Narrative Detection and Feature Analysis in Online Health Communities

Achyutarama R. Ganti and **Steven R. Wilson** and **Zexin Ma**
Oakland University
{ganti,stevenwilson,zexinma}@oakland.edu

Xinyan Zhao
University of North Carolina
ezhao@unc.edu

Rong Ma
Butler University
rma@butler.edu

Abstract

Narratives have been shown to be an effective way to communicate health risks and promote health behavior change, and given the growing amount of health information being shared on social media, it is crucial to study health-related narratives in social media. However, expert identification of a large number of narrative texts is a time consuming process, and larger scale studies on the use of narratives may be enabled through automatic text classification approaches. Prior work has demonstrated that automatic narrative detection is possible, but modern deep learning approaches have not been used for this task in the domain of online health communities. Therefore, in this paper, we explore the use of deep learning methods to automatically classify the presence of narratives in social media posts, finding that they outperform previously proposed approaches. We also find that in many cases, these models generalize well across posts from different health organizations. Finally, in order to better understand the increase in performance achieved by deep learning models, we use feature analysis techniques to explore the features that most contribute to narrative detection for posts in online health communities.

1 Introduction

Narrative forms of communication are widely used for conveying information and building connections. Broadly defined as a representation of someone's experience of a series of events (Bilandzic and Busselle, 2013), narratives take on different formats, ranging from short anecdotes and testimonials to lengthy entertainment TV shows and movies (Kreuter et al., 2007).

In the health context, extensive research has found that narratives are more effective than non-narratives (e.g., statistics, didactic arguments) in communicating health risks (Janssen et al., 2013; Ma, 2021) and promoting health behavior change (Kreuter et al., 2010). Moreover, telling personal illness narratives helps patients to better cope with the illness (Carlick and Biley, 2004) and for health care professionals to better understand the illness (Kalitzkus and Matthiessen, 2009). Given that social media has become a widely used platform for cancer patients and their caregivers to share stories and connect with others (Gage-Bouchard et al., 2017; Hale et al., 2020), it is critical to understand what cancer narratives are told on social media and how they engage social media users.

However, in order to understand the impact of narratives in online communication, narratives must first be *identified* in social media datasets. Doing this often requires annotations from subject matter experts, which can be a costly process and difficult to scale up to massive datasets. In this work, we seek to understand the extent to whether natural language processing methods, specifically, fine-tuned large language models, can be used to automatically detect narratives within social media posts in the health domain using only a relatively small number of expert annotations. Additionally, analyzing models that are able to successfully detect narratives can provide *insights* into the types of textual features that are most related to narrative text within a corpus.

Toward these aims, we collect and annotate a dataset of social media posts created by breast cancer organizations and address the following research questions:

RQ1 Which text classification models provide the best performance for automatic narrative detection for social media texts posted by breast cancer organizations?

RQ2 How does the ability to detect narratives generalize across posts written by different organizations?

RQ3 Which features are most important for automatic narrative detection in this context?

To answer **RQ1**, We compare a range of text classification methods and find that transformer-based deep-learning based methods outperform classical approaches like support vector machines, as well as the previous state-of-the-art method for detecting narratives within health-related social media posts (Dirkson et al., 2019). To answer **RQ2**, we split our dataset so that the same organizations' accounts are not used for both train and test data, finding that in most cases, it is possible for our best models to generalize well across organizations. Finally, to answer **RQ3**, we use machine learning analysis tools to identify which features contribute most to the prediction of narratives, finding that references to people, such as pronouns and names, as well as state-of-being verbs like "is", contributed strongly to cases where models predicted that texts contained narratives.

Our results suggest that automatic detection of narratives in social media posts is a promising application of text classification, and can help ease the burden of manual annotation for researchers seeking to study the relationship between narrative and other variables of interest at scale.[1]

2 Related Work

Online health communities have been computationally studied before in order to understand how users show social support for one another (Andy et al., 2021), to automatically extract information needs of patients (Romberg et al., 2020), and to identify linguistic patterns associated with anxiety (Rey-Villamizar et al., 2016). Additionally, Antoniak et al. (2019) analyzed birth stories from an online forum and demonstrated the utility of these stories for computational work. Machine learning models have been trained using textual health forum data to predict attributes such as the sentiment (Ali et al., 2013) or cancer stage of the patients posting to forums (Jha and Elhadad, 2010). Yet, most work in the area of computational analysis of online medical forums has not considered the importance of narrative. At the same time, computational approaches incorporating and extracting narratives have led to advances in the study of corporate finance (Zmandar et al., 2021), environmental issues (Armbrust et al., 2020), the analysis of clinical records (Jung et al., 2011), and emotion classification within stories (Tanabe et al., 2020).

As NLP datasets, narratives are often directly collected by sampling data from sources that are already known to use narrative based on the genre of the corpus, such as literary works (Hammond et al., 2013), doctors' notes (Elhadad et al., 2015), or fan fiction (Yoder et al., 2021). In the social media domain, data is often sampled in a way to ensure the presence of narratives, e.g., by collecting posts from specific subreddits which typically contain narrative style posts (Yan et al., 2019).

In other cases, the presence or location of narrative content is unknown beforehand and needs to be to detected or extracted. This might be done using filtering criteria like the length of the post or the presence of predefined linguistic patterns (Vijayaraghavan and Roy, 2021). However, some datasets contain a balanced mixture of both narrative and non-narrative content, and quick rule-based filtering is not adequate. In the domain of online health communities specifically, prior work has relied on expert annotations to determine what should or should not be considered a narrative (Dirkson et al., 2019; Verberne et al., 2019). In each of these works, text classification models were trained to automatically determine whether or not a given post contained narratives, and support vector machines (SVM) using bag-of-words or character n-gram features were found to be the best approach.

We build upon this existing work by applying deep learning text classification models to the task of narrative detection in social media posts from breast cancer organizations as an example use case that includes personal narratives, texts for which narrative presence is unknown *a priori*, and provide the potential for enabling larger scale studies of the importance of narratives in health communication. We find that these approaches outperform SVM-based models similar to those used by Dirkson et al. (2019)[2] and Verberne et al. (2019) and explore their effectiveness on our dataset throughout the rest of this paper.

3 Data Collection and Annotation

A list of breast cancer non-profit organizations was identified from the Canadian cancer survivor net-

[1] Code and annotations are publicly available at `https://github.com/ou-nlp/NarativeDetection`.

[2] We contacted the authors of these papers but they could not share their data due to user privacy restrictions. Therefore, we only use the same approach reported by the authors, rather than applying our proposed deep learning models on the same datasets that were used in those studies.

Organization	Posts	Tokens	Narrative
Susan G. Komen	212	10845	65.57%
NBCF USA	144	11433	58.33%
Breast Cancer Now	186	18932	64.52%
AFWBC Canada	116	7636	21.55%
NBCF Australia	191	11161	25.13%
Total	**849**	**60007**	**49.0%**

Table 1: Annotated data set statistics.

work partners page[3]. We selected five organizations with the most Facebook followers and spanning several different countries, including Susan G. Komen For the Cure, National Breast Cancer Foundation USA, the UK-based Breast Cancer Now, A Future Without Breast Cancer (Canadian Cancer Society), and the National Breast Cancer Foundation Australia. Their Facebook posts and engagement metrics from 2016 to 2021 were downloaded using CrowdTangle[4] ($N = 8,580$).

The top 10% posts in terms of total interactions were sampled for annotation. Following standard procedures in content analysis (Riff et al., 2014), two expert coders annotated the presence of narratives (48.83%). All disagreements were resolved by discussion, and the consensus results were used for further analyses (i.e., the highest standard of intercoder reliability) (Krippendorff, 2004). The overall agreement rate was above 0.9. For this study, we omit 9 posts which did not contain any text and only videos or images. The breakdown of the annotated dataset by non-profit organization account is presented in Table 1.

4 Detecting Narrative Style

Next, we set out to determine how well various text classification models could detect the presence of narratives given the expert annotations as training data. For this experiment, we appended data from all five non-profit organizations into a single dataset. All the data points were then shuffled and split using 80% of the data for training, and each 10% for validation and test sets. The metrics that were used for model evaluation are the F1 scores, Precision, and Recall of the narrative class. We consider two categories of models: classical machine learning models using bag-of-words features, and transformer-based deep learning models.

For the **classical models**, we experiment with various preprocessing schemes in terms of lowercasing, lemmatization, and stopword removal, and choose the approach that gave the best performance on our validation set. That process included: lowercasing, removing URLs, lemmatization using NLTK's wordnet (Miller, 1995) lemmatizer, and stopword removal using the NLTK (Bird et al., 2009). However, given the importance of pronouns in narrative detection as evidenced in prior work (Dirkson et al., 2019), we do not remove pronouns as part of our stopword removal step. The models that we consider are Naive Bayes, Logistic Regression, and SVM-classification, using each model's scikit-learn (Pedregosa et al., 2011) Python implementation. Model-specific hyperparameters were also tuned using the validation set as described in Appendix A.

Additionally, we consider the best reported approach from Dirkson et al. (2019), which is the previous best reported narrative detection model for online health forum data. We use the code provided by the authors to both preprocess the data and train the predictive model. The authors used an SVM classifier with a linear kernel and character-level trigram features as input, and so we refer to this model as SVM-trigram in our results.

For the **deep learning models**, we use DistilBERT (Sanh et al., 2019), BERT (Devlin et al., 2019), and RoBERTa (Liu et al., 2019) models based on the `DistilBERT-Base-Uncased`, `BERT-Base-Uncased`, and `RoBERTa-Base` checkpoints available from HuggingFace (Wolf et al., 2019). The tokenizer for each model was automatically determined using the `AutoTokenizer()` class. We use the output representation of the `[CLS]` token as input to the classification layer (the default approach when using the HuggingFace `Trainer` class). Hyperparameters are described in Appendix A.

The results of running each of these models are presented in Table 2. It is evident that deep learning models are capable of distinguishing narratives from non-narratives in the sequences, with BERT showing the best overall performance. Among the classical machine learning methods, the SVM model outperformed others with an F1 score and accuracy of 0.901. Although our classical methods didn't perform poorly, there is a substantial gain in F1-score when using the deep learning approaches. Therefore, for the generalization experiments in the next section, we only consider the best performing model, i.e., the BERT model.

[3] https://survivornet.ca/connect/partners

[4] https://www.crowdtangle.com/

Model	F1	Prec	Recall
Baseline-narrative	0.680	**1.000**	0.512
Naive Bayes	0.879	0.952	0.816
SVM	0.901	0.928	0.876
Log. Regr.	0.891	0.880	0.902
SVM-trigram	0.889	0.935	0.847
DistilBERT	0.964	**1.000**	0.931
BERT	**0.988**	0.977	**1.000**
RoBERTa	0.977	**1.000**	0.956

Table 2: Narrative class F1, Precision, and Recall scores of the text classification models on the narrative detection task, separated into groups of classical ML and deep learning methods. The score of the performing model(s) for each metric is listed in **bold**. SVM-trigram is the best performing model from (Dirkson et al., 2019). Baseline-narrative is the score achieved by labeling all texts as narrative.

4.1 Generalizing across accounts

A model's ability to generalize to unseen data is key to a successful deployment. Our deep learning[5] models can successfully classify the presence of narratives in social-media posts, but it is possible that they overfit to features that are specific to the set of organizations that generated the posts included in our dataset. To evaluate the generalizability of the BERT model to data from unseen organizations, we re-trained the model on data from only four organizations, leaving the fifth one out as test data. We then repeat this process again for each of the five organizations, so that each organization is used as the held-out test set once, and as part of the training set in all other cases.

The results of this experiment are presented in Table 3. The posts from the organization Breast Cancer Now held out as test data were the easiest to generalize to (F1 score of 0.991) compared to the other combinations. On the other hand, the model slightly under-performed when trained on data from all organizations leaving NBCF Australia as test set with a F1 score of 0.900. However, in all cases, this shows that there is good potential for models trained on a subset of organizations to generalize well to others.

Target	F1	Prec	Recall
Susan G. Komen	0.949	0.973	0.927
Breast Cancer Now	**0.991**	0.903	**1.000**
NBCF Australia	0.900	0.903	0.979
NBCF USA	0.976	0.976	0.976
AFWBC Canada	0.936	**1.000**	0.880

Table 3: Generalization performance using the best classifier (BERT) by training on all accounts except for the target account, and testing on the target account.

Train	F1	Prec	Recall
Susan G. Komen	0.917	0.852	**0.993**
Breast Cancer Now	0.777	0.979	0.645
NBCF Australia	**0.953**	0.961	0.945
NBCF USA	0.877	0.791	0.985
AFWBC Canada	0.914	**0.976**	0.859

Table 4: Generalization performance using the best classifier (BERT) by training on one account and testing on the remaining four target accounts.

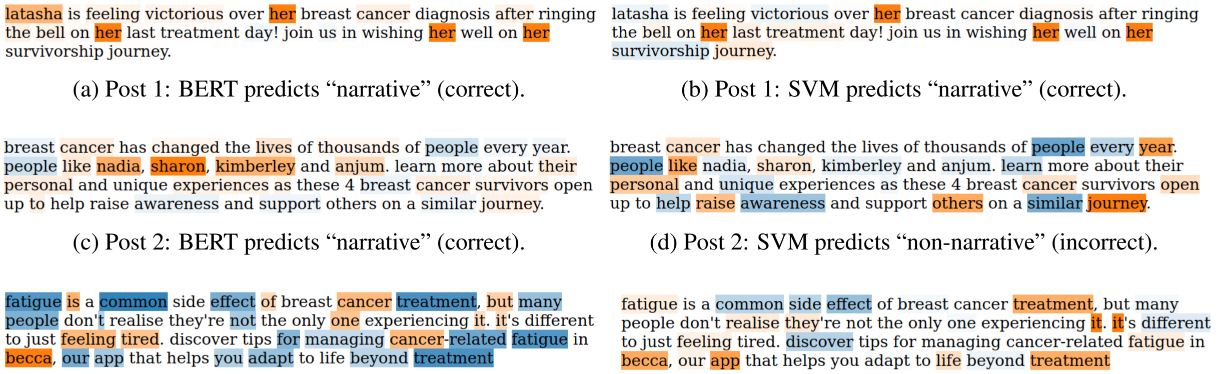
We then performed one slightly varied version of the same experiment to further determine model generalizability. Here, we chose a dataset from only one organization as the training set, and used the remaining four datasets as testing data. As before, we repeat this experiment five times, using each organization as training data once, and testing in all other cases. This experiment helps to determine the potential for cross-organization transfer when we have very limited data or data from a single source. Given the very small amount of data for some of the organizations, we found that the size of the training set was too small to learn effective models in some cases. Therefore, we chose to up-sample our training set by 200%, (duplicating each training instance) which we found empirically to give better results in the low training data case. From the final result (Table 4), we observe that the model trained on NBCF Australia performs the best overall, achieving an F1 score that is within a few points of the model trained on data from all organizations from Table 2. On the other hand, the model trained only on Breast Cancer Now posts had poor generalization performance on the data from the other organizations, suggesting that having data from only a single organization is not always enough to guarantee good generalizability.

5 Analysis of Narrative Detection Models

We have established that deep learning models are very effective at detecting narratives from social media data, substantially outperforming classical machine learning approaches. However, it is not immediately apparent *why* these models are able to achieve better F1 scores. Therefore, in this section, we use model interpretability tools to further examine which features contributed to the ability of our models to detect narratives.

We chose the best performing models in each cat-

[5] We also experimented with our best performing classical ML model, SVM, in the same way, but the results were not as strong (Appendix B).



latasha is feeling victorious over her breast cancer diagnosis after ringing the bell on her last treatment day! join us in wishing her well on her survivorship journey.

(a) Post 1: BERT predicts "narrative" (correct).

latasha is feeling victorious over her breast cancer diagnosis after ringing the bell on her last treatment day! join us in wishing her well on her survivorship journey.

(b) Post 1: SVM predicts "narrative" (correct).

breast cancer has changed the lives of thousands of people every year. people like nadia, sharon, kimberley and anjum. learn more about their personal and unique experiences as these 4 breast cancer survivors open up to help raise awareness and support others on a similar journey.

(c) Post 2: BERT predicts "narrative" (correct).

breast cancer has changed the lives of thousands of people every year. people like nadia, sharon, kimberley and anjum. learn more about their personal and unique experiences as these 4 breast cancer survivors open up to help raise awareness and support others on a similar journey.

(d) Post 2: SVM predicts "non-narrative" (incorrect).

fatigue is a common side effect of breast cancer treatment, but many people don't realise they're not the only one experiencing it. it's different to just feeling tired. discover tips for managing cancer-related fatigue in becca, our app that helps you adapt to life beyond treatment

(e) Post 3: BERT predicts "non-narrative" (correct).

fatigue is a common side effect of breast cancer treatment, but many people don't realise they're not the only one experiencing it. it's different to just feeling tired. discover tips for managing cancer-related fatigue in becca, our app that helps you adapt to life beyond treatment

(f) Post 3: SVM predicts "narrative" (incorrect).

Figure 1: Feature importance visualization for three posts, one per row, that were classified by our top-performing deep learning model (BERT) and classical machine learning model (SVM). Orange (blue) shading indicates the token was found to be important for the "narrative" ("non-narrative") class by LIME, with the color intensity indicating the degree of importance. Post 1 was correctly classified by both models, while posts 2 and 3 were correctly classified by BERT but incorrectly classified by the SVM model.

egory, i.e., BERT for the deep learning approaches, and SVM for the classical models, and use the explainable AI tool for Local Interpretable Model Agnostic Explanation (LIME; Ribeiro et al. (2016)) to understand the significance of text-based features to each model. In both cases, we use the LIME explainer function[6] to learn which features best explain the narrative class and non-narrative class. We chose 5000 samples and 25 features as parameters for the function, based on the suggested default values and our desire to include a reasonable number of features per example.

Each instance in the test dataset is examined using LIME, which generates an importance score for each feature (token) in the input based on how much it contributes to predictions for the positive class (narrative) or negative class (non-narrative). For a given feature j in a given text i, a higher positive score W_{ij} denotes greater importance of that feature in the overall narrative class and a lower positive score denotes a weaker importance of that feature for the same class. Likewise, a greater negative value W_{ij} for a feature indicates a stronger association with predictions of the non-narrative class. Several examples of LIME explanations are presented in Figure 1. We can see that for posts where both models made the correct prediction, the set of important features is approximately the same. However, when BERT made the correct prediction and SVM did not, we notice that BERT places a greater emphasis on first names in the case of narratives, and features like "fatigue" and "common", which refer side effects of breast cancer, are correctly identified as important indicators that the post does not contain a narrative.

While these qualitative results are highly useful, LIME only provides the W_{ij} score for a specific text, i, yet we sought to quantitatively understand which features were important across the entire test set. Therefore, we use Global Aggregations of Local Explanations (GALE; van der Linden et al. (2019)) to aggregate the LIME scores. For the purposes of aggregation, we set a cut-off of $\epsilon = 0.001$ and consider any $W_{ij} < \epsilon$ as a score of 0. A feature importance score of zero indicates that the feature does not explain much of either the narrative or the non-narrative class while making predictions. GALE suggests several different methods for aggregating scores, but we use the Global Average Importance I^{AVG} as it was found to correlate well with external measures of feature importance for model classification. The Global Average Importance I_j^{AVG} for a given feature j is defined as:

$$I_j^{AVG} = \frac{\sum_{i=1}^{N} |W_{ij}|}{\sum_{i:W_{ij}\neq 0} \mathbb{1}}$$

where N is the number of texts in the corpus.

Table 5 shows the top and bottom 10 aggregated feature importance scores for both BERT and SVM. Both the models put more emphasis on pronouns and first names as they are more personal to the storyteller or subject of the narrative. Our feature analysis results align with that of Dirkson et al. (2019) who noted that narratives in health forums are characterized by health related words and first

[6] From https://github.com/marcotcr/lime

BERT		***SVM***	
word	**score**	**word**	**score**
celeste	0.29	her	0.22
she	0.28	taylor	0.20
latasha	0.24	my	0.19
beautiful	0.17	she	0.18
mother	0.16	app	0.15
her	0.15	peace	0.14
barbe	0.14	becca	0.13
hall	0.11	tip	0.13
found	0.09	rest	0.12
is	0.09	his	0.11
s	-0.04	face	-0.10
don’	-0.04	study	-0.11
significant	-0.04	run	-0.11
round	-0.05	mammogram	-0.11
myresearchstory	-0.05	mel	-0.12
awareness	-0.05	addy	-0.15
free	-0.06	steph	-0.15
it	-0.06	listen	-0.16
"	-0.06	mondaymotivation	-0.17
increase	-0.08	song	-0.19

Table 5: Top and bottom ten aggregated feature importance scores for BERT (left side) and SVM (right side) models trained for narrative detection. Larger positive values indicate a greater overall importance for the “narrative” class, while more negative values were more important for predicting the “non-narrative” class.

person pronouns. Also, since breast cancer is more common among women, it is more common to see feminine pronouns and first names related to women with the only exception being the token “his” which can be found as an important feature for the “narrative” class in the SVM model. Upon further inspection, we found that there are instances referring to women as “his wife” and “his mother” which further validates the model’s choice for the token in the positive list. We also note verbs such as “found” (connected to “lump”, which also had a positive score for both models but is not in the top ten for either) and “is”.

Considering the tokens with negative values, indicating that they were more relevant when predicting the “non-narrative class”, we found words related to scientific studies, sharing songs, and describing clinical procedures. Hashtags such as “myreserachstory” and “mondaymotivation” were also present, indicating posts that may have been trying to seek engagement through means other than the use of narrative. While Our BERT model was successful in detecting narratives by learning associations between features like pronouns and first names, the SVM model failed to consistently learn these associations as indicated by the placement of several first names in the non-narrative (negative valued) end of the list.

BERT − SVM	
word	**score**
celeste	0.29
latasha	0.24
barbe	0.14
mother	0.12
hall	0.11
beautiful	0.11
she	0.10
found	0.09
is	0.09
i	0.08
strong	-0.02
diagnosis	-0.02
bell	-0.03
reality	-0.03
be	-0.05
journey	-0.06
her	-0.07
it	-0.08
his	-0.09
my	-0.12

Table 6: Top and bottom ten features that differed in importance the most between the BERT and SVM model. Scores with a larger value had more overall importance for the BERT model, while features with a smaller value had more importance for the SVM model.

While these results illustrate which features were important to each model, they do not directly *quantify* the difference between the BERT and SVM. To investigate that further, we checked the extent to which the degree of importance I_j^{AVG} for each feature differed between BERT and the SVM model (Table 6). For each feature in the list obtained from SVM, we subtract the corresponding aggregated importance score from BERT for that feature. If the result is positive, it indicates that the BERT model puts more emphasis on that feature, whereas if the result is negative, it indicates that SVM gives more importance for that feature compared to BERT model for predicting the “narrative” class. We observe that BERT assigns a higher weight for first names and the pronoun “she” has a higher importance for BERT compared to SVM whereas, the pronoun “her” appears to be given greater importance by the SVM model compared to BERT.

6 Conclusion

In this paper, we show that deep learning models like BERT, DistilBERT and RoBERTa are effective at detecting narratives from social media data. Previous research focused on the use of classical machine learning models to understand narratives in online health discussion forums, but we demonstrate that deep learning models outperform these when detecting the presence of narratives. We studied generalizability of the deep learning models across organizations, finding that overall, models are able to generalize well across accounts, suggesting that deep learning models provided with sufficient data can perform well on an unseen dataset with similar distributions. We also analyze the performance of deep learning models with explainable AI methods, uncovering important features that contribute to narratives in a particular context.

However, there are certain limitations and challenges associated with these models. Although they are quite successful at understanding narratives, performance of deep learning models is directly proportional to the quality of the dataset and they are highly susceptible to annotator and dataset bias.

With the growing amount of health information being shared on social media, understanding narratives becomes extremely important to study public health behavior and estimate health risks. The work described in this paper is a step towards helping researchers automatically annotate narratives in social media posts, thus enabling larger scale studies of the impact of narratives on health conversations.

References

Tanveer Ali, David Schramm, Marina Sokolova, and Diana Inkpen. 2013. Can I hear you? sentiment analysis on medical forums. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 667–673, Nagoya, Japan. Asian Federation of Natural Language Processing.

Anietie Andy, Brian Chu, Ramie Fathy, Barrington Bennett, Daniel Stokes, and Sharath Chandra Guntuku. 2021. Understanding social support expressed in a COVID-19 online forum. In *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis*, pages 19–27, online. Association for Computational Linguistics.

Maria Antoniak, David Mimno, and Karen Levy. 2019. Narrative paths and negotiation of power in birth stories. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

Felix Armbrust, Henry Schäfer, and Roman Klinger. 2020. A computational analysis of financial and environmental narratives within financial reports and its value for investors. In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 181–194, Barcelona, Spain (Online). COLING.

Helena Bilandzic and Rick Busselle. 2013. Narrative persuasion. *The Sage handbook of persuasion: Developments in theory and practice*, 2:200–219.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

Alice Carlick and Francis C Biley. 2004. Thoughts on the therapeutic use of narrative in the promotion of coping in cancer care. *European Journal of Cancer Care*, 13(4):308–317.

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.

AR Dirkson, Suzan Verberne, Wessel Kraaij, AM Jorge, R Campos, A Jatowt, and S Bhatia. 2019. Narrative detection in online patient communities. In *Proceedings of Text2Story—Second Workshop on Narrative Extraction From Texts co-located with 41th European Conference on Information Retrieval (ECIR 2019)*, pages 21–28. CEUR-WS.

Noémie Elhadad, Sameer Pradhan, Sharon Gorman, Suresh Manandhar, Wendy Chapman, and Guergana Savova. 2015. SemEval-2015 task 14: Analysis of clinical text. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 303–310, Denver, Colorado. Association for Computational Linguistics.

Elizabeth A Gage-Bouchard, Susan LaValley, Michelle Mollica, and Lynda Kwon Beaupin. 2017. Cancer communication on social media: examining how cancer caregivers use facebook for cancer-related communication. *Cancer nursing*, 40(4):332–338.

Brent J Hale, Ryan Collins, and Danielle K Kilgo. 2020. Posting about cancer: Predicting social support in imgur comments. *Social Media+ Society*, 6(4):2056305120965209.

Adam Hammond, Julian Brooke, and Graeme Hirst. 2013. A tale of two cultures: Bringing literary analysis and computational linguistics together. In *Proceedings of the Workshop on Computational Linguistics for Literature*, pages 1–8, Atlanta, Georgia. Association for Computational Linguistics.

Eva Janssen, Liesbeth van Osch, Hein de Vries, and Lilian Lechner. 2013. The influence of narrative risk communication on feelings of cancer risk. *British Journal of Health Psychology*, 18(2):407–419.

Mukund Jha and Noémie Elhadad. 2010. Cancer stage prediction based on patient online discourse. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, pages 64–71, Uppsala, Sweden. Association for Computational Linguistics.

Hyuckchul Jung, James Allen, Nate Blaylock, William de Beaumont, Lucian Galescu, and Mary Swift. 2011. Building timelines from narrative clinical records: Initial results based-on deep natural language understanding. In *Proceedings of BioNLP 2011 Workshop*, pages 146–154, Portland, Oregon, USA. Association for Computational Linguistics.

Vera Kalitzkus and Peter F Matthiessen. 2009. Narrative-based medicine: potential, pitfalls, and practice. *The Permanente Journal*, 13(1):80.

Matthew W Kreuter, Melanie C Green, Joseph N Cappella, Michael D Slater, Meg E Wise, Doug Storey, Eddie M Clark, Daniel J O'Keefe, Deborah O Erwin, Kathleen Holmes, et al. 2007. Narrative communication in cancer prevention and control: a framework to guide research and application. *Annals of behavioral medicine*, 33(3):221–235.

Matthew W Kreuter, Kathleen Holmes, Kassandra Alcaraz, Bindu Kalesan, Suchitra Rath, Melissa Richert, Amy McQueen, Nikki Caito, Lou Robinson, and Eddie M Clark. 2010. Comparing narrative and informational videos to increase mammography in low-income african american women. *Patient education and counseling*, 81:S6–S14.

Klaus Krippendorff. 2004. *Content analysis: An introduction to its methodology*. Sage publications.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zexin Ma. 2021. The role of narrative pictorial warning labels in communicating alcohol-related cancer risks. *Health Communication*, pages 1–9.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Nicolas Rey-Villamizar, Prasha Shrestha, Farig Sadeque, Steven Bethard, Ted Pedersen, Arjun Mukherjee, and Thamar Solorio. 2016. Analysis of anxious word usage on online health forums. In *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis*, pages 37–42, Auxtin, TX. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Daniel Riff, Stephen Lacy, and Frederick Fico. 2014. *Analyzing media messages: Using quantitative content analysis in research*. Routledge.

Julia Romberg, Jan Dyczmons, Sandra Olivia Borgmann, Jana Sommer, Markus Vomhof, Cecilia Brunoni, Ismael Bruck-Ramisch, Luis Enders, Andrea Icks, and Stefan Conrad. 2020. Annotating patient information needs in online diabetes forums. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 19–26, Barcelona, Spain (Online). Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Hikari Tanabe, Tetsuji Ogawa, Tetsunori Kobayashi, and Yoshihiko Hayashi. 2020. Exploiting narrative context and a priori knowledge of categories in textual emotion classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5535–5540, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ilse van der Linden, Hinda Haned, and Evangelos Kanoulas. 2019. Global aggregations of local explanations for black box models. *CoRR*, abs/1907.03039.

Suzan Verberne, Anika Batenburg, Remco Sanders, Mies van Eenbergen, Enny Das, Mattijs S Lambooij, et al. 2019. Analyzing empowerment processes among cancer patients in an online community: A text mining approach. *JMIR cancer*, 5(1):e9887.

Prashanth Vijayaraghavan and Deb Roy. 2021. Modeling human motives and emotions from personal narratives using external knowledge and entity tracking. In *Proceedings of the Web Conference 2021*, pages 529–540.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing.

Xinru Yan, Aakanksha Naik, Yohan Jo, and Carolyn Rose. 2019. Using functional schemas to understand social media narratives. In *Proceedings of the Second Workshop on Storytelling*, pages 22–33, Florence, Italy. Association for Computational Linguistics.

Michael Yoder, Sopan Khosla, Qinlan Shen, Aakanksha Naik, Huiming Jin, Hariharan Muralidharan, and Carolyn Rosé. 2021. FanfictionNLP: A text processing pipeline for fanfiction. In *Proceedings of the Third*

Workshop on Narrative Understanding, pages 13–23, Virtual. Association for Computational Linguistics.

Nadhem Zmandar, Mahmoud El-Haj, Paul Rayson, Ahmed Abura'Ed, Marina Litvak, Geroge Giannakopoulos, and Nikiforos Pittaras. 2021. The financial narrative summarisation shared task FNS 2021. In *Proceedings of the 3rd Financial Narrative Processing Workshop*, pages 120–125, Lancaster, United Kingdom. Association for Computational Linguistics.

A Model Hyperparameters

For Naive Bayes, we did not tune any hyperparameters. For the SVM classifier, we considered linear, polynomial, and rbf kernels, and found the polynomial kernel to work the best. We set the regularization parameter $C = 2$. For the Logistic Regression classifier, we tried various values for the regularization parameter C in the range of $\{0.01, 0.1, 0.2, 1, 2, 10\}$ and found that $C = 1$ gave the best results. For the deep learning models, we use a batch size of 16 with a weight decay of 0.01 and a learning rate of 2e-5, training for 5 epochs.

B Generalizability of SVM model

We performed the same experiments from section 4.1 using an SVM model (the best performing classical model from our experiments in section 4). The results are presented in Tables 7 and 8.

Target	**F1**	**Prec**	**Recall**
Susan G. Komen	0.884	0.776	0.972
Breast Cancer Now	**0.901**	0.883	0.921
NBCF Australia	0.830	**0.970**	0.730
NBCF USA	0.851	0.952	0.769
AFWBC Canada	0.830	0.710	**1.000**

Table 7: Generalization performance using the best classical ML model (SVM) by training on all accounts except for the target account, and testing on the target account.

Train	**F1**	**Prec**	**Recall**
Susan G. Komen	0.803	0.946	0.697
Breast Cancer Now	**0.824**	0.886	0.770
NBCF Australia	0.730	0.582	0.981
NBCF USA	0.733	**0.965**	0.591
AFWBC Canada	0.457	0.296	**1.000**

Table 8: Generalization performance using the best classical ML model (SVM) by training on one account and testing on the remaining four target accounts.