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

Dirkson, A.R.

3

DETECTING PERSONAL EXPERIENCES

Edited from: **Anne Dirkson**, Suzan Verberne & Wessel Kraaij (2019), *Narrative Detection in Online Patient Communities.* Proceedings of Text2Story — Second Workshop on Narrative Extraction From Texts co-located with 41th European Conference on Information Retrieval (ECIR 2019). 21-28.

In this chapter, we discuss the extraction of messages containing the experiences of patients (hereafter called narratives) from patient fora. This subset will also include messages in which patients share their experiences with adverse drug events and may thereby aid in their extraction.

Prior to this study, the systematic detection and analysis of patient narratives was limited to a single study in which lower-cased words were used to identify narratives. In contrast, here we examine whether psycho-linguistic features or document embeddings could aid their identification. We also investigate which features distinguish narratives from other social media posts. Moreover, this study is the first to automatically identify the topics discussed in narratives on a patient forum.

We find that for the identification of patient narratives, character 3-grams outperform psycho-linguistic features and document embeddings. Additionally, we find that narratives are characterized by the use of past tense, health-related words and first-person pronouns, whereas non-narrative text is associated with the future tense, emotional support words and second-person pronouns. Topic analysis of the patient narratives uncovered fourteen different medical topics, ranging from tumor surgery to side effects. Future work will use these methods to extract experiential patient knowledge from social media.

3.1. INTRODUCTION

Nowadays, online patient forums are the main medium by which patients exchange their narratives. These narratives mainly recount their own experiences with their condition. As such, they contain experiential knowledge [38], defined as the knowledge that patients gain from their own experiences. In recent years, such experiential knowledge has increasingly been recognized as valuable and complementary to empirical knowledge [50]. Consequently, more health-related applications are making use of patient forum data, for instance to track public health trends [267] and to detect adverse drug events [266]. Experiential knowledge is also valuable for patients themselves: patients indicate that they strongly rely on experiences and information provided on patient forums [277]. This is especially true for patients with a rare disease, for which medical professionals often lack expertise and the number of studies is limited [15].

To understand the experiential knowledge on patient forums, forum posts that contain narratives must first be identified. As of yet, research into systematically distinguishing patient narratives on patient forums is limited to a single study on Dutch forum data [328], which uses words as only features. We expand upon this work using a different data set by examining whether document embeddings and psycho-linguistic features can improve the identification of patient narratives. We expect so, because these aggregated features are less dependent on individual terms, which may overlap significantly between narratives and factual statements about the same topic. Secondly, we explore how narratives differ from other types of posts by studying which features are influential in identifying narratives and which posts are classified incorrectly. Thirdly, we analyze how prevalent narratives are on a cancer patient forum and which topics these narratives discuss.

3.2. RELATED WORK

Narratives on patient forums have mainly been studied qualitatively (e.g., [325]). The automatic identification of narratives on a patient forum is limited to the study by Verberne et al. [328] on a Dutch cancer forum. They identified narratives with a F_1 of 0.911 using only the lower-cased words of the posts as features. They also found that various linguistic factors (1st person singular, 3rd person and negations) and psychological processes (social processes and religion) were correlated with the presence of narratives. These psycho-linguistic features were measured using the Linguistic Inquiry and Word Count (LIWC) method [297].

Additionally, research into self-reported adverse drug events (ADE) has led to the development of classifiers for differentiating between factual statements of ADE and personal experiences of ADE on social media [33, 217, 262]. However, these classifiers are highly specific and thus not suitable for identifying patient narratives in general.

Another closely related field is the classification of personal health mentions on social media, i.e., posts that mention a person who is affected as well as their specific condition, such as: 'my granddad has Alzheimer's'. Presently, only two studies have investigated this task. The first by Lamb et al. [169] focused on separating flu awareness from actual flu reports on social media. More recently, Karisani and Agichtein [152] introduced WESPAD, a classifier for personal health mentions, which attains state-of-the-

art performance for seven different health domains including stroke, depression, and flu infection. Nonetheless, a personal health mention alone is not sufficient to consider the post a narrative, and thus these classifiers are also inadequate for our purpose.

3.3. METHODS

3.3.1. DATA

Our data consists of an open, international Facebook forum for patients with Gastrointestinal Stromal Tumor (GIST)[1]. It is moderated by GIST Support International and consists of 36,722 posts with a median length of 20 tokens.

3.3.2. PREPROCESSING

The data was lower-cased and tokenized with NLTK. Due to the noisy nature of user-generated content, especially in the spelling of medical terms, we applied a tailored preprocessing pipeline[2] to our data. Firstly, an existing normalization pipeline for social media[3] [261] was used to normalize tokens to American English and to expand generic abbreviations used on social media. Hereafter, domain-specific abbreviations were expanded with a lexicon of 42 non-ambiguous abbreviations, generated based on 1000 posts and annotated by a domain expert and the first author. Spelling mistakes were detected using a combination of relative frequency and edit distance to possible candidates and corrected using weighted Levenshtein distance. Correction candidates were derived from the corpus itself. Drug names were normalized using the RxNorm database [314]. Non-English posts were removed using langid [190]. Punctuation was removed, but stop words were not, as we expect function words to play a role in the expression of narratives.

3.3.3. SUPERVISED CLASSIFICATION

Manual annotation of example data We randomly selected 1050 posts for annotation. The annotators were asked to indicate per message whether it contains a personal experience. They were not provided with its context. Personal experiences did not need to be about the author but could be about someone else. This definition was based on earlier work by Verberne et al. [328] and van Uden-Kraan et al. [324]. The first 50 posts were annotated individually by the first author and another PhD student to improve the annotation guidelines.[4] The remaining 1000 posts were divided equally into six sets of 200 posts, with 40 posts (20%) overlapping between all sets. The overlap was used to calculate the pairwise Cohen's kappa. There were seven annotators in total: six PhD students and one GIST patient. Each sample was assigned to an annotator, apart from one sample which was divided between two PhD students. To be able to include the overlapping sample in the classification, we opted to use the annotations of the GIST patient for these 40 posts.[5]

[1] `https://www.facebook.com/groups/gistsupport/`
[2] The preprocessing scripts can be found at: `https://github.com/AnneDirkson/LexNorm`
[3] `https://bitbucket.org/asarker/simplenormalizerscripts`
[4] The annotation guidelines can be found at: `https://github.com/AnneDirkson/NarrativeFilter`
[5] The annotated data is available upon request in order to protect the privacy of the patients

Feature sets Four feature sets were derived from the text data: word unigrams, character n-grams (using the CountVectorizer function in sklearn), psycho-linguistic features, and document embeddings. For both word unigrams and character n-grams, we investigated whether TF-IDF weighting would improve performance compared to raw counts. Additionally, we explored whether stemming or lemmatizing the data prior to extracting the unigrams could improve performance. Psycho-linguistic features were based on the LIWC 2015 [297]. Punctuation categories were discarded, resulting in 82 LIWC features in total. LIWC is a well-known method for investigating psychological processes in text and includes both linguistic (e.g., first-person pronouns) and psychological categories (e.g., positive emotions). The last feature set consisted of document embeddings: a doc2vec model [172] was trained on the labeled training data for each fold in the cross-validation. We combine a distributed memory model with a distributed bag of words model, as recommended by Le and Mikolov [172]. We also attempted to train document embeddings first on the unsupervised data and then retrain on the supervised data, but this led to nonsensical classification features.

Supervised classification algorithms Classifiers were evaluated separately for each feature set. We ignored all posts that had been left empty by the annotator (the annotator chose neither yes nor no): three posts were ignored for this reason. For word unigrams, character n-grams, and psycho-linguistic features, we compared four sklearn classification algorithms: Multinomial Naive Bayes (MNB), linear Support Vector Classification (LinearSVC), Stochastic Gradient Descent (SGD) with log loss, and K Nearest Neighbors (KNN). These were chosen according to the following criteria: (1) known to perform well on text data, (2) recommended for small data sets, and (3) able to calculate probabilistic outcomes. The latter enabled us to use probabilistic ensembles. The doc2vec representations combined with Logistic Regression were used as classifier in itself: the document representations were tagged with the labels of the training data. This model was then used to derive vector representations for new documents. To test if a combination of feature types could improve performance, we evaluated soft voting (argmax of the sums of the predicted probabilities) of the best individual classifiers for the best performing variants of each feature set. Significance testing was done with pair-wise t-tests.

To evaluate the performance, the average F_1 score of a 10-fold cross validation was used. For each run, hyper-parameters were tuned for that specific training set using a 10-fold grid search on the training data. The tuning grids were based on sklearn documentation: C from 10^{-3} to 10^{3} (steps of x10) for LinearSVC and Logistic Regression; number of neighbors from 3 to 11 (steps of 2) for KNN; and max iterations from 2 to 2048 (steps of x2) and alpha from 10^{-8} to 10^{-2} (steps of x10) for SGD. The dimensionality of the document vectors was tuned on a grid of 100 to 400 (steps of 100).

3.3.4. TOPIC MODELING OF THE WHOLE DATA SET

To label the remaining data, the best performing classifier was used with the hyper-parameter settings that were optimal in the majority of the training sets. To investigate which topics are discussed in the patient narratives, we used topic modeling with non-Negative Matrix Factorization of the TF-IDF weighted tokens without stopwords. Topic coherence, measured using TC-W2V [223], was used to select the number of topics. Topic

labels were assigned manually by exploring the words with the highest weights and the top-ranked (i.e., most relevant) messages per topic.

3.4. RESULTS

3.4.1. ANNOTATED DATA

The data was slightly imbalanced, with 37.7% of the posts containing a narrative, resulting in a majority baseline of roughly 0.62. The inter-annotator agreement was substantial ($\kappa = 0.69$).

3.4.2. CLASSIFIER EVALUATION

A Linear SVC on character 3-grams achieves the highest F_1 score (Table 3.1), although character 4-grams ($p = 0.526$), stemmed unigrams ($p = 0.930$) and lemmatized unigrams ($p = 0.587$) do not perform significantly worse. Character 5- and 6- grams also do not perform worse overall ($p = 0.122$ and $p = 0.169$), but their recall is significantly lower ($p = 0.023$ and $p = 0.029$). The classifiers for the best performing document embeddings (DBOW+DM) and psycho-linguistic features, however, are significantly worse overall than character 3-grams ($p = 0.0055$ and $p = 0.026$ respectively). Employing TF-IDF weighting does not aid any of the unigram or character n-gram features. Additionally, neither feature selection (F_1=0.761) nor word boundaries (F_1=0.796) improve the performance of character 3-grams. Using a range of character n-grams, namely 3-to-4 (F_1=0.814), 3-to-5 (F_1=0.814), or 3-to-6 (F_1=0.812), also does not boost performance.

Ensemble classification did not perform better than character 3-grams alone (see Table 3.2). Nevertheless, an ensemble of all four feature types is significantly more precise than all other classifiers ($p = 0.0048$ compared to the second best). To further explore why ensemble classification does not manage to improve overall performance, we investigated the predictions of individual classifiers. As can be seen in Table 3.3, there is a high degree of overlap between the predictions based on character 3-grams and the other feature sets (88.3%, 83.8% and 84.4% respectively). Consequently, the vast majority of the predictions cannot be improved by complementing character 3-grams with these feature sets. Interestingly, 4.7% of the posts are misclassified by all feature sets. Considering the non-overlapping predictions, the percentage of correct predictions was higher for character 3-grams than for either document embeddings or psycho-linguistic features in a pairwise comparison. Thus, it appears that adding these features would be more detrimental than beneficial to narrative classification.

3.4.3. INFLUENTIAL FEATURES

Narratives are typically distinguished by terms relating to the past tense (*was, had, years*), health (*imatinib, tumor, surgeri*) and first-person narrative (*my, i*) (see Figure 3.1). This is corroborated by the character 3-grams, psycho-linguistic features and document embeddings. Some of the important terms for non-narrative texts are also health-related (*patients, gist*) and first-person narrative (*we, us*), which showcases the difficulty of the task at hand. In general, non-narrative texts seem to focus more on emotional support (*prayer, share, may*), second-person narrative (*you, your*) and the future (*may, will*). The psycho-linguistic features additionally reveal that narratives contain more mentions of

Table 3.1: Mean test score (10-fold CV) for best classifiers per feature set

Feature set		Size	Classifier	F_1	R	P
Unigrams	**Original**	4,078	SGD	0.795 ± 0.025	0.788 ± 0.074	0.811 ± 0.055
	Stemmed	3,205	SGD	0.814 ± 0.031	0.793 ± 0.047	0.840 ± 0.049
	Lemmatised	3,777	SGD	0.808 ± 0.039	0.810 ± 0.059	0.813 ± 0.070
Character n-grams	**3-grams**	5,086	SVC	**0.815 ± 0.035**	**0.844 ± 0.047**	0.793 ± 0.058
	4-grams	16,496	SVC	0.811 ± 0.027	0.827 ± 0.068	**0.844 ± 0.029**
	5-grams	36,349	SGD/SVC	0.796 ± 0.023	0.784 ± 0.059	0.817 ± 0.069
	6-grams	60,443	SGD	0.793 ± 0.040	0.797 ± 0.042	0.795 ± 0.079
LIWC		82	SVC	0.773 ± 0.031	0.805 ± 0.044	0.752 ± 0.077
Doc2vec	**DBOW**	400	LogReg	0.737 ± 0.029	0.751 ± 0.056	0.735 ± 0.066
	DM	400	LogReg	0.762 ± 0.039	0.749 ± 0.062	0.785 ± 0.070
	DM+DBOW	800	LogReg	0.77 ± 0.037	0.803 ± 0.064	0.749 ± 0.055

Table 3.2: Mean test score (10-fold CV) for ensemble classification. * DM+DBOW variant.

Feature sets	F_1	R	P
3-grams + LIWC + Doc2vec* + Stemmed Unigrams	0.770 ± 0.029	0.703 ± 0.065	**0.859 ± 0.053**
3-grams + LIWC + Doc2vec*	**0.795 ± 0.037**	**0.772 ± 0.072**	0.829 ± 0.065
3-grams + LIWC	0.706 ± 0.032	0.624 ± 0.059	0.828 ± 0.073
3-grams + Doc2vec*	0.755 ± 0.048	0.735 ± 0.089	0.786 ± 0.040

causality and negative emotions. In contrast, non-narrative texts seem to contain more positive emotions. Lastly, as predicted, function words appear important for classifying narratives in social media, and it is thus advisable to not remove stopwords.

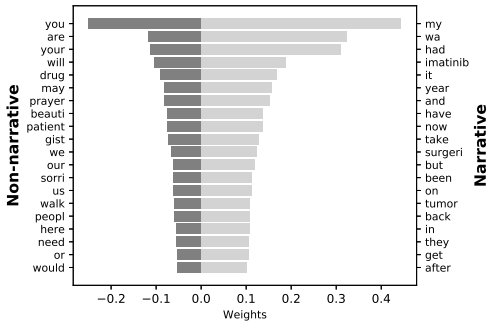
3.4.4. ERROR ANALYSIS FOR THE BEST PERFORMING CLASSIFIER

Error analysis reveals that a significant proportion of the errors is due to incorrect annotation: 36.9% of the false positives and 36.2% of the false negatives were labeled incorrectly (see Table 3.4). Specifically, annotators have difficulty correctly labeling discussions about personal medical facts or side effects as narratives (e.g *'i have been on imatinib 5 months and lost 1/3 of my hair'*). Conversely, annotators may incorrectly judge posts that give emotional support, external information or advice to be narratives while they are not (e.g., *'i may be wrong but total gastrectomy sounds very extreme for two small gist'*).

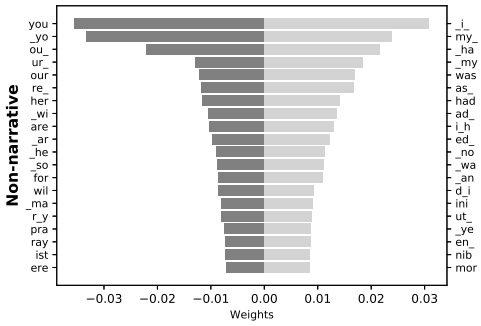
The incorrect labeling may have impacted the automated classification such that these categories are also more difficult for the computer to distinguish. The classifier does, however, appear to outperform human judgment and to some extent 'correct' their mistakes. In fact, its performance may be underestimated by the metrics based on these incorrect labels. Other types of posts that appears challenging for the computer are posts that lack context or contain questions. The former are often answers to unknown questions posed earlier in the thread.

Table 3.3: Comparison of predictions of classifiers for different feature sets. * DM+DBOW variant.

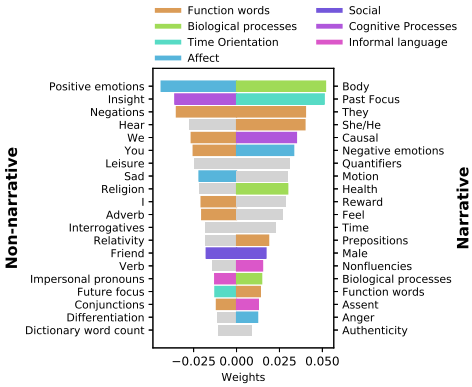
	Compared to	Both		Difference	
		Correct(%)	Incorrect(%)	In Favor of 3-grams(%)	In Favor of Other Method(%)
Character 3-grams	**LIWC**	75.0	8.8	8.4	7.7
	Doc2Vec*	74.8	9.6	8.6	6.9



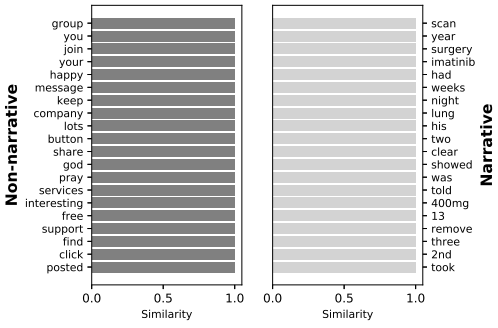
(a) Stemmed Unigrams



(b) Character 3-grams



(c) LIWC



(d) Doc2vec DM model

Figure 3.1: The 20 most influential features in individual classifiers. In (b) underscores represent spaces.

3.4.5. FREQUENCY AND CONTENT OF PATIENT NARRATIVES

Automated narrative detection in unsupervised data The percentage of narratives in the unlabeled data is 37.0 %, which is comparable to the annotated sample. This results in a total of 13.436 posts for topic modeling.[6]

[6]The code for unsupervised narrative filtering is shared at: `https://github.com/AnneDirkson/NarrativeFilter`

Table 3.4: Error analysis for best classifier (character 3-gram classification of narratives)

False positives		**False negatives**	
Reasons for misclassification	Frequency	Reasons for misclassification	Frequency
Mislabeling	24	Mislabeling	17
Emotional support/thanks	15	Unknown	12
Information/advice	13	Lack of context	7
Lack of context	7	Question	5
Question	4	Non-medical narratives	3
Unknown	1	Hypothetical	1
Empty post	1	Empty post	2
TOTAL	65	TOTAL	47

Topic modeling The TC-W2V metric [223] identifies the optimal number of topics to be fourteen. The resulting topics relate to different aspects of the medical process for GIST patients (see Table 3.5). Note that imatinib is the most commonly used medication.

3.5. DISCUSSION

The detection of narratives was most optimal when using character 3-grams. Their strength is in their ability to cluster relevant word types based on suffixes and prefixes. This is especially relevant in the medical domain, e.g., all cancer medication for GIST ends in *'nib'*. In contrast, psycho-linguistic features appear to suffer from oversimplification, because they aggregate words that define *different* classes into one category e.g., *we* and *my* into the umbrella category of first person pronouns (see Figure 3.1). The use of document embeddings may have been hampered by the small size of the data. An alternative explanation could be that incorrect labeling impacts these features more strongly than word-based features.

Narratives could be differentiated most strongly by their use of past tense, first-person narrative and health-related words. The first two are in line with linguistic definition of a narrative. The stronger focus on health, however, may indicate that patients prefer to share their own health experiences than health information from external sources.

Annotating narratives appears a challenging task, despite providing annotators with a guideline based on previous work [328] and validated through initial annotation by two annotators. This is underscored by our inter-annotator agreement ($\kappa = 0.69$) which was comparable to that of Verberne et al. [328] ($\kappa = 0.71$). Our classifier performed less well that their system ($F_1 = 0.91$), which may be explained by their larger sample of annotated data (2.051 posts).

Inevitably, our results depend on the choice of what constitutes a narrative and how the annotators interpret this definition. It appears that especially the line between a medical fact about oneself and a medical experience is fuzzy for annotators. Future studies could perhaps use this knowledge to develop clearer guidelines.

Table 3.5: Most important topics discussed in patient forum narratives. Topic labels were assigned manually. *A type of cancer medication

Topic labels	Top 10 words	Top-ranked post for the topic
Tumor location	tumor stomach removed liver small cm mitotic metastases rate intestine	"i only had one tumor on my stomach"
(Emotional) Coping	take get time doctor like also know imatinib* day would	"i completely understand i started 400 imatinib after surgery in and have lots of bad days [...]"
Duration of Treatment	years imatinib* almost ago 10 taking two still 11 12	"about 1 and 1/2 years"
Types of Scans	scan ct pet results next today last showed week cat	"oops one is a ct scan and one is a pet scan"
Diagnosis of GIST	gist diagnosed cancer specialist oncologist husband anyone ago surgeon found	"that was my gist"
Other Medication	sunitinib* regorafenib* sorafenib* imatinib* working 37 exon nilotinib* trial stopped drug	"i have this on sunitinib"
Side Effects	side effects imatinib* effect different fatigue eyes bad 400mg time	"and no side-effects"
Tumor Surgery	surgery remove since weeks first post surgeon second shrink done	"just had surgery"
Absence of Tumor Recurrence	disease evidence still years today post since resection year far	"no evidence of disease no evidence of disease"
Recurrence of Work, Medication or Tumor	back came come hair go went weeks took coming lost	"i started imatinib after i went back to work"
Emotional support	good luck news best far hope bad goes well keep pretty	"all my best and good luck"
Dosage of Medication	mg 400 800 imatinib* 600 take day taking since started	"11 years of imatinib since 2003 at 600 mg and since november 2009 at 800 mg [...]"
Timing of Scans	months every scans three ct six year two first month	"my doctor said 3 years"
Ingesting imatinib	one year last took imatinib* day another old got time	"take imatinib"

3.6. CONCLUSION

For the detection of patient narratives on social media, psycho-linguistic features and document embeddings are outperformed by character 3-grams. These narratives are associated with the past tense, health and first-person pronouns, whereas non-narrative text is associated with the future tense, emotional support and second-person pronouns. The patient narratives could be subdivided into discussions of fourteen different medical topics, ranging from surgery to side effects. Future work will develop automated methods for the extraction of patient knowledge from the narratives.