Introduction

The COVID-19 pandemic has put an unprecedented strain on the healthcare system in the USA. For our proposed project, our objective is to measure how emotions and morale have changed amongst nurses before and during the COVID-19 pandemic. We analyzed this via Reddit data in the major nursing subreddit (r/nursing) using natural language processing. Through this process, we hope to gain a better understanding of the mental strain experienced by nurses as a result of this pandemic.

Survey

Prior studies that measure nurse morale have mainly been done using qualitative surveys. For example, a study of nurses in Wuhan used a qualitative survey distributed over a popular chat app to measure the risk of developing mental health issues (Lui et al., 2020). Similarly, a 2001 study on nurses used a survey to determine which factors lead to the most stress (McGowan, 2001). While both studies have insights that can aid us in our own study, surveys have notable limits. The sample size of survey-takers is often limited, and survey questions can have unintentional bias that affects the answers given. Concerning emotion in particular, a survey-taker may present themselves differently in a survey than they would in casual online conversation. Venting online is an emotionally driven act where those strong feelings will bleed into the words; taking a survey, however, may sacrifice strong emotional language in an attempt to maintain professionality.

Natural language processing is commonly used in many fields to measure sentiment, emotion, and find patterns in text data. The typical data processing pipeline for natural language processing is the bag of words approach, where a text document is tokenized, cleaned and standardized, and vectorized for use in algorithms (Bonaccorso, 2017). With this data, we can use topic modeling to find common topics amongst documents; latent sentiment analysis, based on matrix decomposition, and latent dirichlet allocation, a probabilistic approach, are the most common models (Jelodar, 2019). Topic coherence can then be used to judge the quality of different numbers of topics (Stevens et. al, 2012). By applying these techniques to our Reddit corpus, we can quickly analyze thousands of posts to better understand healthcare employee morale before and during the pandemic.

In addition, we want to leverage emotion detection algorithms to generate insights into the emotional state of nurses over time. COVID-19 has inspired other researchers to perform emotion detection on Reddit to better understand how the emotional content of posts in general has changed. Authors in one study utilized the NRC Emotion Lexicon, which rates 10,000 words on 8 emotions on a scale from 0 to 1, to find the dominant emotion in millions of Reddit comments (Mohammad, 2018). This dictionary based approach is relatively simple and easy to implement, but there is room for improvement in that it does not take into account negation nor amplifiers (Basile et al., 2021). In our work, we intend to use the NRC Emotion Lexicon and a supervised machine learning technique, as recent research has shown that the combination of the two is the most effective emotion detection method (Acheampong et al., 2020).

Detection of words and mental/physical health topics such as anxiety and depression have been done in multiple experiments such as: (Shen & Rudzicz, 2017), (Tadesse et al., 2019), (Guntuku et al., 2017), (Osadchiy et al., 2020), and (Coppersmith et al., 2018). Those papers have used techniques such as LDA which uses a probability distribution of topics across the topics in one document or Linguistic inquiry and Word Count (LIWC) which use dictionaries to generate a feature vector for each of the documents.

Proposed Method

Innovations

We intend to improve upon the emotion detection algorithm used in comparable research (Basile et al., 2021). While other research simply uses the NRC Emotion Lexicon to determine the dominant emotion in a Reddit post, we will use a texts' ratings in the lexicon's 8 emotion categories as a feature in a supervised machine learning model. We will also leverage Latent Sentiment Analysis (LSA) and Doc2Vec as additional features to emphasize key words in our corpus and to numerically represent a text's concept, including a consideration of its word order. By using these features with a supervised learning model, we are hoping to label our Reddit posts more accurately than comparable research has.

Most of the papers that we read in preparation for this project used surveys or Twitter as a source of emotion detection within the medical persons. Our approach is to use Reddit as it is faster to collect data, and it has a wider range of words for each post or comment that can be much more declarative than the tweet limitations. Depending on social media over surveys will help the experiment further to check for new inputs much easier and compare over the previous time frame.

Another way that our approach improves upon past studies is our plan to visualize our data as an interactive timeline. By publishing an interactive visualization of our data to the web, we hope to not only present our findings to the public, but also to allow viewers to explore the data themselves and answer questions of their own.

Algorithms

For data collection, we used the PushShift Reddit API, which is an archive of all Reddit posts and comments. For this project, we collected all posts and comments from January 1st, 2019 to October 19th, 2021 on both the Nursing and the Medicine subreddits. For each subreddit, a for-loop applied over each month was used to collect all posts and comments pertaining to that month. Each month was put into a separate CSV file. Ultimately, we only utilized posts and comments from the Nursing subreddit, allowing us to go into more depth in our analysis.

For posts, we collected strings representing the author, title, body text, and url of the post, and integers representing the creation time and unique Reddit ID of the post. For comments, we collected strings representing the author and body text of the comment, and integers representing the creation time and unique Reddit ID of the comment. We also collected integers for the unique Reddit IDs linking the comment to the post it was found under, and any parent comment it was replying to.

Pre-Processing

In pre-processing, we relied on NLTK in addition to other steps. For each file as follows:

- Collect {id, author, created} of each record
- If the file is a post it will:
    - Collect data from selftext field and set the field type as "Post"
- If the file is a comment it will :
    - Collect data from body field and set the field type as "Comment"
- Set the field category to either "Medicine" or "Nurses" depending on the parent directory
- Tokenize the post/comment into words
- Remove words less than 2 characters and punctuation and Get the stem of each word

Store a json file under processed/Project_data with a similar hierarchy of the input files for the output of the processed file with the schema of: {id, category, type, author, url, created, words}.
We added another step to generate data frames and stats out of the json files. Output from this stage is stored in the GATech box that we use to share data for our project.

Emotion Detection

Our emotion detection algorithm leverages two text vectorization techniques, Latent Semantic Analysis (LSA) and Doc2Vec, and the NRC emotion lexicon to create rich features. LSA is a common text vectorization technique that has two components: TF-IDF and singular value decomposition (SVD). TF-IDF vectorizes a corpus, highlighting words that are featured frequently within a document, but not frequently amongst the entire corpus, and SVD performs dimensionality reduction, creating features with the most amount of variance explained. Doc2Vec is a text vectorization technique that leverages a simple neural network to create vectors that represent the concept and context of each document. Lastly, we created features based off of the 8 emotions in the NRC emotion lexicon, accounting for varying length of text by normalizing the scores across the 8 emotions.

These features were used to train supervised learning models that predicted the emotion of a text. These models included some that naturally work well on multiclass data, including Random Forest, Gradient Boosting (XGBoost), and Naive Bayes, and those where we needed to employ a one vs rest strategy, including Linear SVM and Logistic Regression. To find the best model, we leveraged cross validation to minimize the risk of overfitting, performed hyperparameter tuning, and then chose the model with the best accuracy. By utilizing a diverse set of features and trying many models, we hoped our chosen model would capture the nuance in each document, leading to stronger predictions than models in comparable studies.

Topic Modeling

For topic modeling, our group used a Latent Dirichlet Allocation (LDA) model, which is a probabilistic model, via the Python Gensim package. Our first step was to filter the documents to only include a length of 50 words or above after preprocessing. Because topic modeling tries to determine the topics of individual documents, short documents can throw off the model. Several runs were performed with different word length filters, and using human judgement of topic relevance, 50 words was the minimum to provide substantial topics. This decreased the size of the data to approximately 200,000 posts and comments.

We generated a word dictionary from the documents of the top 100,000 words that appeared more than 15 times and in less than 50% of the documents. We then generated bigrams from the words to determine common two-word phrases. To increase the relevance of words, we also performed filtering on the words themselves, eliminating all words appearing less than five times and eliminating words from the dictionary that appear in over 50% of documents. Similarly to the emotion detection section of our project, TF-IDF was used as well to highlight important yet uncommon words, which are the words most indicative of different topics.

To help determine the optimal number of topics, we used Coherence scores, which measure similarity between high-scoring words in topics. Generally, Coherence scores align with human perception of well-defined topics, though some human guidance is needed concerning more precise distinctions (e.g. deciding between seven and eight topics, for example, where coherence scores are more similar).

Chosen for speed, our Coherence metric used was the UMass metric, which looks at document co-occurrence. In other words, it assigns a higher score to models which have more words in the same topic appear in the same document. The values range from -14 to 14, with higher values having more high document co-occurrence. However, because the documents used here are on the shorter side (50 words minimum), the coherence scores likely skew lower because fewer words means less of an opportunity for word co-occurrence, even if those words are generally related. Thus, human judgement was used for choosing the final set of topics among similarly well-performing topic models. (Stevens et al., 2012)

For simplicity, posts and comments were then assigned to one topic each based on probability-likelihood. For example, a post that has a 90% likelihood of being topic 1 and a 10% likelihood of being topic 2 will be assigned topic 1.

Visuals

Our visuals have been created in Tableau Public, allowing us to easily publish interactive visualizations of our data to the web. For topic modeling, we made line graphs showing the number of posts/comments made in each topic over a timeline on the x-axis. Hovering over each line will display the word list generated by the topic modeling algorithm. For emotion detection, we created line graphs showing, for each month, the percentage of posts/comments with a certain emotion. These graphs allow viewers to easily see topic and emotion trends on Reddit over time. We also created an interactive emotion detection line graph, where clicking on a particular emotion in a specific month will bring up the top ten posts from that month associated with that emotion. Finally, we created line graphs combining our emotion detection data with our topic modeling data, showing the count of posts/comments associated with each emotion per topic. This displays the data in such a way that allows us to see if any topic trends and emotion trends correlate.

Experiments/ Evaluation

In our project we want to address the following questions: How have emotions in nursing Reddit posts changed over the course of the pandemic? How does this compare to pre-pandemic posts? What are the dominant topics in these posts and how have those changed over time? And lastly, is there a link between the dominant topics and emotional content?

Working with a huge data set of approx.1.8 million records has influenced the implementation of pre-processing data, so we tried to stage the steps of the preprocessing pipelines. Removing stop words is a normal step in data cleansing. However, in our case, negation (which is part of stop words) can influence the data, so we decided not to remove stop words. Some punctuations, like ” ”, were not cleaned up with removing punctuation, so we added a condition to leave only words with more than 2 characters.

One observation made during the data collection process was that while most months had similar posting and commenting rates, there was a notable spike during the beginning of the pandemic, particularly for comments. Though this still needs to be confirmed by modelling, our hypothesis is that COVID will be the most popular topic by far during this period between March and May 2020, and negative emotions will be dominant.

We initially started to remove duplicate words, but later we found that leaving duplicate words helped in topic modeling and enhanced the results especially with using bigrams.

Emotion Detection

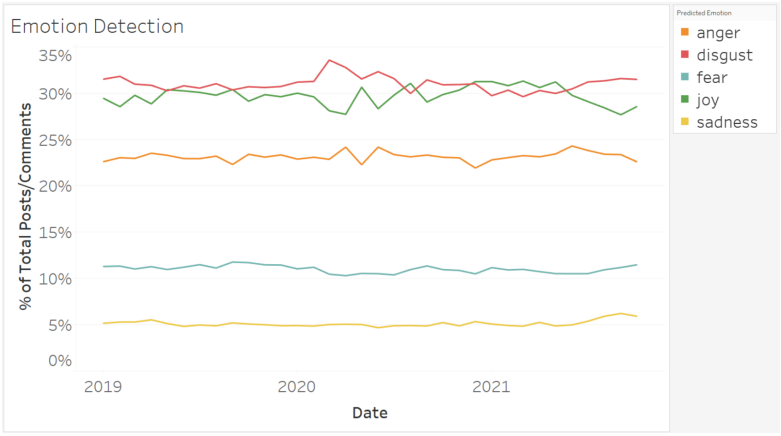
Our emotion detection model uses a well known training dataset, the ISEAR dataset, which has 8000 documents that are categorized as one of 7 emotions. We performed the same preprocessing steps on this dataset as we performed on our Reddit corpus. In addition, we filtered the ISEAR dataset to documents categorized as one of the emotions in the NRC Emotion Lexicon: anger, disgust, fear, joy, and sadness, so we could compare our improved model to a baseline model.

After creating the modeling features described in the algorithms section and performing hyperparameter tuning with 10-fold cross validation, we determined the strongest performing model was a Linear SVM model (see appendix for model comparison). Compared to the NRC lexicon model used in comparable research, our Linear SVM model has an accuracy that was 20% points higher (63.5% vs

43.1%). We used accuracy as our metric because the ISEAR dataset was balanced between the 5 classes and there isn't a compelling reason to more heavily penalize false negatives or positives.

After retraining the Linear SVM model on the entire dataset, we began preparing the Reddit corpus for modeling. For emotion detection modeling, each post and comment was considered its own document. We removed documents without any words in the NRC lexicon, as this was an indicator that the document did not have any emotional content, which our model was not trained to detect.

Graphing the results, the most obvious trend is that there's more volatility in the % of posts with a given emotion after the pandemic began. The volatility is not surprising, as there have been at least four waves in the US, some expected and others not, and each wave had their own narrative and root causes.



The model results showed surprising trends around joy, the sole positive emotion in our training set. The % of predicted joy posts reached its lowest levels in April 2020, during the peak of the first wave, and in September 2021, during the peak of the fourth wave. Interestingly, even considering pre-pandemic posts, the % of predicted joy posts reached its highest levels in December 2020 to May 2021, despite the worst of the pandemic in the US being December 2020 to January 2021. There are a few possible explanations for this. First, nurses started receiving vaccines in December 2020, so more posts may have had optimistic tones as an end was in sight. Second, the algorithm often interpreted emotional support and gratitude as joy, so users may have been supporting each other more frequently during the second wave.

Topic Modeling

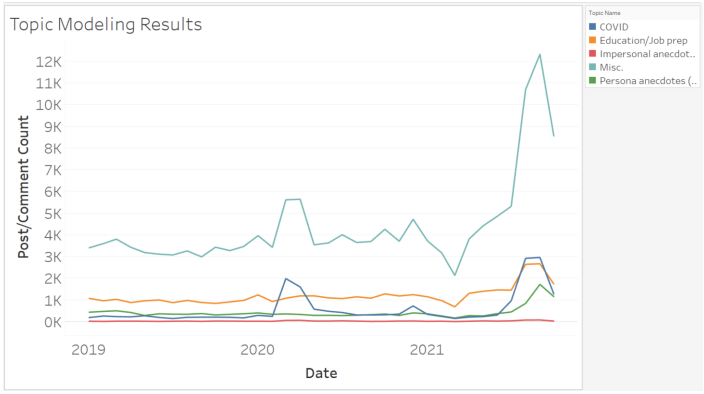
Using topic modeling, the two final models generated had seven and five topics, which had coherence scores of -2.009722747 and -4.078468722, respectively. These two models were chosen because of a combination of topic readability (i.e. each topic's words made logical sense) and a relatively good coherence score. (See appendix for a chart of all coherences.) Both models had clear topics for COVID, jobs/education, and anecdotes about patients and coworkers. However, the most striking difference between the two is that the seven-topic model created two topics that were exclusively Reddit bots. While it does make logical sense as Reddit bots tend to have similar post structures and word choice, it does reveal an oversight in our experiment.

The final model chosen was the five-topic one, as it had a cleaner visualization and did not have the bot influence that the seven-topic model does. The five-topic model's coherence score is also higher than the seven-topic model's coherence score, which shows that the former's top topic words have a higher co-occurrence.

Miscellaneous	feel, nurse, this, patient, work, just, been, know, like, what
COVID	vaccine, covid, people, this, mask, are, they, who, test, not
Education/job prep	program, job, nurse, school, year, ani, work, would, experience, current

Personal Anecdotes	she, her, was, him, his, had, night, patient, call, told
Impersonal Anecdotes	wear, scrub, blood, shoe, mask, use, needle, bag, skin, catheter

As expected, there was a significant spike in the COVID topic beginning March 2020, which coincided with the first wave of COVID. A smaller spike occurred around December 2020, which was near the introduction of the vaccine.



Additionally, the topic that was the most distinct from the others was Jobs/Education, which had a significant number of posts/comments with over 90% probability. This could be explained by the popularity of the topic itself, as Reddit has a significant younger population who would be likely to want to break into the nursing field and ask advice. Additionally, while other topics can have a considerable overlap (e.g. "vaccine" could hypothetically refer to either the COVID vaccine or a routine vaccine), posts about education are more likely to have terminology not found in other topics.

Conclusions and Discussions

Overall, we were pleased to find that the results of our modeling aligned with the timeline of COVID: we started seeing more negative emotions and the COVID topic in our results starting March 2020, we saw the COVID topic, but also more positive emotions, starting November 2020, when vaccinations drove optimism, and we saw the resurgence of the COVID topic and negative emotions starting August 2021, potentially suggesting burnout. These results align with the first, third, and fourth COVID waves in the United States, respectively, validating our work and the models' usefulness.

In future work, the results of emotion detection could be improved by using a more specific and current dataset. The ISEAR dataset was aggregated in the 1990s from students reporting situations where they felt the prompted emotion. As a result, these responses lack the medical context found on the nursing subreddit. Looking at responses for disgust in ISEAR, many respondents talked about bodily functions and anatomy; both of these topics are part of day to day work for nurses, so nurses are likely to use them in a different context than people outside the medical field.

The emotion detection algorithm could also be improved by trying additional pre-processing techniques. For example, topic modeling saw an improvement in results when using bigrams while emotion detection only used 1-grams (although Doc2Vec does consider the order of text).

As noted in the topic modeling results, one oversight in the project was the influence of bots on certain topic models. In future projects, filtering out bots would be a good practice to keep the results to purely people. Unfortunately, bots can be difficult to detect on Reddit as there is no universal flag to distinguish them, but using one of several bot databases available online can at least filter out the usernames of the most popular bots.

During the project, all team members contributed a similar amount of effort.

The link to our dashboard is here.

References

Acheampong, F. A., Wenyu, C., & Nunoo-Mensah, H. (2020). Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, *2*(7). https://onlinelibrary.wiley.com/doi/10.1002/eng2.12189

Basile, V., Cauteruccio, F., & Terracina, G. (2021). How Dramatic Events Can Affect Emotionality in Social Posting: The Impact of COVID-19 on Reddit. *Future Internet*, *13*(2). https://www.mdpi.com/1999-5903/13/2/29

Bonaccorso, G. (2017). *Machine Learning Algorithms* (pp. 242-287). Packt Publishing. https://ebookcentral.proquest.com/lib/gatech/reader.action?docID=4926962&ppg=257

Coppersmith, G., Leary, R., Crutchley, P., & Fine, A. (2018, August 27). Natural Language Processing of Social Media as Screening for Suicide Risk. *Biomedical Informatics Insights*, *10*. https://journals.sagepub.com/doi/full/10.1177/1178222618792860

Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2017, December). Detecting depression and mental illness on social media: an integrative review. *ScienceDirect*, *18*, 43-49. https://www.sciencedirect.com/science/article/pii/S2352154617300384?via%3Dihub

Jelodar, H., Wang, Y., Yuan, C. et al. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimed Tools Appl* 78, 15169–15211 (2019). https://doi.org/10.1007/s11042-018-6894-4

Lui, Z., Han, B., Jiang, R., Huang, Y., Ma, C., Wen, J., Zhang, T., Weng, Y., Chen, H., & Ma, Y. (2020). Mental Health Status of Doctors and Nurses During COVID-19 Epidemic in China. *The Lancet*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3551329

McGowan, B. (2001). Self-reported stress and its effects on nurses. *Nursing Standard*, *15*(42), 33-8. https://www.proquest.com/docview/219820833?fromopenview=true&pq-origsite=gscholar

Mohammad, S. (2018, May). Word Affect Intensities. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, *(LREC 2018)*. https://aclanthology.org/volumes/L18-1/

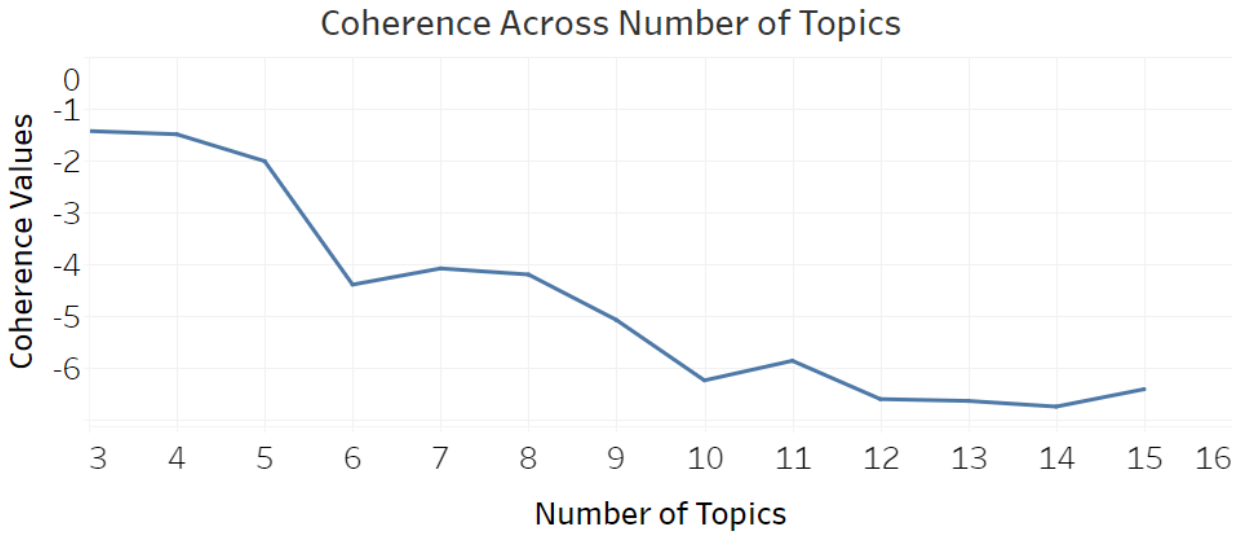
Osadchiy, V., Mills, J. N., & Eleswarapu, S. V. (2020, March). Understanding Patient Anxieties in the Social Media Era: Qualitative Analysis and Natural Language Processing of an Online Male Infertility Community. *Journal of Medical Internet Research*, *22*(3). http://www.jmir.org/2020/3/e16728/

Shen, J. H., & Rudzicz, F. (2017, August). Detecting anxiety on Reddit. *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology {---} From Linguistic Signal to Clinical Reality*, 58-65. https://aclanthology.org/W17-3107/

Stevens, K., Kegelmeyer, P., Andrzejewski, D., Buttler, D. (2012, July) Exploring topic coherence over many models and many topics. E*MNLP-CoNLL '12: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning,* 952–961. https://dl.acm.org/doi/10.5555/2390948.2391052

Tadesse, M. M., Lin, H., Xu, B., & Yang, L. (2019, April 04). Detection of Depression-Related Posts in Reddit Social Media Forum. *IEEE Access*, *7*, 44883-44893. https://ieeexplore.ieee.org/abstract/document/8681445

Appendix:



Emotion Detection Modelling Results:

Model	**Accuracy on K-folds**	**Accuracy on Validation Set**
Random Forest	63.0%	
XGBoost	65.6%	
Naive Bayes	58.4%	
Linear SVM	68.5%	63.5%
Logistic Regression	68.3%	
NRC Lexicon		43.1%