Using Deep Learning to Correlate Reddit Posts with Economic Time Series during the COVID-19 Pandemic

Philip Hossu[*] and **Natalie Parde**
Department of Computer Science
University of Illinois at Chicago
{phossu2, parde}@uic.edu

Abstract

The COVID-19 pandemic produced uniquely unstable financial conditions that permeated many sectors of the economy, creating complex challenges for existing financial models. In this investigation we tackle a subset of these challenges, seeking to determine the relationship between Reddit posts and actual economic time series during the economically tumultuous year of 2020. More specifically, we compute correlation between language used in Reddit's `/r/personalfinance` forum and initial or continuing unemployment claims in the United States throughout the year. We collect a novel dataset for the task, complete with annotations distinguishing between unemployment-specific and general employment inquiries. We also train a Convolutional Neural Network (CNN) based deep learning model to distinguish between these categories, achieving a maximum F1 score of 0.857. Finally, we compute Pearson correlation coefficients between these model predictions and real-world financial data relative to a number of reasonable baselines, yielding correlations of up to 0.796 for initial claims and 0.747 for continuing claims.

1 Introduction

The ever-increasing amount of online social media data, paired with recent advances in deep learning and natural language processing, enables a new generation of analysis previously unattainable for analyzing public response to a global pandemic and consequent financial crisis. In this investigation, we seek to shed light on how a specific subset of social media data can be used to draw parallels with real world economic trends. We select unemployment data for our case study. Our work leverages a deep learning model to discover strong correlations between Reddit's[1] `/r/personalfinance` posts and actual initial and continuing unemployment metrics in the United States during the anomalous, turbulent COVID-19 year of 2020. Our contributions are as follows:

- We create a novel dataset of 667 posts from `r/personalfinance` annotated with fine-grained categories corresponding to types of employment inquiries.
- We build a CNN-based model to discriminate between unemployment categories, achieving a binary classification F1 score of 0.857.
- We compute correlation scores between our model predictions and real-world employment data, achieving strong Pearson's correlation scores of 0.796 and 0.747 for initial claims and continuing claims, respectively.

To enable replication and encourage additional follow-up work, we make our data available upon request. The remainder of this paper is organized as follows. We begin by presenting a brief summary of related work which served as inspiration. We then describe our data collection and annotation procedure, as well as define and discuss training procedures for a CNN classification model. Predictions from the top performing model are subsequently extended to the full year of data, enabling us to compute correlations with real unemployment economic time series, focusing on both initial and continuing claims in the USA. We analyze and discuss these results before commenting on intriguing future work directions.

2 Related Work

Although a historically less utilized platform than Twitter in academic works, Reddit contains a wealth of data in the modern online social network age. Reddit reportedly has around 430 million active users,[2] with just under 50% of these users being based in the United States,[3] our target country for this investigation. The primary advantage of using Reddit data for analysis and modeling in the context of the work presented here is that users can opt in to various topic-focused discussion forums, ranging from general topics like `/r/food` or `/r/music`, to specific topic forums like `/r/whitesox` or `/r/CHIcubs`, giving researchers a significant amount of flexibility to investigate topics.

Within natural language processing research settings, the use of Reddit has been slowly increasing in popularity, with

[*] Contact Author

[1] https://www.reddit.com

[2] https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

[3] https://www.statista.com/statistics/325144/reddit-global-active-user-distribution/

Proceedings of the Third Workshop on Financial Technology and Natural Language Processing
(FinNLP@IJCAI 2021), pages 8-14, Online, August 19, 2021.

recent studies leveraging Reddit data for mental health discourse analysis [Choudhury and De, 2014; Valizadeh *et al.*, 2021], psychological personality prediction of users [Gjurković and Šnajder, 2018], and even recent COVID-19 symptom and sentiment detection [Murray *et al.*, 2020]. From an economic or financial perspective, one clear and prevalent use of social media text in academic research has been that of analyzing crowd sentiment for investment and trading strategies, whether related to equities or other asset classes. One recent example of this is the work of Wooley *et al.* (2019), where the authors focus on a number of research questions revolving around cryptocurrency trading prices and related Reddit discussion boards. The results and discussion from these and other works support our notion of using Reddit as a data source in our case study of unemployment analysis.

The `/r/personalfinance` sub-reddit specifically has indeed seen some limited mention in academic works; however, these works have largely been geared towards psychological mental health assessment of online users. Two such works are those of Shen and Rudzicz (2017) and Low *et al.* (2020), both which directly reference data from `/r/personalfinance` in their work — but as a control sub-reddit, rather than a target for determining anxiety or mental health.

3 Dataset

We systematically collected data from Reddit since no existing datasets were available for our specific needs, and we utilized existing unemployment data from official government sources. The following subsections explain this process in greater detail.

3.1 Text Data Collection

Reddit post data was collected using the PushShift library [Baumgartner *et al.*, 2020] in Python 3 from the `/r/personalfinance` sub-reddit using a search query from January 1 to December 31 of 2020. Reddit's API guidelines[4] allow for the collection and use of this data. While this API returns a sample of all posts, we filtered out any posts which had been removed from the site for any reason in order to respect the decision and privacy of users and site moderators. Similarly, we made the choice to not analyze any personal information related to the posters themselves. This yielded a total of 99,282 posts. One of the aspects which makes this sub-reddit especially interesting and useful for analysis is that a vast majority of posts are user-tagged with pre-defined categories, like "investing" or, more relevantly in this work, "employment" (4935 posts).

3.2 Annotation

For our predictive model and ultimate unemployment time series correlation, we labeled a sample of posts which were poster-tagged with "employment." Our annotation scheme is as follows:

- **Unemployment (U):** Individuals who are obviously unemployed, recently laid-off, and who have questions about their unemployment (e.g., regarding government benefits or otherwise). Other examples may include:
  - Job seekers posting questions related to filling in unemployment assistance forms
  - Individuals asking about late or missed unemployment assistance payments
- **Cut/Furlough (C):** Individuals who are furloughed or who have had hours or pay cut. Their employee status is limited, but they have not been fully laid-off. Other examples may include:
  - Company offering to bring back employees for reduced pay
  - Individuals filing for unemployment after being only temporarily let go
- **Employment (E):** Currently employed individuals who have questions about their current employment or a new employment offer; or, currently employed individuals debating leaving their current job for another job. Other examples may include:
  - Adding a part-time job
  - Considering or negotiating an offer while currently employed
- **Other (O):** Individuals not fitting in categories U/C/E, or hypothetical questions. Other examples may include:
  - Students having part-time on-campus jobs
  - Individuals willingly quitting their job while showing no obvious signs of financial need

To compute inter-annotator agreement, a shared set of 100 data samples of the "employment" tagged posts were double-annotated by two U.S. born, English speaking, graduate-level college-educated individuals, which we determined was sufficient background to accurately interpret English social media posts. One of the annotators was an author of this paper, and the other was an external volunteer. Following a training phase that involved iterating on the annotation guide to improve clarity, particularly with respect to vague edge cases, a Cohen's Kappa score of 0.88 was obtained, which denotes substantial agreement [Landis and Koch, 1977]. Given this quantitative measure of confidence in the annotation guide, the remaining instances were single-annotated. The final dataset comprised 667 instances.

3.3 Unemployment Data

Unemployment time series data was obtained from one primary source — the St. Louis Federal Reserve's F.R.E.D. (Federal Reserve Economic Data) portal.[5] Both raw weekly initial ("ICSA") and continuing claims ("CCSA") data are available for public download. *Initial claims* refers to a government tally of individuals seeking unemployment assistance for the first time, whereas *continuing claims* refers to the number of people who have already filed an initial claim, and after a week of joblessness, must continue filing claims for government aid. To maintain temporal alignment with our Reddit posts, we downloaded unemployment data corresponding to the time frame from January 1 to December 31 of 2020, yielding 52 data points.

[4] https://www.reddit.com/wiki/api-terms

[5] https://fred.stlouisfed.org

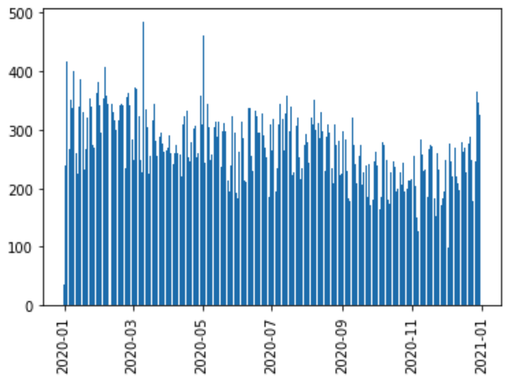


Figure 1: Post frequency on `/r/personalfinance` ranging from January 1, 2020, through December 31, 2020.

Word	Occurrences	Word	Occurrences
would	75065	know	32276
i'm	58644	year	32244
credit	55896	account	31759
get	51274	want	30928
pay	49026	years	28186
money	43567	make	28047
like	41422	loan	27515

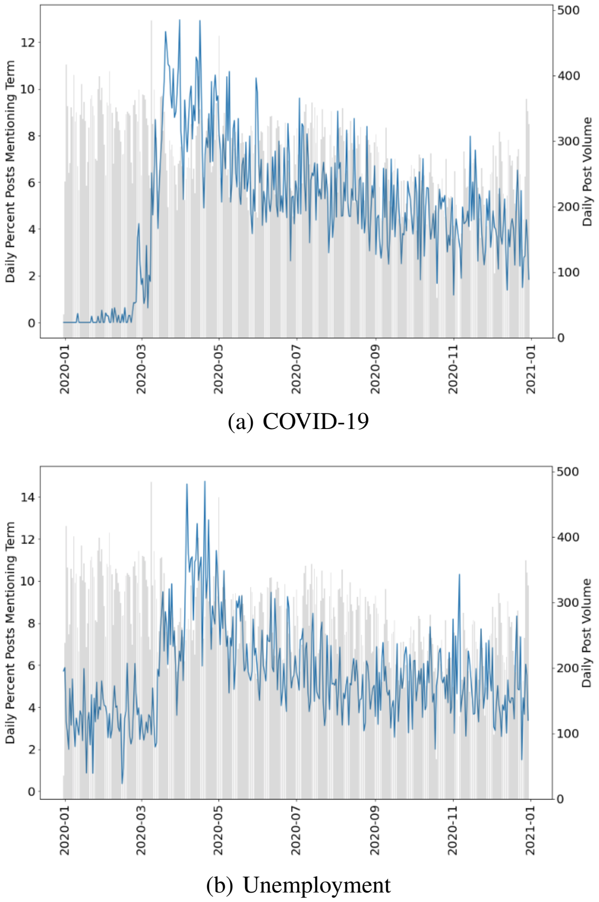
Table 1: The most frequent words appearing in `/r/personalfinance` posts during the year 2020, after removing stopwords.

3.4 Exploratory Data Analysis

We conducted exploratory, descriptive analyses of our data and outline a few findings in this subsection. Figure 1 shows post frequency over time on `/r/personalfinance` during our data collection range. We note the somewhat downward trajectory of this data, with a notable lower volume of posts around November.

Considering post title and text, we observe word count distributions (min, max, mean) for titles as (1, 41, 6.24) and texts as (0, 2897, 83.92), after removing stopwords using the NLTK stopwords list [Bird *et al.*, 2009]. Also excluding stopwords, we report the most frequent post text words in Table 1. We note that some common words among those that are most frequent are indicative of the underlying question/answer style of the posts, like "i'm," "like," "know," "want," etc.

Finally, we implemented a series of functions to compute the post frequency with title or text matching a set of query words. Figure 2 shows two queries — (a) for a set of COVID-19 related keywords {*covid*, *covid19*, *coronavirus*, *virus*, *corona*, *covid-19*, *2019-ncov*, *2019ncov*, *sars-cov-2*, *sarscov2*}, and (b) for a set of unemployment related keywords {*unemployed*, *laid-off*, *layoff*, *layoffs*, *unemployment*, *benefits*}. In both plots, the blue line denotes the daily percent of posts matching one of these query terms, and the background grey bars reiterate the post volume and frequency.



(a) COVID-19

(b) Unemployment

Figure 2: In blue, the line tracks the daily percentage of posts matching queries corresponding to (a) COVID-19 or (b) unemployment during the year 2020. In grey, the bars indicate the overall post volume and frequency over the same time period.

It is interesting to visually note the steadily decreasing frequency of COVID-19 words after the initial spike — and overall relative infrequency of these keywords with a maximum of around 12% of posts — despite what we know to be the actual prevalence of COVID-19 cases in the United States,[6] shown in Figure 3.

4 Methods & Models

For the task of correlating our `/r/personalfinance` posts with initial and continuing unemployment claims, we experiment with different iterations of a CNN deep learning model. Our model structure is influenced heavily by Semeval 2017's Task 5 [Cortis *et al.*, 2017], which focused on sentiment scoring of financial headlines. More specifically, we draw inspiration from the works of Mansar *et al.* (2017) and Kar *et al.* (2017), which experimented with various deep learning models and word embeddings to achieve the top two scoring results.

The primary advantage of using a simple CNN for our prototype model is that these models can work effectively on

[6] https://covid.cdc.gov/covid-data-tracker/

Row	Model	Macro F1	Weighted F1	Details
TB1	Random Forest	0.710 (0.323)	0.772 (0.443)	GloVe 300, Stopword Removal, Regex
TB2	Logistic Regression	0.772 (0.443)	0.796 (0.481)	GloVe 300, Stopword Removal, Regex
TB3	Linear SVM	0.792 (0.509)	0.814 (0.545)	GloVe 300, Stopword Removal, Regex
CNN1	CNN	0.807	0.823	GloVe 300
CNN2	**CNN**	**0.837 (0.544)**	**0.857 (0.598)**	**GloVe 300, Stopword Removal, Regex**
CNN3	CNN	0.821	0.844	GloVe 300, Stopword Removal, Regex, VADER

Table 2: Model Training Results: *Two-Class F1 (Four-Class F1)*. *Two-class F1* refers to the F1 results achieved under a simplified two-class labeling scheme that merges U+C and E+O into two broad categories, whereas *four-class F1* (in parentheses) reports F1 results under the original labeling scheme that considers U, C, E, and O all as separate classes.

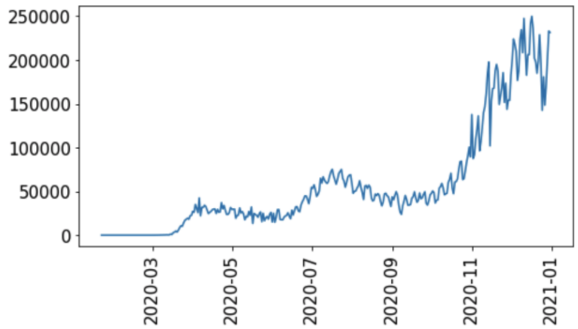


Figure 3: Daily COVID-19 cases in the United States during the year 2020, reported by the CDC.

small datasets, having fewer parameters to train than Recurrent Neural Network-based models. This is important since our access to supervised training data is limited to the small-scale corpus developed specifically for this work. A worthwhile side benefit of using more efficient models is that training time is cut down significantly, which makes a noteworthy difference in our training setting — a 2011 Mac Mini with 16gb of memory and a dual-core i5 processor. Our model contains the following layer structure: (1) Convolution 1D, (2) Max Pool, (3) Dense F.C., (4) Dropout, (5) Dense F.C., (6) Dropout, (7) Softmax Output.

All models were trained and evaluated using randomly assigned 90/10 train/test splits. We compared six conditions, three of which were classical machine learning models (considered as baseline alternatives) and three of which were variations of our CNN architecture with different training or preprocessing settings:

- **TB1:** A random forest classification model [Ho, 1995].
- **TB2:** A logistic regression classification model [McCullagh and Nelder, 1989].
- **TB3:** A linear support vector machine (SVM) classification model [Boser *et al.*, 1992].
- **CNN1:** Our CNN model with no extra preprocessing steps.
- **CNN2:** Our CNN model with stopword removal, as well as regular expression-based removal of special characters.
- **CNN3:** Our CNN model with all of the preprocessing steps applied in CNN2. Additionally, post-level VADER [Hutto and Gilbert, 2014] sentiment scores were concatenated to the output of the max pooling layer.

These conditions are further described in Table 2, and additional feature details are provided in §4.2. Scikit-Learn [Pedregosa *et al.*, 2011] was the primary package for implementing the classical machine learning models (TB1, TB2, and TB3), and Keras [Chollet and others, 2015] was the primary package for implementing the CNN models (CNN1, CNN2, and CNN3). CNN models were trained using stochastic gradient descent and a learning rate of 0.0001, for 150 training epochs. For all models, text input consisted of the post title concatenated with the first 25 words of the post text, encoded using pretrained 300-dimensional GloVe embeddings trained on Wikipedia 2014 and Gigaword 5 [Pennington *et al.*, 2014].

4.1 Model Evaluation

Results of our model comparison are shown in Table 2. We evaluate models using macro and weighted F1 scores (harmonic mean of precision and recall), computed using Scikit-Learn. Non-parenthesized scores report model performance on a simplified two-class labeling scheme where we combined U+C and E+O posts into two distinct categories. The motivation behind this decision was an unfortunate and significant class imbalance — our dataset distribution was ultimately 19.8% U, 12.2% C, 50.1% E, and 17.1% O. While clearly not empty, there simply were not as many posts in our target categories of interest, U and C, as we expected. Four-class classification scores are also reported for our top performing model in row CNN2, as well as our baseline models, enclosed in parentheses. All four-class scores are notably lower than the two-class scores.

4.2 Ablation Studies

In our CNN conditions, we experimented with a variety of settings to assess the utility of preprocessing steps and lexicon-based features (i.e., sentiment scores), with these settings represented as CNN1, CNN2, and CNN3. To incorporate sentiment scores into the model, we concatenated post-wise VADER sentiment lexicon scores [Hutto and Gilbert, 2014] to the output of the max pooling layer, similar to that

Row	Item 1	Item 2	r
R1	Frequency of Tagged Employment Posts	Initial Claims	0.560
R2	Frequency of Tagged Employment Posts	Continued Claims	0.309
R3	**Keyword Unemployment Post %**	**Initial Claims**	**0.796**
R4	Keyword Unemployment Post %	Continued Claims	0.701
R5	CNN U/C Predicted Posts	Initial Claims	0.691
R6	**CNN U/C Predicted Posts**	**Continued Claims**	**0.747**

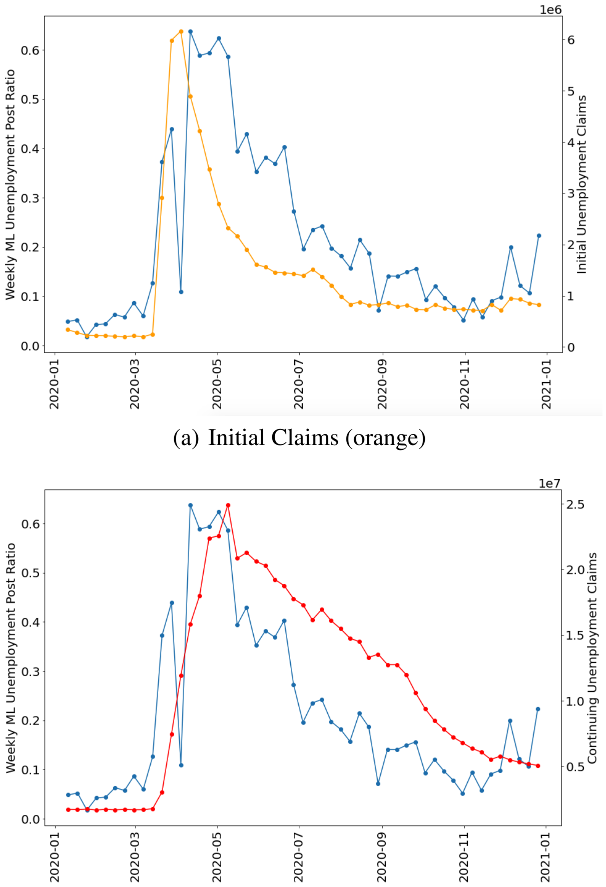
Table 3: Pearson correlation coefficients computed between model predictions on `/r/personalfinance` unemployment posts (Item 1) and official initial and continued unemployment claims (Item 2).

done by Mansar *et al.* (2017). Contrary to our expectations, this did not consistently improve model performance, yielding a weighted F1 score of 0.844 which is lower than the performance without VADER scores (CNN2).

However, it was clear from our comparison that the task benefited from some basic text preprocessing steps. In CNN2, stopwords were removed using NLTK [Bird *et al.*, 2009] and regular expressions were utilized to remove special characters such as commas and dashes, whereas in CNN1 no preprocessing steps were performed. CNN2 outperformed CNN1; moreover, these actions lowered the number of missing words encountered by our pretrained GloVe embedding model (13.56% originally, reduced to 1.50%).

5 Results & Discussion

Extending the predictions of our top performing model to the entire set of "employment" `/r/personalfinance` posts, we computed correlation with official initial and continuing unemployment claims to assess sub-reddit posting behaviors with respect to real-world financial trends. Pearson correlation coefficients (r) were computed using the Stats module of Scipy [Virtanen *et al.*, 2020] and are shown in Table 3. Item 1 in the table corresponds to the condition used to predict (un)employment, and Item 2 corresponds to the data with which it was correlated (either *Initial Claims* or *Continuing Claims*). Rows R1 & R2 show baseline results which consider the frequency or volume of posts self-tagged with "employment." Rows R3 & R4 show the correlation results when considering the unemployment keyword percentages from Figure 2(b). Rows R5 & R6 show the CNN predicted U/C class. All "Item 1" data is aggregated weekly to match up with the 52 weeks of "Item 2" data.



(a) Initial Claims (orange)

(b) Continuing Claims (red)

Figure 4: Relationship between CNN model predictions (blue) and official unemployment trends (orange or red, corresponding to initial or continuing claims, respectively).

We observe that the strongest correlation with initial claims stems from the simple keyword search (described in §3.4) at a value of 0.796. This does not necessarily indicate poor performance from our model, which achieved a lower score at 0.691. Rather, we speculate that the way in which categories were defined, mixed with the actual sub-reddit post content, simply does not as closely reflect initial unemployment metrics. On the other hand, our CNN model predictions of the Unemployment/Cut merged class correlate better with continuing unemployment claims than the alternative approaches, at a value of 0.747. The correlations between our CNN model predictions and real-world unemployment trends are illustrated through the plots in Figure 4.

Visually, it is clear that the elongated peak of the CNN predictions around April and May (shown in blue) match up better with continuing claims (red) than with initial claims (orange). Qualitatively, this behavior appears reasonable given the context of the online forum — in fact, with almost all posts being question/answer focused, we observed this same pattern while annotating posts. Individuals seemed relatively unlikely to post immediately after being laid-off with questions about filing for unemployment. Instead, for unemployment related inquiries we observed notably higher frequency of individuals asking questions about events within the last month or so. These posts frequently referenced mistakes made on unemployment filing forms, for example by asking how long they needed to wait before they received their first assistance payment after filing.

6 Summary, Conclusion & Future Work

In this work, we successfully obtained strong correlation with official unemployment statistics during the economically anomalous year of 2020. We contributed a new dataset of `r/personalfinance` posts labeled with *Unemployment*, *Cut/Furlough*, *Employment*, and *Other* designations. We then trained a number of machine learning models on this data, including both classical baselines (Random Forest, Logistic Regression, Linear SVM) as well as a more advanced CNN model, which achieved a best weighted F1 score of 0.857 on two-class classification of posts. These predictions were extended to the full slice of user "employment" tagged posts. We then computed correlation between these predictions and initial and continuing unemployment claims, sourced from the St. Louis Federal Reserve, relative to two baselines (frequency of posts tagged with "employment," and unemployment keyword percent search). Our experiments yielded top correlations of 0.796 for initial claims and 0.747 for continuing claims. We make our data and source code available to the research community to foster additional work in this area and to facilitate replication.

With respect to future work directions, on a micro level, there are clear next steps which could be taken to encourage improved performance. Our current models serve as a prototype, with the primary objective being to establish proof of concept rather than achieve state of the art performance. While our methodology for training the CNN model was efficient and F1 scores indicate that the model structure and inputs were enough to achieve strong performance measures, extending model input to include longer word sequences or further expanding our labeled dataset could similarly help to achieve higher performance. This could also enable the use of more advanced machine learning models like gated recurrent units (GRUs), long short-term memory networks (LSTMs), or BERT/Transformer-based models [Devlin *et al.*, 2019], which we largely avoided due to the small size of our labeled data sample.

On a macro level, this study provides an interesting case study towards harnessing natural language in Reddit posts for economic time series correlation, specifically focused on unemployment and the COVID-19 pandemic in the USA. However, further work could consider including a larger time frame to encompass non-pandemic phenomena as well. Even within the context of our specific sample of `/r/personalfinance` data, there is still ample room to explore and analyze altogether different time series, as there is no shortage of user-tagged posts covering a diverse range of topics including "retirement," "housing," "investing," "auto," and others, offering a productive starting point for follow-up work.

References

[Baumgartner *et al.*, 2020] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The Pushshift Reddit Dataset. *arXiv:2001.08435 [cs]*, January 2020. arXiv: 2001.08435.

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

[Boser *et al.*, 1992] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, page 144–152, New York, NY, USA, 1992. Association for Computing Machinery.

[Chollet and others, 2015] François Chollet et al. Keras. https://github.com/fchollet/keras, 2015.

[Choudhury and De, 2014] Munmun De Choudhury and Sushovan De. Mental Health Discourse on reddit: Self-Disclosure, Social Support, and Anonymity. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1), May 2014.

[Cortis *et al.*, 2017] Keith Cortis, André Freitas, Tobias Daudert, Manuela Huerlimann, Manel Zarrouk, Siegfried Handschuh, and Brian Davis. SemEval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 519–535, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*, May 2019. arXiv: 1810.04805.

[Gjurković and Šnajder, 2018] Matej Gjurković and Jan Šnajder. Reddit: A Gold Mine for Personality Prediction. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 87–97, New Orleans, Louisiana, USA, June 2018. Association for Computational Linguistics.

[Ho, 1995] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, pages 278–282 vol.1, 1995.

[Hutto and Gilbert, 2014] Clayton J. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Eytan Adar, Paul Resnick, Munmun De Choudhury, Bernie Hogan, and Alice H. Oh, editors, *ICWSM*. The AAAI Press, 2014.

[Kar *et al.*, 2017] Sudipta Kar, Suraj Maharjan, and Thamar Solorio. RiTUAL-UH at SemEval-2017 Task 5: Sentiment Analysis on Financial Data Using Neural Networks. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 877–882, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[Landis and Koch, 1977] J. Richard Landis and Gary G. Koch. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174, 1977.

[Low *et al.*, 2020] Daniel M. Low, Laurie Rumker, Tanya Talkar, John Torous, Guillermo Cecchi, and Satrajit S. Ghosh. Natural Language Processing Reveals Vulnerable Mental Health Support Groups and Heightened Health Anxiety on Reddit During COVID-19: Observational Study. *Journal of Medical Internet Research*, 22(10):e22635, October 2020.

[Mansar *et al.*, 2017] Youness Mansar, Lorenzo Gatti, Sira Ferradans, Marco Guerini, and Jacopo Staiano. Fortia-FBK at SemEval-2017 Task 5: Bullish or Bearish? Inferring Sentiment towards Brands from Financial News Headlines. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 817–822, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[McCullagh and Nelder, 1989] Peter McCullagh and John A. Nelder. *Generalized Linear Models, Second Edition*. Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series. Chapman & Hall, 1989.

[Murray *et al.*, 2020] Curtis Murray, Lewis Mitchell, Jonathan Tuke, and Mark Mackay. Symptom extraction from the narratives of personal experiences with COVID-19 on Reddit. *arXiv:2005.10454 [cs, stat]*, May 2020. arXiv: 2005.10454.

[Pedregosa *et al.*, 2011] 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. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

[Shen and Rudzicz, 2017] Judy Hanwen Shen and Frank Rudzicz. Detecting Anxiety through Reddit. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, pages 58–65, Vancouver, BC, August 2017. Association for Computational Linguistics.

[Valizadeh *et al.*, 2021] Mina Valizadeh, Pardis Ranjbar-Noiey, Cornelia Caragea, and Natalie Parde. Identifying medical self-disclosure in online communities. In *Proceedings of the 2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2021)*, Online, 2021. Association for Computational Linguistics.

[Virtanen *et al.*, 2020] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[Wooley *et al.*, 2019] Stephen Wooley, Andrew Edmonds, Arunkumar Bagavathi, and Siddharth Krishnan. Extracting cryptocurrency price movements from the reddit network sentiment. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 500–505, 2019.