23rd International Symposium INFOTEH-JAHORINA, 20–22 March 2024

Common N-Gram Method (CNG): A Promising Approach to Detecting Mental Health Disorders on Social Media

Harshit Agarwal and Vlado Kešelj
Faculty of Computer Science, Dalhousie University
6050 University Ave, NS, Canada
9arshit@gmail.com vlado@cs.dal.ca

***Abstract*—Technology and social media's growing prevalence, especially during the COVID-19 pandemic, has contributed to a rise in mental health issues. This research discusses novel applications of natural language processing which can help develop more effective and accessible diagnostic tools for mental health illnesses. To enhance the realism of our model, we created a biased dataset that reflects the real-world ratios of mental illness prevalence. The proposed solution is the Common N-grams (CNG) method that offers comparable results to the state-of-the-art CNN-LSTM model and is less resource-intensive. The CNG method performs better than the CNN-LSTM model and Support Vector Machine (SVM), baseline model, in multi-classification tasks. The CNN-LSTM surpasses performance in binary tasks compared to the best score reported in the previous study with the same dataset. The study also highlights the usefulness of the Relative N-Gram Signature method to analyze the classification decision of the CNG technique. The proposed solutions offer practical and accessible options for individuals seeking reliable and accurate mental health support.**

Keywords*—Natural Language Processing, Mental Health, Common N-grams Method, CNG, Word Embeddings, Neural Networks, Deep Learning, Machine Learning, Social Media Analytics

I. INTRODUCTION

In the field of psychology, prior research work has been conducted to study the relationship between social media and mental health and how they affect each. Pantic [1] specifically investigated the relationship between Facebook usage and symptoms of depression and found that certain online behaviors may serve as predictive indicators in identifying and assessing depression.

Mental illnesses can significantly impact a person's writing and speaking patterns, and gaining an in-depth understanding of the characteristics and effects of these disorders is crucial for improving early identification, management, and treatment options [2], [3]. Therefore, this study aims to provide valuable insights into these mental illnesses and contribute towards achieving better mental health outcomes.

This research discusses novel applications of natural language processing and machine learning techniques which can help develop more effective, accessible, and integrable diagnostic tools for mental health illnesses using text-based data. The objective is to enrich the well-being of people who are dealing with mental health problems and to give people the tools they need to better understand and manage their mental health. The proposed solutions offer comparable results to state-of-the-art systems and are less resource-intensive, making them a practical and accessible option for individuals seeking reliable and accurate mental health support. By implementing these solutions, we hope to contribute to a future where mental health resources are widely available and mental health conditions can be detected and treated with greater efficiency and accuracy.

The main contributions of this research are:
1) Novel implementation of CNG for multiclassification task in the field of health care and comparing its performance to deep learning methods.
2) Preparing and using a biased dataset in the real-world ratios of mental illness to simulate true prevalence, ensuring results that truly represent real-life scenarios. This approach enhances the model's evaluation, accurately reflecting its performance across diverse classes in the dataset.
3) Applying relative n-gram signature method for analysis which improves the interpretability of the model by providing a more meaningful representation of the data, leading to improved understanding and potential breakthroughs.

II. RELATED WORK

The accessibility of widespread language on social media has intrigued researchers studying the linguistic expressions of individuals with mental health conditions [4], [5]. Many studies have turned to Facebook, Twitter, and Reddit for data collection due to the cost and bias issues present in survey-based studies. Twitter's (currently X) character limit for messages can pose a challenge for deep learning techniques that require substantial amounts of data, as the concise nature of the messages may not allow for adequate information to be captured [6]. Various studies have been conducted to use social media data to identify and analyze specific mental illnesses such as depression [7], [5], [8], schizophrenia [9], and postpartum emotions in mothers [10], suicide ideation [11], [12], [4], [13], [14]. Cohan et al. [15] expanded upon the work of Yates et al. [5] and created a more extensive dataset that incorporated nine classes of illnesses and a targeted control group from Reddit.

Similar research has been done in the field to detect mental illness using text and by studying online social media behavior using machine learning and deep learning methods. Jina et al. [16] implemented a deep learning model using XGBoost and CNN-based classification on Reddit posts to accurately identify mental disorders like depression, anxiety, bipolar, borderline personality disorder, schizophrenia, and autism.

Ivan et al. [17] have analyzed how a Hierarchical Attention Network (HAN) can be better at analyzing social media posts under certain conditions than traditional models. The study found that the performance of ML models is highly dependent on the availability of data, with results worsening as the amount of available data decreases.

Relative n-gram signatures detect document properties, they can provide a visual explanation of a CNG classifier's reasoning [18]. This method was used to explain and improve CNG classifiers decisions in the authorship attribution domain

979-8-3503-2994-0/24/$31.00 ©2024 IEEE

and we adopt it in our research problem to better explain the significance of character n-gram signatures in detecting different disorders. The n-gram signature method provided some understanding of the model's reasoning, although it struggled with ADHD classification, which was also noted in the original research.

Cohan et al. [15] conducted an in-depth study on binary and multi-label classification of various illness classes, along with a control group, using models such as logistic regression, Support Vector Machine, CNN, and supervised FastText. Their best f1-score of 0.54 for binary classification and 0.27 for multi-label classification serves as a benchmark for similar research in this domain. On the other hand, Kim et al. [19] focused on binary classification tasks, and their CNN-based model produced an impressive average f1-score of 0.85, which significantly outperformed XGBoost's average score of 0.50. It is worth noting that, unlike Cohan's dataset, Kim's data was extracted from subreddits that were dedicated to each illness thus text included an explicit declaration by the user about their condition and their experience, while Cohan's dataset excluded such text which makes it more applicable for creating models that can be deployed in real work. Other positive attempts at using Reddit posts to detect depression and anxiety came to the same conclusion that NLP models which are based on N-gram models as well as deep learning systems can be a practical solution for the detection of mental health disorder and early prevention [20], [21].

III. Methods

Dataset Description: This study uses Self-Reported Mental Health Diagnoses dataset whose collection approach is based on RSDD dataset [5] and further builds on it. The SMHD dataset comprises posts made on Reddit by users (referred to as "diagnosed users") who acknowledge having been given a diagnosis for one or more of nine mental health problems, as well as posts made by matched control users. The data of diagnosed users had every post made to a mental health-related subreddit or having a keyword associated with a mental health issue removed; as a result of the selection process, the data of control users do not have such postings [15]. The entire SMHD dataset consists of 116 million posts from 335,000 users [15] but this study uses a smaller sample dataset due to limited computation and memory resources.

We focus on six mental illnesses included in DSM-5 standards [22]. For the sampled dataset used in this study, patients with and without a different diagnosis of mental illness were balanced at 5% for depression, 3.8% for anxiety, 2.8% for ADHD, 1% for autism, 0.6% for bipolar fisorder, 0.45% for schizophrenia and rest 86.35% for the control group to simulate true mental illness prevalence [23], [24], [25], [26]. Table I gives more details about the dataset.

TABLE I: Label-wise Data Statistics

Label	# of Users	# of posts	Total Word Count	Total # of char	Total # of Non Alpha char
Control	8636	1,754,943	64,720,509	357,904,732	17,146,263
Depression	500	253,894	4,437,815	22,847,497	677,775
Anxiety	380	39,850	2,361,869	12,752,601	530,154
ADHD	280	31,695	1,827,233	10,056,686	403,166
Autism	100	12,184	767,305	4,209,406	181,642
Bipolar	60	6,138	390,626	2,108,075	82,527
Schizophrenia	45	5,361	314,569	1,749,242	79,053
Total	**10,000**	**2,104,065**	**74,819,926**	**411,628,239**	**19,100,580**

Dataset Pre-Processing: The raw data is cleaned by removing data points with missing text or labels and by removing corrupt data. We assume that letter case does not provide important information, so all letters are converted to lowercase to reduce data sparsity. Furthermore, in order to ensure that the dataset contains sufficiently informative posts for analysis, any posts with fewer than 50 words were dropped. In the data transformation phase, the data is tokenized and padded.

For the supervised learning methods, the tokens are vectorized by turning each text into a sequence of numbers following the standard term frequency inverse document frequency (TF-IDF) method.

Padding transforms sequences to be equal to the desired length by adding 0 before or after the sequence. If the sequences are longer then they are truncated so that they fit the desired length [27]. Many studies that implement deep learning models for classification tasks remove stopwords [28], [29] from the dataset. Still, for this research, this step was not done as stopwords and punctuation usage and frequency can be indicators of mental health illness [30]. After the preprocessing was completed, transformed data was passed to classification models.

Supervised Learning Models: The two supervised learning algorithms used in this study are SVM and CNN-LSTM models. Process pipelines for both models have similar phases. For the SVM model, no embedding layer is present and a linear kernel is chosen as a nonlinear kernel such as Radial Basis Function does not scale well with the big dataset. CNN layers for feature extraction on input data are paired with LSTMs to facilitate sequence prediction in the CNN-LSTM architecture.

For the CNN-LSTM model, the data is passed to an embedding layer before it is passed on to the first convolutional layer which is followed by the max pooling layer. The embedding layer helps us convert words into vectors of fixed length which can help the model better understand words and reduce dimensionality. After the max pooling layer data is passed to the LSTM layer which is connected to a fully connected layer and has softmax as the output activation function. The model also consists of a dropout layer [27] which deactivates random neurons during training to avoid overfitting.

Common N-grams Method (CNG): An n-gram is a contiguous sequence of N items from a given text. N-grams can be letter or word-based. N defines the length of the sequence. Unigram will contain one item, bigram will have 2 items, and so on. Character n-grams are language-independent and hence character-level n-gram language models can be easily applied to any language and even non-language sequences such as DNA and music.

We include all symbols in character n-grams, including letters, digits, punctuation, and whitespace symbols. The scope of this study is limited to character n-gram only. Fig. 1 illustrates the extraction of character n-gram via sliding window where the value of N is four.

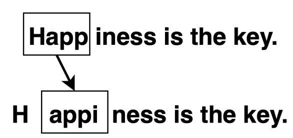


Fig. 1: Character N-gram Extraction Illustration

This study uses the CNG algorithm as proposed by Keselj et al. [31], where the class profiles consists of L most frequent n-grams in training data with their relative frequencies, and the profile distance is computed as the sum

$$\sum_{x}\left(\frac{f_1(x) - f_2(x)}{(f_1(x) + f_2(x))/2}\right)^2$$

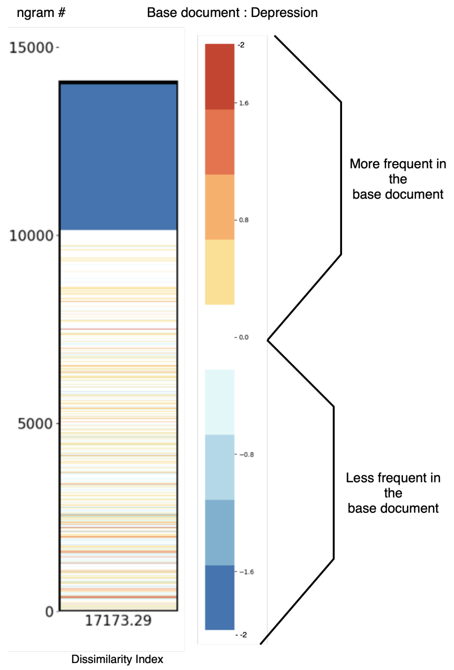


Fig. 2: Relative n-gram signature of Schizophrenia profile with Bipolar as a base document with parameters $n = 4$ and $L = 10000$

over all n-grams x in two profiles, where $f_1(x)$ and $f_2(x)$ are frequencies of x in two profiles respectively, or they are taken as 0 in a profile if x is not present in that profile. The CNG distance (i.e., dissimilarity) thus balances out effect of frequency differences of more frequent and less frequent n-grams, as explained in the original study [31]. Algorithm 1 outlines the computing method for the CNG dissimilarity. where $f_1(n)$ and $f_2(n)$ are frequencies of an n-gram n in the illness and the document profile.

Algorithm 1 Profile Dissimilarity (p_1, p_2)

$sum \leftarrow 0$
for all n-grams x contained in p_1 and p_2 **do**
  let f_1 and f_2 be frequencies of x in p_1 and p_2 (zero if they are not included)
  add square of the normalized difference of f_1 and f_1 to sum: $sum \leftarrow sum + (2 \cdot (f_1 - f_2)/(f_1 + f_2))^2$;
Return sum

Relative N-gram Signature: Jankowska et al. [32] proposed a visualization system called Relative N-gram Signature analysis methods the based on the Common N-grams method for text classification. The system was designed to conduct a detailed investigation of the characteristics of n-grams in the documents and provide greater insights into the working of the classifier in an intuitive manner.

A relative n-gram signature is built based on two n-gram profiles that reflect the usage frequency of n-grams between them. Suppose P_1 and P_2 are two n-gram size L profiles to be analyzed. The signature is created in the following way: First, all the n-grams in P_1 (which is the base document that serves as the background for the signature) are ordered based on their frequency followed by the n-grams that appear only in P_2 in a similar manner preserving their order in P_2. The n-gram at i-th index will be the i-th most common n-gram in the base document when $i \leq L$. If $i > L$, it means the i-th n-gram is i-th most common n-gram in the second document since the n-gram with a number greater than L doesn't appear in the base document. The lower the value of i the more common the n-gram is the class. The dissimilarity index is calculated using formula present in Algo. 1 for n-grams used to create the signature.

The system visually represents the difference in n-gram frequency usage using dual contrast color mapping. The white color indicates zero or near-zero distance and equal frequency of the n-gram in both documents. Red indicates a higher frequency in the base document, while blue indicates a higher frequency in the other document. The darker the strip, the greater the distance between the profiles for that specific n-gram.

IV. Results

The training time for the CNG model was notably efficient, taking only about half the time compared to the CNN-LSTM model. Additionally, memory consumption for the CNG model varied, but it was generally around 20% less, depending on the specific parameters chosen for the CNG model. We employed the f1-score to compare the models' overall performance. We evaluated these models on test data, which was not used while training the models. Table III shows the performance of the three models tested in this experiment and Supervised FastText which was the best-performing model in the study [15] which created the dataset used in the research.

A grid search for finding optimal values of n and L was done for the CNG model, $3 \leq n \leq 5$ and $L \in \{10000, 20000, 50000, 100000\}$. The reason for these limited values was, first as observed from past studies and experiments done with sample dataset, the highest performance was seen when $3 \leq n \leq 6$. Second, for $n \geq 7$ computation is slow and the memory requirement is high.

Even though our best f1-score (0.23) is for parameter $n = 5$ and $L = 100,000$ the model has a zero prediction rate for autism, bipolar control, and schizophrenia and overpredicts anxiety and control.

	N-gram size		
Profile Size	**3**	**4**	**5**
10,000	0.18	0.19	**0.20**
20,000	0.13	0.19	**0.20**
50,000	0	**0.20**	**0.20**
100,000	0	0.16	0.23

TABLE II: Pilot study F1-score for the multi-classification task using CNG

F1-score of 0.20 was obtained for four different pairs of n and L. We observed that for parameters $n = 5$ and $L = 10,000$, the model predicts 6 out of 7 labels with decent accuracy. While in other instances model has a similar f1-score for depression and control but a very low score for other classes. Due to this reason, CNG model's optimal parameters were selected as $n = 5$ and $L = 10,000$.

Binary classification task is performed between each illness and control group. For the binary classification task except for the depression class, CNN-LSTM outperforms SVM and CNG models. For the depression class, SVM has the highest f1-score of 0.89 and CNN-LSTM has 0.83. CNN-LSTM provides consistent results across all labels for binary labels. While CNG's performance in the binary classification task may not be ideal, it still exhibits a noteworthy precision score. Although the f1-score falls below 0.5 for all classes, this does not detract

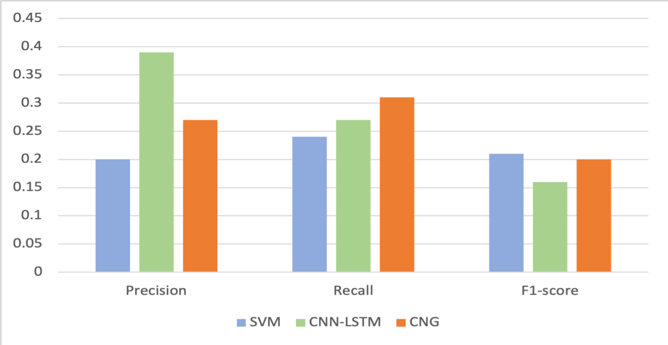


Fig. 3: Model Comparison

from the fact that the model is able to correctly identify a significant number of positive samples.

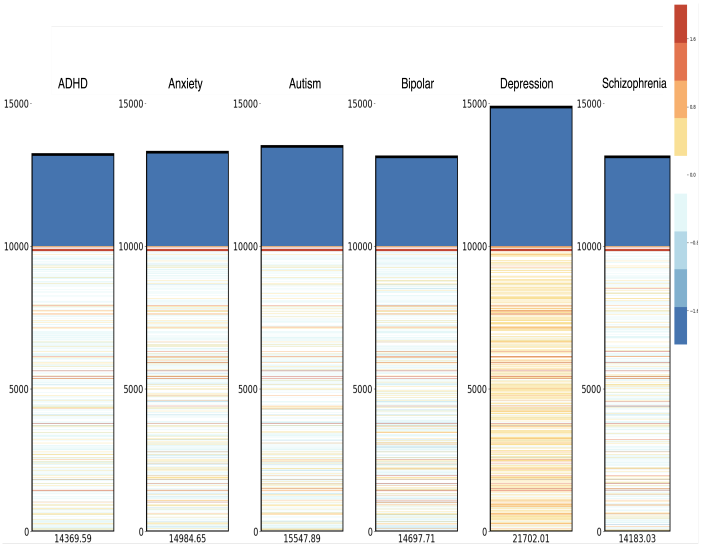
For multiclassification tasks, however, CNG method (0.20) outperforms CNN-LSTM (0.16). CNN-LSTM model is unable to predict any other class other than anxiety and depression. Even though SVM, which is acting as our baseline model, has a similar f1-score as CNG method it suffers from the same problem as CNN-LSTM does and is only able to predict depression and control class. This is most likely due to SVM overfitting on two majority classes and being unable to find boundaries for the minority classes which is evident from higher precision and recall rate but lower f1-score.

In a previous study [15], ML models that outperformed our method in terms of f1-score also showed inconsistency between precision and recall measures. Models would often report high precision and low recall rate or vice versa which is evidence that their models had an inconsistent performance where models would either return many results which are incorrect or very few results but they are correct.

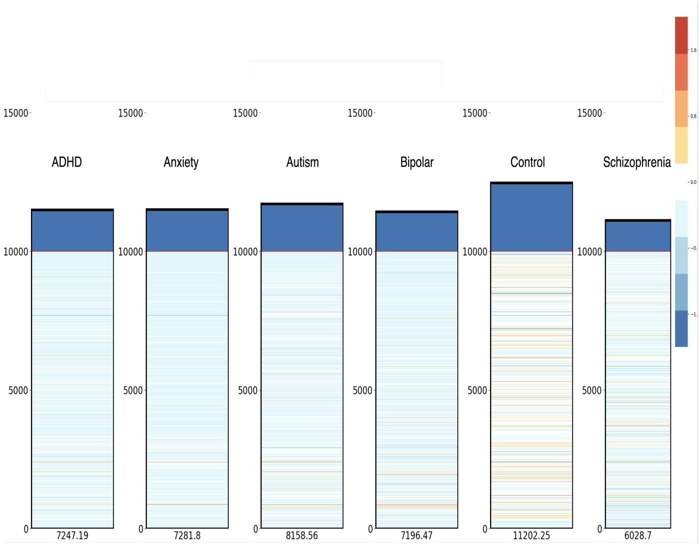
Model Interpretation: One advantage of the CNG classifier is that using the Relating N-Gram signature method, we can visualize class profiles and provide explanation for classifier decisions. Fig. 4 shows the relative n-gram signature for control and depression as a background profiles for parameters $n = 5$ and $L = 10,000$ for which CNG achieved the best results. It was observed that "depression" was the label with the highest accuracy, which can be attributed to the largest number of distinct n-grams in its profile (blue top) and more prominent frequency differences in common n-grams (more colorful lower common portion) Fig. 4a. "Schizophrenia" has the lowest classification rate and is misclassified as control, bipolar, and anxiety labels which can be attributed to the fact that schizophrenia has the lowest dissimilarity index of all when these labels act as base documents. The n-gram signature method provided some understanding of the model's reasoning, although it struggled with ADHD classification, which was also noted in the original research.

When observing the top 20 n-grams using the signature method it was observed that the top n-grams for anxiety, autism, bipolar, and schizophrenia are similar to one another and have almost the same frequencies. This is not an uncommon thing observed since stopwords were not removed which are usually in abundance in text written in the English language. An interesting phenomenon when using depression as base class is that the top 20 n-grams are more common in depression profiles than in any other label.

N-gram "_YOU_" (the symbol '_' represents a space character), capturing the second-person pronoun, is more prominent in the depression profile than in the control group. It has



(a) Relative N-gram Signature with Control as Base Document



(b) Relative N-gram Signature with Depression as Base Document

Fig. 4: Relative N-gram Signatures for n = 5 and L = 10,000

been noted in the literature that the first and second-person pronouns are more commonly used by people with mental health illnesses. Phrases such as "act of kindness" and "impact of betrayal" generate n-grams such as "ACT_O", "IMPAC" and "KIND_" which are in the high-frequency depression profile suggesting that people suffering from depression were describing situations with extreme emotions. The n-gram "ITING" was observed in sentences conveying emotions of excitement or worry, as words such as "waiting", "visiting", and "limiting" were present. These finds are consistent were previous work where similar emotions were observed from other data sources such as Facebook [33].

Past tense usage was more prevalent in profiles of individuals with mental disorders, as reflected in the n-gram "OULD" containing words like "could," "would," and "should." N-grams "YYYYY" and "OOOOO" are present in ADHD profiles with high frequency and mostly absent in other profiles. These particular n-grams were extracted from words like "heyyyyy" and "sooooo" and are thought to be linked to the repetitive behavior that is often observed in people with

TABLE III: Models Classification Performance

Model	Label						
	Depression	**Anxiety**	**ADHD**	**Autism**	**Bipolar**	**Schizophrenia**	**Multi-class**
Support Vector Machine	**P = 0.91**	P = 0.60	P = 0.59	P = 0.60	P = 0.62	P = 0.61	P = 0.20
	R = 0.89	R = 0.60	R = 0.59	R = 0.60	R = 0.62	R = 0.60	R = 0.24
	F = 0.89	F = 0.60	F = 0.59	F = 0.60	F = 0.61	F = 0.60	F = 0.21
Common N-Gram Method	P = 0.61	P = 0.62	P = 0.61	P = 0.59	P = 0.65	P = 0.58	**P = 0.27**
	R = 0.55	R = 0.55	R = 0.55	R = 0.54	R = 0.54	R = 0.53	**R = 0.31**
	F = 0.47	F = 0.48	F = 0.47	F = 0.46	F = 0.42	F = 0.45	**F = 0.20**
CNN Long Short-Term Memory Network	P = 0.82	**P = 0.65**	**P = 0.64**	**P = 0.65**	**P = 0.65**	**P = 0.66**	P = 0.29
	R = 0.85	**R = 0.65**	**R = 0.64**	**R = 0.65**	**R = 0.64**	**R = 0.66**	R = 0.27
	F = 0.83	**F = 0.65**	**F = 0.64**	**F = 0.65**	**F = 0.64**	**F = 0.66**	F = 0.16
Supervised FastText	P = 0.66	P = 0.67	P = 0.62	P = 0.68	P = 0.62	P = 0.69	P = 0.23
	R = 0.44	R = 0.44	R = 0.37	R = 0.39	R = 0.42	R = 0.33	R = 0.44
	F = 0.53	F = 0.53	F = 0.46	F = 0.49	F = 0.50	F = 0.45	F = 0.27

ADHD.

To interpret the CNN-LSTM model this research analyzed the embedding layer of the model. The embeddings were trained from scratch instead of using a pre-trained embedding model. To visualize embedding in 3D space, dimensionality reduction was done with help of t-SNE algorithm [34]. As shown in Fig. 5a the deep learning model is also able to link words that are related to feelings of loneliness and sadness and other related words such as “useless”, “alone”, “social” and “trash”. The model is also able to build a relation between words such as “king” and “president” and words related to some fee structure such as “tax”, “charge”, “fee” and “line”. These words are part of the sentences there described bills or tax payments as the reason for distress and worry.

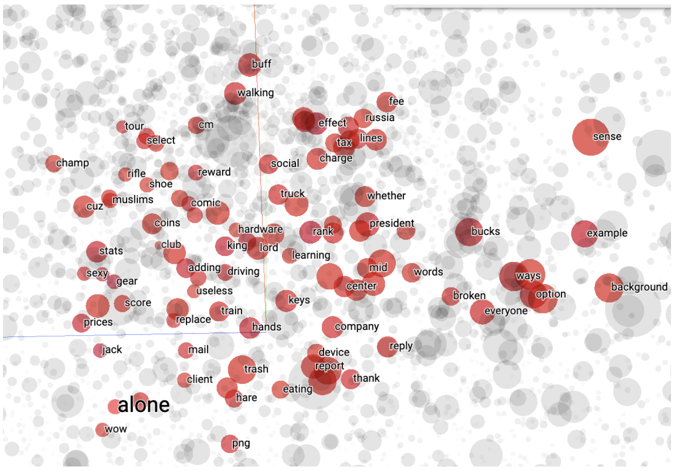
Word Embeddings extracted from CNN-LSTM model have some similarity with top n-grams that were observed from the CNG method. Words like “the”, “could”, “with”, “for”, and “this” are close neighbours in 3D space. CNN-LSTM model’s good recall rate for depression class can be explained by this phenomenon that the deep learning model was able to learn that these words are often present in text written by people with mental disorders.

Limitations: To note some limitation of presented results, we note that due to limited computational resources, a subset of the complete dataset is used, and a study on the full dataset should be conducted. Another limitation, shared with other studies on this dataset, is reliance on the subredit groups, where it is difficult to independently verify self-declared claims about mental disorders.

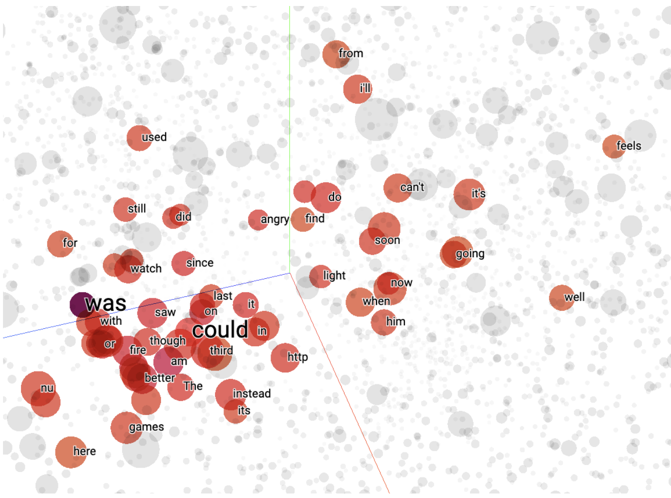
V. Conclusion and Future Work

To our knowledge, this is the first study to introduce the CNG method in classification of DSM-5 mental disorders, incorporating a control group for evaluation. Compared to state-of-the-art machine learning models, the CNG method demonstrated superior performance with an f1-score of 0.20, outperforming CNN-LSTM (0.16) in multi-classification tasks. Although the CNG model didn’t surpass CNN-LSTM in binary classification, it matched the score of the Supervised FastText method.

The Relative N-Gram Signature method was used to analyze the classification decisions of the CNG method and to provide interesting and useful insights. It was observed that classification decisions ar based on features that are consistent with findings in previous studies and that are used by professional psychologists to diagnose their patients. The CNG model has the potential to be easily integrated in real-time with other



(a) Neighbours words of ”ALONE”



(b) Neighbours words of ”COULD”

Fig. 5: Word Embeddings from CNN-LSTM model

health research methods for better treatment mental disorders, such as faster diagnosis and improvement evaluation.

In future work, CNG's performance can be further boosted by the creation of profiles tailored specifically to markers of an illness which can be facilitated using the n-gram visualization technique. Considering our work showed binary models are better at classifying mental disorders, using Ensemble methods to combine multiple binary classification models to create a multi-class and multi-label classification system should be an interesting avenue of research since illnesses like depression or anxiety are symptoms or secondary illness accompanied when suffering from ADHD or schizophrenia.

REFERENCES

[1] I. Pantic, "Online social networking and mental health," *Cyberpsychology, Behavior, and Social Networking*, vol. 17, no. 10, pp. 652–657, 2014.

[2] R. M. Ramos, P. G. F. Cheng, and S. M. Jonas, "Validation of an mhealth app for depression screening and monitoring (psychologist in a pocket): Correlational study and concurrence analysis," *JMIR MHealth UHealth*, vol. 7, no. 9, p. e12051, 2019.

[3] L. A. Clark, B. Cuthbert, R. Lewis-Fernández, W. E. Narrow, and G. M. Reed, "Three approaches to understanding and classifying mental disorder: ICD-11, DSM-5, and the national institute of mental health's research domain criteria (RDoC)," *Psychol. Sci. Public Interest*, vol. 18, no. 2, pp. 72–145, 2017.

[4] A. Cohan, S. Young, A. Yates, and N. Goharian, "Triaging content severity in online mental health forums," *Journal of the Association for Information Science and Technology*, vol. 68, no. 11, pp. 2675–2689, 2017. [Online]. Available: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.23865

[5] A. Yates, A. Cohan, and N. Goharian, "Depression and self-harm risk assessment in online forums," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2017, pp. 2958–2968. [Online]. Available: https://www.aclweb.org/anthology/D17-1321

[6] G. Coppersmith, M. Dredze, C. Harman, and K. Hollingshead, "From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses," in *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Denver, Colorado: Association for Computational Linguistics, Jun. 5 2015, pp. 1–10. [Online]. Available: https://aclanthology.org/W15-1201

[7] G. Coppersmith, M. Dredze, and C. Harman, "Quantifying mental health signals in Twitter," in *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Baltimore, Maryland, USA: Association for Computational Linguistics, Jun. 2014, pp. 51–60. [Online]. Available: https://aclanthology.org/W14-3207

[8] Z. Jamil, D. Inkpen, P. Buddhitha, and K. White, "Monitoring tweets for depression to detect at-risk users," in *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*. Vancouver, BC: Association for Computational Linguistics, 2017, pp. 32–40.

[9] M. Mitchell, K. Hollingshead, and G. Coppersmith, "Quantifying the language of schizophrenia in social media," in *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Denver, Colorado: Association for Computational Linguistics, Jun. 5 2015, pp. 11–20. [Online]. Available: https://aclanthology.org/W15-1202

[10] M. De Choudhury, S. Counts, and E. Horvitz, "Major life changes and behavioral markers in social media: Case of childbirth," in *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, ser. CSCW '13. New York, NY, USA: Association for Computing Machinery, 2013, p. 1431–1442. [Online]. Available: https://doi.org/10.1145/2441776.2441937

[11] B. Desmet and V. Hoste, "Online suicide prevention through optimised text classification," *Information Sciences*, vol. 439-440, pp. 61–78, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S002002551830094X

[12] R. Kshirsagar, R. Morris, and S. Bowman, "Detecting and explaining crisis," *arXiv preprint arXiv:1705.09585*, 2017.

[13] M. De Choudhury and E. Kıcıman, "The language of social support in social media and its effect on suicidal ideation risk," *Proc. Int. AAAI Conf. Weblogs Soc. Media*, vol. 2017, pp. 32–41, 2017.

[14] J. Jashinsky, S. H. Burton, C. L. Hanson, J. West, C. Giraud-Carrier, M. D. Barnes, and T. Argyle, "Tracking suicide risk factors through twitter in the US," *Crisis*, vol. 35, no. 1, pp. 51–59, 2014.

[15] A. Cohan, B. Desmet, A. Yates, L. Soldaini, S. MacAvaney, and N. Goharian, "SMHD: a large-scale resource for exploring online language usage for multiple mental health conditions," in *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 1485–1497. [Online]. Available: https://aclanthology.org/C18-1126

[16] K. Jina, L. Jieon, P. Eunil, and H. Jinyoung, "A deep learning model for detecting mental illness from user content on social media." *Scientific Reports*, vol. 10, no. 11846(2020), 2020. [Online]. Available: https://doi.org/10.1038/s41598-020-68764-y

[17] S. Ivan and S. Michael, "Adapting deep learning methods for mental health prediction on social media," in *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 322–327. [Online]. Available: https://aclanthology.org/D19-5542

[18] M. Jankowska, V. Kešelj, and E. Milios, "Relative n-gram signatures: Document visualization at the level of character n-grams," in *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2012, pp. 103–112.

[19] J. Kim, J. Lee, E. Park, and J. Han, "A deep learning model for detecting mental illness from user content on social media," *Scientific Reports*, vol. 10, no. 1, Jul. 2020. [Online]. Available: https://doi.org/10.1038/s41598-020-68764-y

[20] J. H. Shen and F. Rudzicz, "Detecting anxiety through reddit," in *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2017, pp. 58–65.

[21] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria, "MentalBERT: Publicly available pretrained language models for mental healthcare," *arXiv [cs.CL]*, 2021.

[22] A. P. Association, *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Association, May 2013. [Online]. Available: https://doi.org/10.1176/appi.books.9780890425596

[23] W. H. Organization, "Depression," https://www.who.int/news-room/fact-sheets/detail/depression, accessed: 2023-2-12.

[24] S. Dattani, H. Ritchie, and M. Roser, "Mental health," https://ourworldindata.org/mental-health, 2021.

[25] P. D. Sharon Saline, "ADHD statistics: New ADD facts and research," https://www.additudemag.com/statistics-of-adhd/, Oct. 2006, accessed: 2023-02-12.

[26] W. H. Organization, "Schizophrenia," http://www.who.int/news-room/fact-sheets/detail/schizophrenia, accessed: 2023-2-12.

[27] F. Chollet, "Keras," https://keras.io, 2015, accessed: 2023-02-12.

[28] E. Loper and S. Bird, "NLTK: The natural language toolkit," in *Proceedings of the ACL Interactive Poster and Demonstration Sessions*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 214–217. [Online]. Available: https://aclanthology.org/P04-3031

[29] S. Sarica and J. Luo, "Stopwords in technical language processing," *PLOS ONE*, vol. 16, no. 8, pp. 1–13, 08 2021. [Online]. Available: https://doi.org/10.1371/journal.pone.0254937

[30] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of liwc2015," The University of Texas at Austin, Tech. Rep., 2015.

[31] V. Kešelj, F. Peng, N. Cercone, and C. Thomas, "N-gram-based author profiles for authorship attribution," *Proceedings of the Conference Pacific Association for Computational Linguistics PACLING'03: 2003*, 09 2003.

[32] M. Jankowska, V. Kešelj, and E. Milios, "Relative n-gram signatures: Document visualization at the level of character n-grams," in *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2012, pp. 103–112.

[33] J. C. Eichstaedt, R. J. Smith, R. M. Merchant, L. H. Ungar, P. Crutchley, D. Preoţiuc-Pietro, D. A. Asch, and H. A. Schwartz, "Facebook language predicts depression in medical records," *Proceedings of the National Academy of Sciences*, vol. 115, no. 44, pp. 11 203–11 208, 2018. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.1802331115

[34] G. E. Hinton and S. Roweis, "Stochastic neighbor embedding," *Advances in neural information processing systems*, vol. 15, 2002.