University of Arkansas, Fayetteville

ScholarWorks@UARK

Computer Science and Computer Engineering Undergraduate Honors Theses

Computer Science and Computer Engineering

5-2021

# Using Large Pre-Trained Language Models to Track Emotions of Cancer Patients on Twitter

Will Baker

Follow this and additional works at: https://scholarworks.uark.edu/csceuht

Part of the Data Science Commons, Health Communication Commons, and the Numerical Analysis and Scientific Computing Commons

Citation

Baker, W. (2021). Using Large Pre-Trained Language Models to Track Emotions of Cancer Patients on Twitter. *Computer Science and Computer Engineering Undergraduate Honors Theses* Retrieved from https://scholarworks.uark.edu/csceuht/92

This Thesis is brought to you for free and open access by the Computer Science and Computer Engineering at ScholarWorks@UARK. It has been accepted for inclusion in Computer Science and Computer Engineering Undergraduate Honors Theses by an authorized administrator of ScholarWorks@UARK. For more information, please contact ccmiddle@uark.edu.

Using Large Pre-Trained Language Models to Track Emotions of Cancer Patients on Twitter

A thesis submitted in partial fulfillment
of the requirements for the degree of
Bachelor of Science in Computer Science

By

William Baker

April 2021
University of Arkansas

This thesis is approved for recommendation to the Undergraduate Council

[redacted]

Dr. Justin Zhan
Thesis Advisor

[redacted]

Dr. Matthew Patitz
Committee member

[redacted]

Dr. Brian Primack
Committee member

## Abstract

Twitter is a microblogging website where any user can publicly release a message, called a tweet, expressing their feelings about current events or their own lives. This candid, unfiltered feedback is valuable in the spaces of healthcare and public health communications, where it may be difficult for cancer patients to divulge personal information to healthcare teams, and randomly selected patients may decline participation in surveys about their experiences. In this thesis, BERTweet, a state-of-the-art natural language processing (NLP) model, was used to predict sentiment and emotion labels for cancer-related tweets collected in 2019 and 2020. In longitudinal plots, trends in these emotions and sentiment values can be clearly linked to popular cancer awareness events, the beginning of stay-at-home mandates related to COVID-19, and the relative mortality rates of different cancer diagnoses. This thesis demonstrates the accuracy and viability of using state-of-the-art NLP techniques to advance the field of public health communications analysis.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

## 1 Introduction

Twitter has exploded in popularity since its inception in 2006 to become one of the most used social media sites in the world. As of 2021, the platform boasts almost 200 million daily active users worldwide[1]. The character constraint on the text posts on Twitter lends itself well to spontaneous and organic interaction, with the majority of tweets being simply what the author is feeling at the moment. This type of communication is invaluable in the healthcare space. Patients spend little time with their care providers in comparison to their other frequent activities and it can be difficult for healthcare workers to address needs or feelings that patients often find uncomfortable disclosing to others.

Therefore, tweets, which are available through easily-acquired developer API tokens, represent an untapped resource for patient communication. Tweets from cancer patients have the potential to be especially valuable as feedback to healthcare providers, given the length of time that cancer patients are in contact with the same team of healthcare workers. Previous researchers [1, 2, 3, 4, 5, 6, 7, 8] have conducted sentiment and emotion analysis on tweets containing the keyword 'cancer' in several manners. For instance, by-hand analyses, examining network structure, using word embeddings, and applying rudimentary machine learning language models to the task.

In this thesis, more advanced language models will be implemented than in previous research. BERTweet [9] is a large-scale transformer model based on the RoBERTa [10] pre-training and fine-tuning process. RoBERTa is itself an improvement over the original BERT transformer model [11]. BERTweet is the first model of its kind to be released to other researchers for further improvements and novel applications; BERTweet was trained on 850 million English tweets collected from 2012 to 2019, as well as 23 million additional English tweets related to

[1] `https://www.oberlo.com/blog/twitter-statistics`

COVID-19, which prepares it well for downstream classification tasks on tweets. This pipeline of pre-training on a large text corpus and then fine-tuning the model for classification tasks is called transfer learning. It has been shown that pre-training is integral to model performance on downstream tasks, and it follows that pre-training a model on material which is similar to the texts in the downstream task will yield improved performance.

Using BERTweet, two downstream transfer learning tasks will be carried out. Both tasks come from Task 1 of the 12th International Workshop on Semantic Evaluation (SemEval 2018) [12], which involves predicting sentiment and emotion in tweets. Subtask 4 is to predict the sentiment of a tweet on a seven-point scale from negative to positive and subtask 5 is a multi-label classification task to predict the emotions found in a tweet from a list of 11 different emotion labels. Participants in the competition for these tasks were allowed to train their models on human-labeled data and then submit their predictions on the test data to be scored and placed on a leaderboard. Using BERTweet trained on these competition datasets, two research aims will be explored in this thesis. The first aim is to conduct sentiment and emotion analysis on tweets based on the type of cancer mentioned. It is hypothesized that tweets which mention cancers with higher mortality rates will be more negative on average than tweets which mention cancers with low mortality rates. The second aim is to compare the average sentiment and emotion labels of tweets mentioning cancer before and after the beginning of major lockdowns due to COVID-19. It has been discussed that lockdowns have lowered the hopefulness of cancer patients and introduced perceived barriers to care [13]. It is hypothesized that patients' reluctance to visit crowded hospitals and clinics has decreased their average tweet sentiment.

## 2 Previous Work

Using patient tweets to mine sentiment and analyze content types is not a novel practice in journals regarding heathcare communication. It is common to take random samples of tweets collected using keyword matching and annotate them by hand, noting the types of accounts that write the tweets and the subjective content of the tweets. Then analysis can be performed on the correlations between, for example, accounts owned by organizations and their proportion of tweets related to breast cancer awareness in contrast to the proportion tweeted by individuals [4]. A similar approach was conducted by Sutton et al., who collected 1.3 million tweets matching common cancer keywords, took a random sample of 3,000 tweets stratified by cancer diagnosis (e.g. lung, breast, etc.), and then coded tweets in this sample with markers for both content and account type [5]. Labels for content type included ‘awareness’, ‘prevention and risk’, ‘diagnosis’, and ‘treatment’. Labels for account type were ‘individual’, ‘media’, ‘organizational’, and ‘unknown’. An analysis of 1.7 million tweets collected during breast cancer awareness month by Thackeray et al. also quantified differences in tweet content between organizational and individual accounts [8]. Vraga et al. explored the amount of discussion related to female and male cancers (primarily breast cancer and prostate cancer/Movember) on both Instagram and Twitter over an entire year [2]. Expectedly, increases in activity matching keywords related to breast cancer occurred during October, which is breast cancer awareness month. Also, increases in posts related to Movember, a men’s healthcare movement emphasizing prostate cancer, testicular cancer, mental health, and suicide prevention, occurred during November, the month after which the movement is named. One finding of note is that Movember was mentioned much more frequently on Instagram than Twitter, which the authors attribute to the selection of moustaches and other facial hair as the visual representation of support for Movember, something that is not possible for visuals that directly show breast and reproductive organs.

Other papers included sentiment classification in their analysis of text within tweets. Crannell et al. separated tweets by cancer diagnosis [1]. Each set of diagnosis-specific tweets was further divided and sets for each individual patient were created. Then, happiness values for each patient's tweets were calculated using a quantitative hedonometric analysis with the Mechanical Turk (LabMT) word list. LabMT is a word happiness list of the most frequently occurring ∼10k English words based on several large corpora. The average happiness value for a single tweet, therefore, is a weighted arithmetic mean of each word's frequency and the word's corresponding average happiness score from LabMT. One important note by Crannell et al. is that the research is limited by the keyword "cancer"; diseases such as melanoma, leukemia, and lymphoma are excluded in the cases where the user does not also include the word "cancer" in the tweet [1]. Hedonometric analysis via LabMT was also used by Clark et al. to conduct a longitudinal study of sentiment in tweets which contain keywords "breast" and "cancer", in which important dates pertaining to legislature and awareness events are taken into consideration [3].

While hedonometric analysis with resources such as LabMT is relatively fast and easy, the approach does not allow sentiment to be mined from delicate semantic usages of language. LabMT does not take into account the order of words, the combination of individual words into phrases, or idiomatic expressions. A better approach is to use a lexicon with rules to account for these more advanced semantics. Gabarron et al. use the SentiStrength algorithm to compare sentiment values between tweets mentioning type 1 and type 2 diabetes [14]. SentiStrength, like LabMT, has a word list with positive and negative terms classified with a sentiment strength value from 2 (slightly positive/negative) to 5 (very positive/negative). Unlike LabMT, however, SentiStrength modifies these per-word sentiment values with a spelling correction algorithm, a booster word list (e.g. very, extremely), a negating word list (e.g. not), boosting strength of words with repeated letters, an emoticon list, and inclusion of punctuation in the calculation of sentiment [15]. Sewalk et al. use VADER to track longitudinal trends by geographic location in tweets which were collected matching a set of keywords related to medical patient experiences [16]. VADER (Valence Aware Dictionary and sEntiment Reasoner) is a modern version of SentiStrength. It uses a sentiment lexicon with additional rules which modify sentiment values in texts, but it has more rules than SentiStrength, and more nuanced rules, as well [17]. For example, VADER examines tri-grams preceding words with large sentiment values to detect negation, not just the presence of negating words like "not". Wolny takes the emoji processing of algorithms such as SentiStrength and VADER to the extreme, classifying emotions in tweets based solely on emojis and emoticons [18]. Wolny argues that only a limited set of emoticons is needed for effective emotion analysis, because the top 20 emoticons account for a large amount of tweets.

Others have also taken into account the network structure of Twitter replies and retweets to engineer features used for tweet classification. Retweets, replies, and mentions between users can easily be adapted into large directed graphs. Volkova et al. incorporate both tweet text and social network graph features into a single deep learning model to label tweets and Twitter accounts as propaganda, hoax, clickbait, and satire [19]. Garimella et al. use network graphs to quantify controversy on social media. The authors built graphs for different topics, based on hashtags, then partitioned each graph into exactly two components [20]. To measure controversy, the degree of separation between the two pieces was quantified with a formula called random walk controversy (RWC), coined by the authors. Sentiment analysis was also used to quantify the difference between two sides of a controversial topic and sentiment disparity between controversial and non-controversial topics. Wang et al. analyzed emotions in a month's worth of tweets which contained cancer-related hashtags [6]. Correlation between certain graphical measures and emotion in tweets was explored, and the authors found that tweets with joy, sadness, and hope received more likes, whereas tweets with joy and anger were more retweeted. Himelboim et al. used network analysis to discover clusters of users in graphs created from tweets collected pertaining to either breast or prostate cancer [7]. At the center of these clusters, which they call hubs, lie the most followed users. The descriptions in the user profiles of these hubs were used to categorize these popular users as either media, academic organization, health organization, grassroots (individual/blogs), or celebrity. While the inclusion of network analysis in text classification tasks, especially sentiment analysis, has been shown to offer marginal improvements in accuracy, its largest benefit is the ability to visualize data and transform tweets and text data into spatial data that can be made into insightful figures.asdlkfj;

## 3 Methods

After discussion of methods for sentiment and social network analysis used in past years, we now approach the state-of-the-art. Sequential models such as Long Short-Term Memory (LSTM), Bidirectional LSTM (Bi-LSTM), and Recurrent Neural Network (RNN) pass and update a hidden state from one token to the next as they process a sequence, and the final hidden state contains the representation of the entire sequence. The sequential nature of the calculation of these hidden states makes pretraining LSTM and RNN models very computationally expensive. In an experiment to boost the performance of these models, the attention mechanism was developed [21], which takes hidden states at all stages of a recurrent network into account to reach a final representation of the entire sequence. Using a Bi-LSTM architecture equipped with a multi-layer self-attention mechanism, Baziotis et. al. achieved competitive results in the SemEval-2018 Task 1 competition, the same one being used in this thesis, by pretraining on 550m English tweets [22]. In 2018, Vaswani et al. proposed the transformer model, which relies entirely on self-attention to draw global dependencies between input and output [23]. This model architecture allows for significantly more parallelization compared to sequential and recurrent models, dramatically decreasing the amount of time needed to pretrain on large text corpora. In the recent past, deep learning models following the transformer architecture have achieved state-of-the-art performance in many natural language processing tasks. In 2019, Devlin et al. introduced BERT (Bidirectional Encoder Representations from Transformers), which improves upon the original transformer model by learning token representations in both directions [11]. To achieve this, BERT uses a revamped pretraining procedure which includes masked language model and next sentence prediction objectives. Many BERT models pretrained on a variety of texts and languages are available to the public. This creates an easy ‘plug-and-play’ approach to using these pretrained models for downstream NLP tasks such as text classification and question answering. After BERT was introduced, Liu et al. developed RoBERTa, a Robustly optimized BERT pretraining approach [10]. RoBERTa is the result of the authors' experimentation with hyperparameters while pretraining BERT. It was found that BERT was significantly undertrained, and that with some tweaks, it is able to outperform newer and even larger transformer models. Pretraining optimizations in RoBERTa include dynamic masking, large mini-batches, larger byte-pair encodings, and using full sentences across documents. Like BERT, many pretrained variations of RoBERTa are available online.

In this thesis, the transformer model BERTweet will be used for sentiment analysis and emotion classification. BERTweet is a large-scale BERT model which was pretrained using the RoBERTa procedure on a corpus of 850 million English tweets collected over several years [9]. Because downstream NLP tasks in this thesis will involve only English tweets, it is advantageous to have such a large pretrained model exposed to the type of language used in the tweets. It is expected that accuracy of text classification on tweets would not be as high if the model was pretrained on text such as movie reviews or Wikipedia articles, which is the case with many publicly available pretrained models. Nguyen et al. declare that BERTweet is the first public large-scale pretrained language model for English Tweets.

To conduct sentiment analysis for individuals and based on cancer type, the datasets for SemEval-2018 Task 1 were used to create a model which could generate predictions for tweets collected independently of the competition datasets [12]. Task 1 involved predicting affect in tweets in several ways, including emotion classification and valence ordinal classification, which apply to this thesis. SemEval-2018 Task 1 subtasks were available in English, Spanish, and Arabic, as mentioned above, only English sets were used in this thesis.

SemEval-2018 Task 1-4 is valence ordinal classification in English, meaning predicting the sentiment of a tweet on a seven-point scale from negative to neutral to positive. For this task, BERTweet was fine-tuned using the train and dev datasets for 30 epochs with batch sizes of 32 and a learning rate of 0.00005. All other hyperparameters were left to the default values defined for classification models in the Simple Transformers API[1], which was used to initialize, train, and predict labels with BERTweet. Simple Transformers works with a large number of pre-trained models available through HuggingFace's model repository[2], including RoBERTa and BERTweet. Using this fine-tuning procedure, BERTweet achieved first place in the SemEval-2018 Task 1-4 leaderboard[3], with a Pearson Correlation Coefficient of 0.856, beating the next best of 0.836.

SemEval-2018 Task 1-5 is multi-label emotion classification in English. This task is, based on the text content of a tweet, to assign any number out of 11 different emotions: anger, anticipation, disgust, fear, joy, love, trust, surprise, sadness, pessimism, and optimism. For example, a tweet may be given labels of anticipation and optimism, while another may be assigned anger, disgust, and pessimism. The evaluation metrics for this task were accuracy and F1 score, both micro-averaged and macro-averaged. BERTweet was fine-tuned on the dev and train datasets for this task using the same procedure as in Task 1-4. While BERTweet's accuracy is ranked fifth, with 0.587 versus first place at 0.604, BERTweet's micro-averaged and macro-averaged F1 scores are in third and fourth place respectively, with 0.706 and 0.575, respectively, behind first place at 0.713 and 0.584, respectively.

Using the models produced by fine-tuning base BERTweet on SemEval-2018 Tasks 1-4 and 1-5, predicted labels were generated for collected tweets. For tweet collection, the Twython library was used as a pure Python wrapper around the Twitter streaming API. Tweets were collected over the entire years of 2019 and 2020 which matched a set of cancer related keywords based on the approach by Clark et al. [3] to exclude horoscope and astrology-related tweets. This gives over one year of COVID-19-free tweets and nine months of tweets which potentially mention COVID-19. Filtering by Twitter handle was also needed to remove tweets

[1] `https://simpletransformers.ai/`

[2] `https://huggingface.co/`

[3] `https://competitions.codalab.org/competitions/17751#results`

made by automated accounts. Using these predictions, longitudinal analysis was conducted to answer the research questions addressed in the introduction.

# 4 Results

## 4.1 Sentiment Analysis

To conduct analysis to address the research aims discussed in the introduction, sentiment predictions of individual tweets were averaged by day to reach a daily sentiment value for every day in 2019 and 2020. These values were then plotted in a connected line in order to reveal trends over the past two years. The total tweet set was also divided into smaller sets which each mentioned a different type of cancer. Cancer types used to make subsets of the total data were breast, prostate, skin, colorectal, lung, and pancreatic. It is important to note that the tweet sets labeled with a specific cancer type are subsets of the total data set, but the total data set labeled 'all cancers' in the figures is not simply the sum of the individual tweet sets, but it is all tweets which matched the single keyword 'cancer'. As mentioned by Crannell et al. [1], the usage of 'cancer' as the only keyword when collecting tweets likely causes under-representation for diseases such as melanoma, leukemia, and lymphoma, where patients are not as likely to use 'cancer' in their tweet. It is also very difficult to filter out tweets using 'cancer' as a slang term to describe something in an intensely negative way, such as with phrases like "⟨person⟩ is cancer" or "⟨thing⟩ is a cancer on society."

The plot using the raw sentiment data is found in Figure 4.1. The title 'raw data' refers to the lack of any smoothing or averaging on this plot. As detailed in the methods section, sentiment values were predicted on a seven-point scale, with an integer of $-3$ being the most negative, 3 being the most positive, and 0 being neutral. As can be seen, this plot is very noisy and sentiment fluctuates often for all cancer tweet sets. To make this data easier to read, exponential smoothing was applied to the original plot and this representation can be found in Figure 4.2. Because of the high level of noise in the raw data sets, exponential smoothing will be used in all longitudinal plots in the rest of this thesis. In addition, the average sentiment of all tweets over 2019-2020 for each set of tweets is presented in Table 4.1 along with the 5-year relative survival rate of each corresponding type of cancer for diagnoses in the years 2010-2016 across all sexes (where applicable), ages, and races in the United States. Data for mortality rates is provided by the National Cancer Institute[1]. It is interesting to note that prostate cancer, with the highest survival rate of 98.3%, has the highest average sentiment of $-0.5$. Conversely, pancreatic cancer, with the lowest survival rate of 10.5%, does not have the lowest sentiment value; however, the lowest sentiment value of $-1.07$ belongs to lung cancer, with the second-lowest survival rate of the cancers listed. These findings partially support the hypothesis that average sentiment for tweets mentioning a specific type of cancer is correlated with the mortality of that cancer type. All tweets matching the keyword 'cancer' have an average sentiment of $-0.94$, which is to be expected given the negative connotation of the word, regardless of the context in which it is used.

In Figure 4.2, several clear longitudinal trends and events can be seen. First, there is a notable decrease in sentiment across all cancer-related tweets in late March 2020. This decrease corresponds with the beginning of nationwide lockdown initiatives and stay-at-home mandates. This clear increase in negativity gives hard evidence supporting thoughts expressed by Norman Sharpless, current director of the National Institute of Cancer. In an editorial article from June 2020 he explored the possible increase in future cancer-related deaths due to the COVID-19 pandemic [13]. Sharpless explained that the fear of contracting the coronavirus in healthcare settings was dissuading people from the screening, diagnosis, and treatment for non–COVID-19 diseases such as cancer. Also, at many hospitals, so-called 'elective' cancer treatments and surgeries were being deprioritized to preserve clinical capacity for COVID-19 patients. For example, some patients began to receive less intense chemotherapy and/or radiotherapy, and in other cases, patients'

[1] `https://seer.cancer.gov/csr/1975_2017/browse_csr.php`

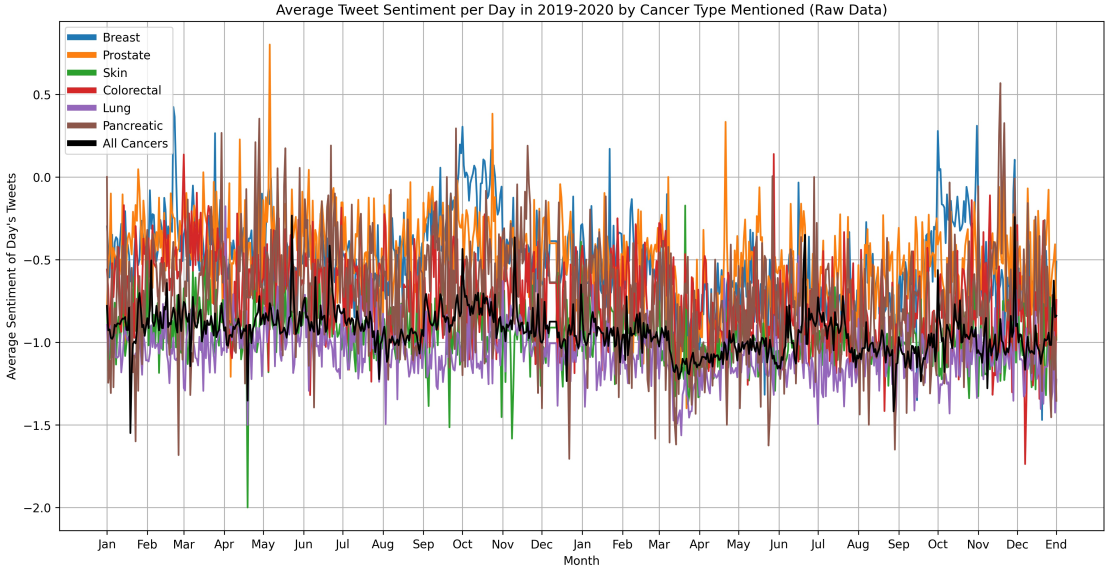

**Figure 4.1**: Sentiment over time in 2019 and 2020, by type of cancer mentioned. Raw data with no smoothing. 0 is neutral sentiment, −3 is strongly negative, and 3 is strongly positive.

| Cancer | Sentiment | Survival (%) |
|---|---|---|
| Prostate | -0.50 | 98.3 |
| Skin | -0.96 | 95.0 |
| Breast | -0.51 | 91.4 |
| All | -0.94 | 69.5 |
| Colorectal | -0.73 | 66.1 |
| Lung | -1.07 | 21.8 |
| Pancreatic | -0.73 | 10.5 |

**Table 4.1**: Average sentiment of all tweets combined in 2019-2020 by type of cancer mentioned. Sentiment value of 0 is neutral, −3 is very negative, and 3 is very positive. Survival is 5-year relative survival (percent) for diagnoses between 2010-2016 of the corresponding cancer type, across all ages, races, and sexes (where applicable).

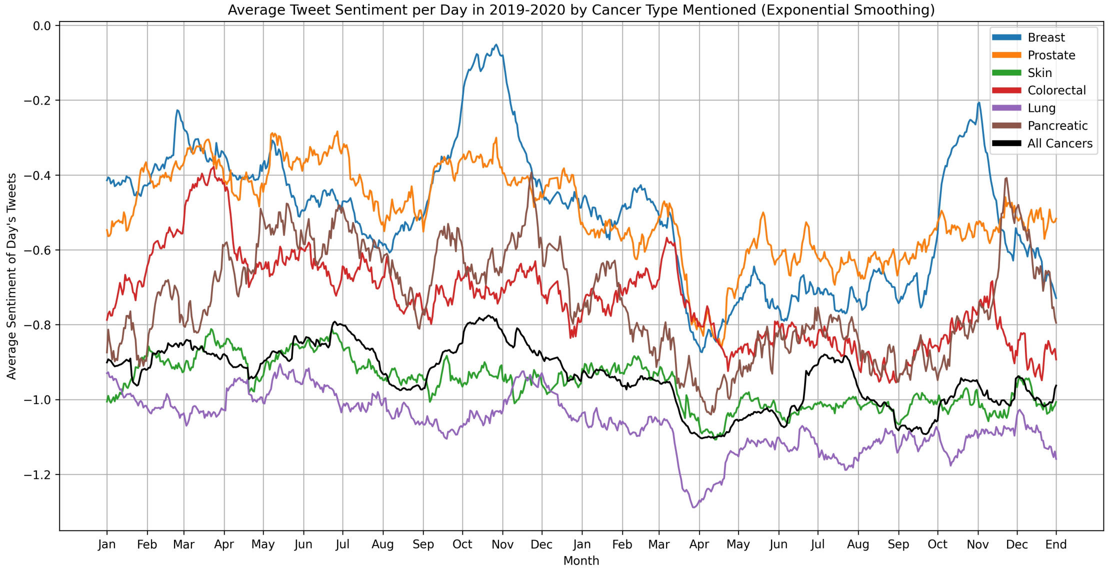

**Figure 4.2**: Sentiment over time in 2019 and 2020, by type of cancer mentioned. Exponential smoothing applied to data. 0 is neutral sentiment, $-3$ is strongly negative, and 3 is strongly positive.

operations to remove a newly detected tumor were delayed. After the sentiment drop seen in Figure 4.2, sentiment values for all cancer types rebounded, but not quite to the levels seen in the previous year of 2019.

Common cancer awareness events and holidays can also be seen represented by positive spikes in sentiment in Figure 4.2. For tweets mentioning breast cancer, sentiment becomes much more positive during the month of October, which is National Breast Cancer Awareness Month, than the rest of the year. Breast cancer awareness month is commonly celebrated with support for current breast cancer patients and breast cancer survivors across social media. The highest sentiment values across both 2019 and 2020 for pancreatic cancer occur on the third Thursday of November, which is World Pancreatic Cancer Day. For tweets mentioning colorectal cancer, the highest sentiment values across the year occur in the month of March, which is National Colon Cancer Awareness Month. In the next section, longitudinal plots will be displayed based not on sentiment, but on the frequency of classification of tweets with a specific emotion label. Looking at a range of different emotion labels instead of a simple scale of sentiment from positive to negative gives a more nuanced view of the trends and behaviors of Twitter users mentioning cancer.

### 4.2 Emotional Analysis

Eleven emotions were possible when labeling tweets: trust, surprise, sadness, pessimism, optimism, love, anger, anticipation, disgust, fear, and joy. Any number of these emotions could be assigned to a given tweet; for example, a tweet could be labeled with ‘love’ and ‘optimism’ or ‘disgust’,‘fear’, and ‘sadness’. The emotions of trust, surprise, and love will not be explored in this analysis due to the extremely low frequency with which these labels were applied.

In the last section, average sentiment of tweets per day was used to show longitudinal trends and indicate the effects of notable events on tweet content. A time-series plot showing the proportion of the day’s tweets labeled as fearful, rather than the average sentiment, paints a much more dramatic picture of the impact of stay-at-home mandates on tweets mentioning cancer. In Figure 4.3, a very large spike is seen across all tweet sets in late March 2020, corresponding with the start of lockdown protocols in the United States. Across all cancer-related tweets, it can be seen that fearful language stayed more frequent after March 2020 than in the previous year of 2019. Similarly, in Figure 4.4, a spike in pessimism is seen in March 2020, and levels of pessimism after this spike appear higher, on average, than in all of 2019. On a more positive note, the plot for optimism excellently shows the occurrence of cancer awareness events and their celebration on social media. In Figure 4.5, notable increases in optimism for tweets mentioning breast cancer occur in the months of October, which, as mentioned before, is National Breast Cancer Awareness Month. An increase in optimism for pancreatic cancer tweets happens around the third Thursday of November, which is World Pancreatic Cancer Day, and a spike in optimism for colorectal cancer tweets happens in March,

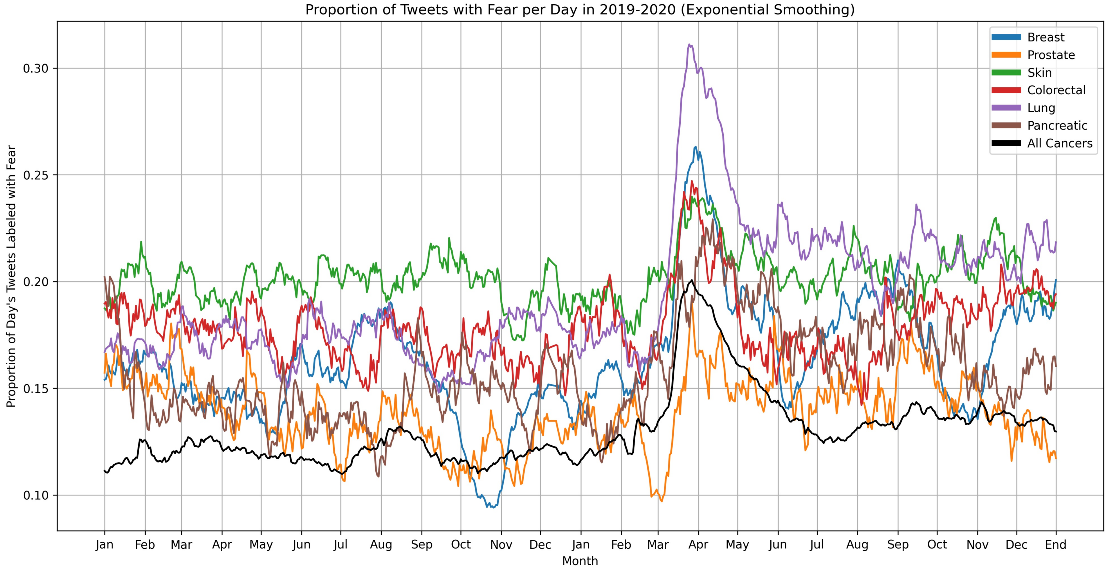

**Figure 4.3**: Proportion of each day's tweets labeled fearful in 2019 and 2020, by type of cancer mentioned. Exponential smoothing applied to data.

which is National Colorectal Cancer Awareness Month. All three of these markers were visible in Figure 4.2, but are exaggerated in Figure 4.5. It is worthy to note that higher peaks in optimism occurred in 2019 than in 2020 for all the cancer awareness events mentioned above.

To explore the hypothesis that the frequency of positive and negative emotions in cancer-related tweets is correlated with the mortality rate of the types of cancers mentioned in the tweets, Table 4.2 shows the average proportion of tweets assigned each emotion label across the entire two-year span of 2019-2020, divided by type of cancer mentioned in the tweet set. These values are placed next to the 5-year relative survival rate for each type of cancer for comparison, with the highest number for individual cancer types in bold for each emotion label. Interestingly, most of the emotions do not seem to be correlated with mortality rate at all. However, when correlation coefficients are calculated between survival rate and emotion prevalence, it is found that the emotion labels of sadness and anger have the strongest correlation with survival rate, with Pearson coefficients

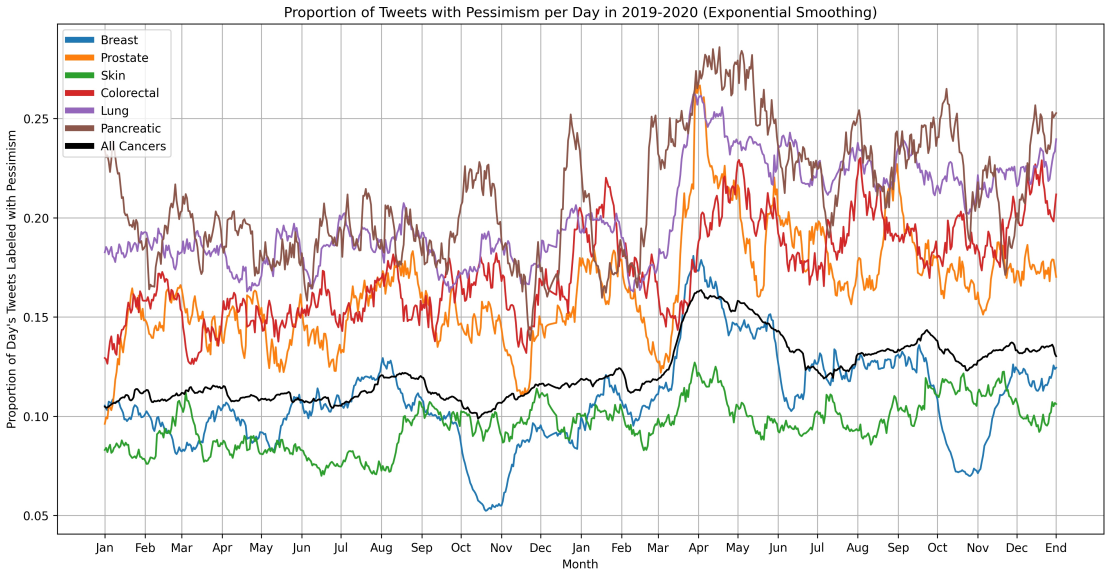

**Figure 4.4**: Proportion of each day's tweets labeled pessimistic in 2019 and 2020, by type of cancer mentioned. Exponential smoothing applied to data.

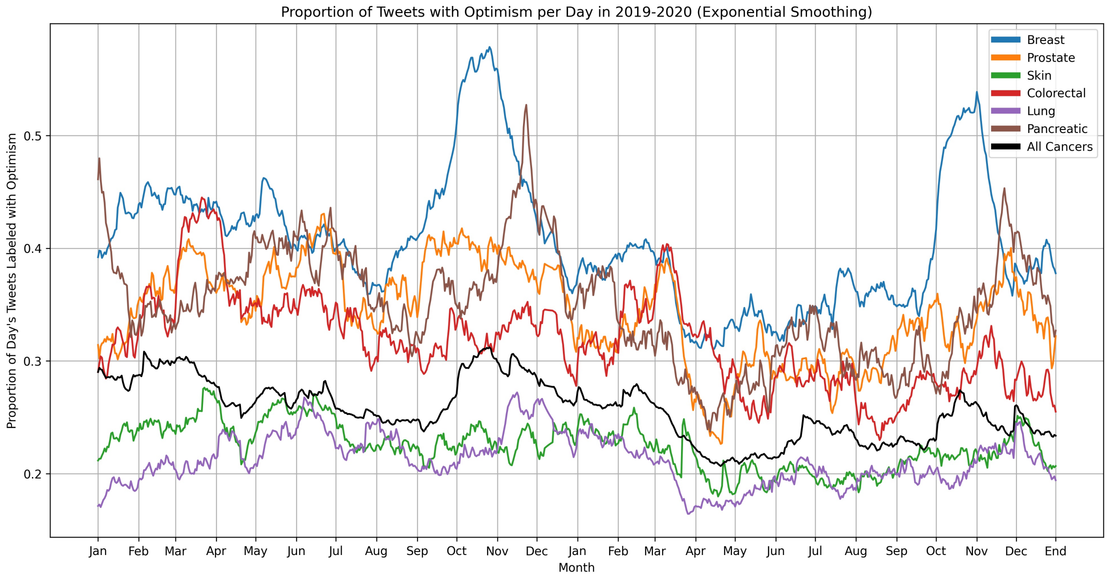

**Figure 4.5**: Proportion of each day's tweets labeled optimistic in 2019 and 2020, by type of cancer mentioned. Exponential smoothing applied to data.

| Cancer | Fear | Sadness | Anger | Disgust | Pessimism | Optimism | Joy | Anticipation | Survival (%) |
|---|---|---|---|---|---|---|---|---|---|
| Prostate | 0.139 | 0.375 | 0.120 | 0.241 | 0.165 | 0.344 | 0.236 | **0.374** | 98.3 |
| Skin | **0.202** | 0.391 | **0.269** | **0.423** | 0.096 | 0.225 | 0.188 | 0.205 | 95.0 |
| Breast | 0.165 | 0.503 | 0.155 | 0.245 | 0.109 | **0.407** | **0.323** | 0.306 | 91.4 |
| Colorectal | 0.179 | 0.493 | 0.138 | 0.243 | 0.175 | 0.316 | 0.218 | 0.287 | 66.1 |
| Lung | 0.195 | 0.500 | 0.234 | 0.375 | 0.203 | 0.213 | 0.162 | 0.172 | 21.8 |
| Pancreatic | 0.157 | **0.609** | 0.130 | 0.179 | **0.211** | 0.346 | 0.250 | 0.291 | 10.5 |
| All | 0.131 | 0.408 | 0.326 | 0.405 | 0.123 | 0.256 | 0.221 | 0.186 | 69.5 |

**Table 4.2**: Average proportion of tweets per day labeled with each emotion across both 2019 and 2020, by type of cancer mentioned in the tweet set. Survival is 5-year relative survival (percent) for diagnoses between 2010-2016 of the corresponding cancer type, across all ages, races, and sexes (where applicable).

of $r = -0.809$ and $r = -0.833$, respectively. This negative correlation coefficient means that tweets mentioning cancers with lower survival rates tend to contain more sad and angry language, which supports the hypothesis that tweets which mention more serious cancer diagnoses are likely to contain more negative emotion than cancer types with higher survival rates.

## 5 Conclusion

The vast number of tweets written every day by those with cancer go largely unnoticed by the healthcare teams that treat such patients. Asking cancer patients for participation in surveys about the emotions surrounding their care and their situation does not often yield many volunteers, and patients might even hide these feelings from their professional healthcare teams. The COVID-19 pandemic led to tightened restrictions around hospitals and healthcare locations, as well as nationwide stay-at-home orders. These circumstances significantly reduced the amount of cancer screening and diagnoses in the United States and introduced both perceived and real barriers to care for those already diagnosed with cancer [13]. It was hypothesized that an increase in negative sentiment and emotions in tweets related to cancer would be seen around March 2020, when COVID-19 lockdown policies went into effect around the country. It was also predicted that tweets mentioning cancer diagnoses with lower survival rates would use more negative language that other cancer diagnoses. Clear increases in pessimism and fear in tweets, linked to COVID-19 lockdown policies, was shown in longitudinal plots covering the years of 2019 and 2020, as well as a drop in the overall sentiment of tweets during that time. Out of all emotion labels possible for tweets collected, sadness and anger were the two most correlated with cancer diagnosis survival rate, which matches the prediction that negative emotions in tweets and survival rate are inversely correlated.

In future iterations of this work, human-coded subsets of the tweets collected marking them with relevance, sentiment, and emotion labels would improve the accuracy of the BERTweet model used. In particular, a machine learning classifier to filter out non-relevant tweets would be of great value when looking at tweets with the keyword ‘cancer’, because of the word’s common usage in astrological horoscopes and as a colloquial way to describe something as extremely negative. Although the analysis performed in this thesis were conducted in retrospect, it would be possible to conduct sentiment and emotion analysis of tweets collected in real-time, due to the streaming nature of the Twitter official API. Such analysis would allow healthcare providers to know within weeks or days the real effects of changes in policy or care availability, such as in the case of COVID-19 lockdowns. As discussed in the section regarding related work, easily available, state-of-the-art natural language processing and machine learning techniques are being under-utilized in the field of public health communications. This research aims to demonstrate the viability of integrating such techniques into the existing longitudinal analysis pipelines of many public health researchers and to showcase insights that these machine learning models are capable of providing.

## Bibliography

[1] W. C. Crannell, E. Clark, C. Jones, T. A. James, and J. Moore, "A pattern-matched twitter analysis of us cancer-patient sentiments," *Journal of Surgical Research*, vol. 206, no. 2, pp. 536–542, Dec 2016. [Online]. Available: https://doi.org/10.1016/j.jss.2016.06.050

[2] E. K. Vraga, A. Stefanidis, G. Lamprianidis, A. Croitoru, A. T. Crooks, P. L. Delamater, D. PFOSER, J. R. Radzikowski, and K. H. Jacobsen, "Cancer and social media: A comparison of traffic about breast cancer, prostate cancer, and other reproductive cancers on twitter and instagram," *Journal of Health Communication*, vol. 23, no. 2, pp. 181–189, 2018, pMID: 29313761. [Online]. Available: https://doi.org/10.1080/10810730.2017.1421730

[3] E. Clark, T. James, C. A. Jones, A. Alapati, P. Ukandu, C. Danforth, and P. Dodds, "A sentiment analysis of breast cancer treatment experiences and healthcare perceptions across twitter," *ArXiv*, vol. abs/1805.09959, 2018.

[4] M. S. Sedrak, M. M. Salgia, C. Decat Bergerot, K. Ashing-Giwa, B. N. Cotta, J. J. Adashek, N. Dizman, A. R. Wong, S. K. Pal, and P. G. Bergerot, "Examining public communication about kidney cancer on twitter," *JCO Clinical Cancer Informatics*, no. 3, pp. 1–6, 2019, pMID: 30860867. [Online]. Available: https://doi.org/10.1200/CCI.18.00088

[5] J. Sutton, S. C. Vos, M. K. Olson, C. Woods, E. Cohen, C. B. Gibson, N. E. Phillips, J. L. Studts, J. M. Eberth, and C. T. Butts, "Lung cancer messages on twitter: Content&#xa0;analysis and evaluation," *Journal of the American College of Radiology*, vol. 15, no. 1, pp. 210–217, Jan 2018. [Online]. Available: https://doi.org/10.1016/j.jacr.2017.09.043

[6] J. Wang and L. Wei, "Fear and hope, bitter and sweet: Emotion sharing of cancer community on twitter," *Social Media + Society*, vol. 6, no. 1, p. 2056305119897319, 2020. [Online]. Available: https://doi.org/10.1177/2056305119897319

[7] I. Himelboim and J. Y. Han, "Cancer talk on twitter: Community structure and information sources in breast and prostate cancer social networks," *Journal of Health Communication*, vol. 19, no. 2, pp. 210–225, 2014, pMID: 24111482. [Online]. Available: https://doi.org/10.1080/10810730.2013.811321

[8] R. Thackeray, S. H. Burton, C. Giraud-Carrier, S. Rollins, and C. R. Draper, "Using twitter for breast cancer prevention: an analysis of breast cancer awareness month," *BMC Cancer*, vol. 13, no. 1, p. 508, Oct 2013. [Online]. Available: https://doi.org/10.1186/1471-2407-13-508

[9] D. Q. Nguyen, T. Vu, and A. Tuan Nguyen, "BERTweet: A pre-trained language model for English tweets," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations.* Online: Association for Computational Linguistics, Oct. 2020, pp. 9–14. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.2

[10] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pre-training approach," *ArXiv*, vol. abs/1907.11692, 2019.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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).* Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://www.aclweb.org/anthology/N19-1423

[12] S. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko, "SemEval-2018 task 1: Affect in tweets," in *Proceedings of The 12th International Workshop on Semantic Evaluation.* New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 1–17. [Online]. Available: https://www.aclweb.org/anthology/S18-1001

[13] N. E. Sharpless, "Covid-19 and cancer," *Science*, vol. 368, no. 6497, pp. 1290–1290, 2020. [Online]. Available: https://science.sciencemag.org/content/368/6497/1290

[14] E. Gabarron, E. Dorronzoro, O. Rivera-Romero, and R. Wynn, "Diabetes on twitter: A sentiment analysis," *Journal of Diabetes Science and Technology*, vol. 13, no. 3, pp. 439–444, 2019, pMID: 30453762. [Online]. Available: https://doi.org/10.1177/1932296818811679

[15] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, "Sentiment strength detection in short informal text," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 12, pp. 2544–2558, 2010. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.21416

[16] K. C. Sewalk, G. Tuli, Y. Hswen, J. S. Brownstein, and J. B. Hawkins, "Using twitter to examine web-based patient experience sentiments in the united states: Longitudinal study," *J Med Internet Res*, vol. 20, no. 10, p. e10043, Oct 2018. [Online]. Available: http://www.jmir.org/2018/10/e10043/

[17] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, no. 1, May 2014. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/14550

[18] W. Wolny, "Emotion analysis of twitter data that use emoticons and emoji ideograms," in *ISD*, 2016.

[19] S. Volkova, K. Shaffer, J. Y. Jang, and N. Hodas, "Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on Twitter," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers).* Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 647–653. [Online]. Available: https://www.aclweb.org/anthology/P17-2102

[20] K. Garimella, G. D. F. Morales, A. Gionis, and M. Mathioudakis, "Quantifying controversy on social media," *Trans. Soc. Comput.*, vol. 1, no. 1, Jan. 2018. [Online]. Available: https://doi.org/10.1145/3140565

[21] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2016.

[22] C. Baziotis, A. Nikolaos, A. Chronopoulou, A. Kolovou, G. Paraskevopoulos, N. Ellinas, S. Narayanan, and A. Potamianos, "NTUA-SLP at SemEval-2018 task 1: Predicting affective content in tweets with deep attentive RNNs and transfer learning," in *Proceedings of The 12th International Workshop on Semantic Evaluation.* New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 245–255. [Online]. Available: https://www.aclweb.org/anthology/S18-1037

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *ArXiv*, vol. abs/1706.03762, 2017.

[24] R. Gaonkar, H. Kwon, M. Bastan, N. Balasubramanian, and N. Chambers, "Modeling label semantics for predicting emotional reactions," *ArXiv*, vol. abs/2006.05489, 2020.

[25] N. Kant, R. Puri, N. Yakovenko, and B. Catanzaro, “Practical text classification with large pre-trained language models,” *ArXiv*, vol. abs/1812.01207, 2018.