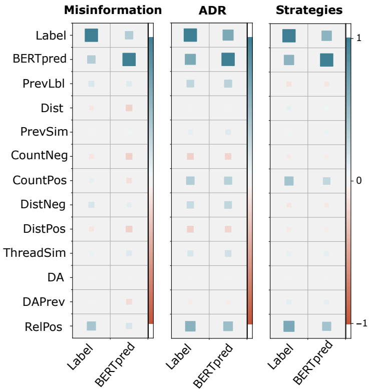


Figure 4.2: Correlation matrix of ground truth labels and BERT predictions with the manually engineered features. The size and colour of the squares corresponds to the strength of the correlation

thread context often does not aid performance. One consistent exception is when manually engineered features are combined with a non-sequential SVM layer on top of a BERT model. This combination can improve recall for all target classes, although not significantly so. An additional layer on top of a BERT model that is able to capture the thread context appears to offer flexibility in tailoring the model towards a higher recall or precision. In future work, we plan to investigate the benefit of including conversational context for other tasks such as concept normalization of ADEs.

For all data sets included in this study, a preselection of discussion threads was made prior to annotation to ensure a higher proportion of target posts. We expect that both sequential models and manually engineered features of thread context may prove more beneficial when such a preselection does not occur and the target class is even more imbalanced. Thus, our results may be an underestimation of the benefit of conversational context for finding 'needles in the haystack'.

Finally, our findings call into question the practice of splitting data into folds without taking the discussion context into account. In this study, we split the folds per discussion thread and we recommend others to consider doing so when dealing with multiple posts from the same thread, as neglecting to do so when there are dependencies between posts may bias model performance. This is especially important when threads contain duplicate posts.