Multimodal Analysis and Modality Fusion for Detection of Depression from Twitter Data

Nalin Semwal[1], Manan Suri[1], Divya Chaudhary[2], Ian Gorton[2], Bijendra Kumar[1]

[1]Department of Computer Engineering
Netaji Subhas University of Technology
New Delhi, India
[2]Khoury College of Computer Sciences
Northeastern University
Seattle, USA
semwalnalin@gmail.com, manansuri27@gmail.com, d.chaudhary@northeastern.edu, i.gorton@northeastern.edu,
bizender@gmail.com

Abstract

It is established that social media is not only a possible cause of mental health disorders, but a strong indicator. Great predictive information is contained in both text posts and images posted, which can be exploited by classification models. In this work, we evaluate and compare a number of different approaches to the detection of depression from social media activity. We use publicly available twitter posts and profile and background images as our predictive features. We test classical machine learning approaches, sequential models such as LSTMs, and Convolutional Neural Networks. Additionally, we implement and test a modality fusion model which fuses textual and image-based features to achieve greater accuracy. This fusion model outperforms the best textual and image models tested by a full 17.36 percentage points and 35.99 percentage points respectively, indicating that the information contained in text and images is complementary and is best exploited in conjunction.

Introduction

Not only has it been postulated that social media sites may be a source of depressive symptoms and low self-esteem (Pantic 2014), posts on popular social media sites such as Twitter and Reddit have also been acknowledged as a viable indicator for depression (Martínez-Castaño, Pichel, and Losada 2020; Coppersmith, Harman, and Dredze 2014). Classification of depression, sentiment analysis and detection of suicidal tendencies using Twitter posts are popular tasks (Stephen and P. 2019; Coppersmith et al. 2018; Safa, Bayat, and Moghtader 2022). These are often achieved by encoding text in the form of fixed-length vectors, and then applying classifiers such as logistic regression and random forests (Rajaraman and Ullman 2011). Apart from text posts, other forms of data, such as profile and background images, biographical data, etc. have also been used, albeit less commonly, for classification (Guntuku et al. 2019; Safa, Bayat, and Moghtader 2022; Wang et al. 2020; Chiu et al. 2021; Kumar and Garg 2019; Gallo et al. 2020). These multi-modal approaches have been found to provide reasonably accurate inference to support, if not supplant, text posts.

However, there are two important considerations to be studied in the context of analyzing mental health on social media: 1) The use of more sophisticated embeddings such as Word2Vec and sequential models such as LSTMs for the classification task (Le and Mikolov 2014; Mikolov et al. 2013; Hochreiter and Schmidhuber 1997), 2) The possibility of fusing multiple modalities, such as text posts and images, to draw joint inference.

In this work, we conduct four sets of experiments: 1) Applying classical Machine Learning (ML) techniques, such as logistic regression and decision trees to simple text posts for classification, 2) Applying learned embeddings and sequential models such as LSTMs and GRUs to simple text posts for classification, 3) Performing classification using different modalities, including publicly available profile and background images, and biographical data, 4) Performing modality fusion for classification by fusing text posts with profile and background images.

Data collected for the experiments includes over 10 million publicly available posts and approximately 10,000 images. Labeling for classification is done based on self-diagnosis of depression by the user.

Methodology

Dataset

Our data collection strategy was similar to that of Safa et al. (Safa, Bayat, and Moghtader 2022).

Diagnosis Group We first collected tweets containing some self-reported diagnosis of depression over a span of time from 01/01/2017 to 01/06/2022 using the regular expressions - "i have/was (just) (been) diagnosed with depression". We performed an initial filtering by removing all retweets and all duplicate tweets. The resulting set of 8754 tweets was used to construct the 'Diagnosis' user group, denoted by U_D. U_D was then refined by removing all users who has posted less than 100 tweets or had not posted a profile or background image. U_D finally contained 2970 users. U_D was used to obtain the Diagnosis tweet dataset T_D by scraping $max(T_i, 3000)$ tweets for each user $i \in U_D$, where T_i is the number of tweets by user i. After removing all

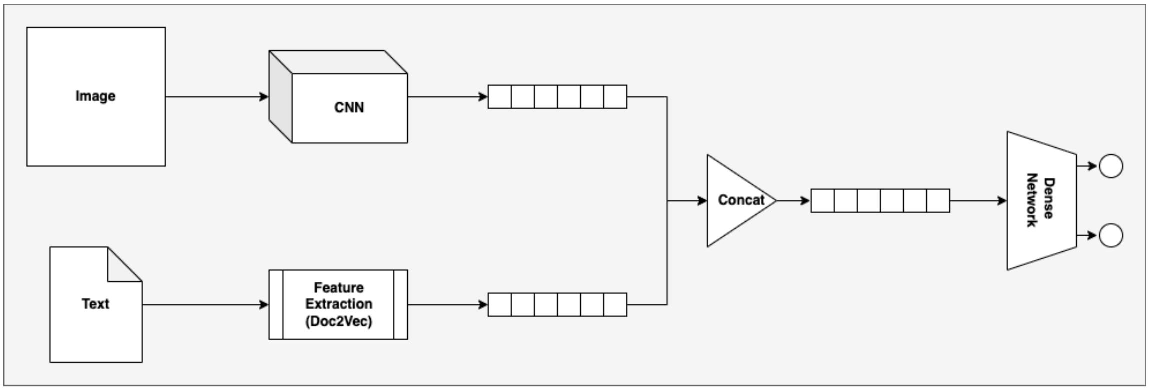

Figure 1: Modality Fusion Concept

retweets and duplicates from T_D, we ended up with a final dataset of 6.1 million Diagnosis tweets. We constructed the Diagnosis image datasets P_D and B_D, containing profile and background images respectively, from U_D. All images in P_D and B_D were resized to 512x512 pixels.

Control Group In order to construct the 'Control' group of users, denoted by U_C, we first collected tweets containing the word 'the' for a single day (01/06/2022). The same preprocessing as the Diagnosis group was applied, removing retweets and duplicates to get a set of 40,789 tweets. U_C was then constructed by filtering the users similarly to U_D. Additionally, all users overlapping with U_D were removed, resulting in a set of 2273 users. U_C is then used to construct T_C in the same way as T_D, giving us 4.6 million Control group tweets. The Control image datasets P_C and B_C were then constructed from U_C and processed similarly to P_D and B_D.

Feature Extraction

In both T_D and T_C, we first perform the standard preprocessing pipeline of Tokenization, Stopword Removal, Lemmatization (using WordNet) and Snowball Stemming. For our tasks, all tweets in T_D are given a label of 0 and those in T_C are given a label of 1.

TF-IDF Vectors: These reflect a word's importance in a document based on the word's frequency in the given document and the number of documents the word appears in. We construct Character 2-grams, Character 4-grams, Word 1-gram, Word 2-grams, Word 3-grams.

Word2Vec: A fixed size vector representation is learned corresponding to each word by optimising for a 'pseudo-task'. We use the Skip-gram method.

Doc2Vec: A fixed-length embedding of the complete document is obtained. This is done by adding a document vector feature to the Word2Vec algorithms.

Model Architectures

Classical Methods Since T_C and T_D combined contain $>$10 million tweets, we hold back only a small fraction (1%) of the data for validation, and use the rest (99%) for training.

We use TF-IDF features to train the following classical ML models: Logisitc Regressor, Ridge Classifier, Gradient Boosted Trees, Random Forest, Artifical Neural Network.

The ANN is designed with 2 hidden layers, with 256 neurons and 32 neurons respectively, and with 2 output neurons with softmax activation denoting class probabilities. Dropout is applied with a probability of 0.2.

We also train an identical ANN using the Doc2Vec features, with two hidden layers, dropout with a probability of 0.2 and softmax output activation, for 50 epochs.

Sequential Models Long Short Term Memory (LSTM) models are able to exploit the sequential and temporal information present in data. Additionally, they avoid the vanishing gradient problem faced by simple RNNs through the use of explicit gates. We use three Bidirectional LSTM Cells with 100 neurons each, followed by a 1D Convolution with 100 output channels and a Dense layer with 16 neurons. We use binary cross-entropy loss and the Adam optimizer (Kingma and Ba 2014). The model is trained with a batch size of 1024 and a learning rate of 0.001 for 100 epochs, but accuracy is observed to plateau after $\approx$ 20 epochs.

Image-based Model We train a CNN classifier using P_D, P_C, B_D, and B_C. For the backbone we use the EfficientNetV2 family of models pretrained on the ImageNet dataset. To this model we attach a classification head consisting of a single dense layer with a sigmoid activation. During the training process, the backbone model's weights are frozen and the classification head is fine-tuned on our dataset. We use binary cross-entropy loss and the Adam optimizer (Kingma and Ba 2014). The model is fine-tuned with a batch size of 24 and a learning rate of 0.001 for 20 epochs.

Modality Fusion Based on the work of Gallo et al. (Gallo et al. 2020), we train an early fusion model making use of both tweets and images. To achieve this, we concatenate the feature vector representation of each tweet with a feature vector corresponding to the author's profile and background images. A conceptual diagram of the fusion model is shown in Figure 1.

For textual tweets, we use the Doc2Vec model to obtain a 1024-unit vector corresponding to each tweet. Let this be represented as $\varphi_{text}(t_i)$ where $t_i \in (T_D \cup T_C)$.

For images, we use the same EfficientNetV2S architecture as described above, pooling and flattening the final output into a 1280-unit vector. Let these be represented as $\varphi_{image}(p_i)$ and $\varphi_{image}(b_i)$ where $p_i \in (P_D \cup P_C)$ and $b_i \in (\cup B_D \cup B_C)$.

For each tweet in the dataset, we obtain two final 2304-unit vectors - one by concatenating the tweet's feature vector with the feature vector of the author's profile image, and the second using the author's background image. These 2304-unit vectors serve as the input to our classification model. Let the final features be denoted by

$$\varphi(u_i^p) = Cat(\varphi_{text}(t_i), \varphi_{image}(p_i)) \quad (1)$$

$$\varphi(u_i^b) = Cat(\varphi_{text}(t_i), \varphi_{image}(b_i)) \quad (2)$$

For the classification model, we use an ANN with two hidden layers, containing 256 and 32 neurons respectively, followed by a two unit output layer with softmax activation. We train the model with binary cross-entropy loss with the Adam optimizer (Kingma and Ba 2014) and a learning rate or 0.001 for 50 epochs with a batch size of 32.

Results and Discussion

Results for all models and metrics tested are presented in Table 1.

Classical Methods Among the classical methods tested with purely textual data, the Artificial Neural Network with the Character 4-gram is found to achieve the highest accuracy at 81.94%, followed by the Artificial Neural Network with the Character 2-gram at 81.3%. The Gradient Boosted Tree with the Word 3-gram achieves the lowest accuracy at 52.38%. While the Artificial Neural Network achieves impressive accuracy overall, we postulate that this could be increased significantly by designing a more sophisticated and deeper architecture.

The ANN with Doc2Vec features, after training for 50 epochs, achieves an accuracy of 64.3%. We believe that this too may be improved by a more intricately designed model.

Sequential Models After training for 100 epochs, the LSTM based model with Word2Vec embeddings achieves an accuracy of $\approx 65\%$. It is interesting to note that this model achieves a slightly higher accuracy for the version without pre-processing, unlike the usual case with TF-IDF vectors. The pre-processed version achieves an accuracy of 63.01% after identical training.

Image-based Model Among the CNN models tested, the best accuracy is achieved with the EfficientNetV2M backbone after training for 20 epochs, at 63.31%. This number is comparable to and even greater than the accuracy achieved by some of the text-based models. As noted by (Safa, Bayat, and Moghtader 2022), images are seen to contain a surprisingly large amount of information about the user's mental health.

Modality Fusion With the Doc2Vec features and EfficientnetV2B0 CNN backbone, the fusion model tested is seen to achieve a surprisingly high accuracy of 99.3%, outperforming all other models by a wide margin. This number is a full 17.36 percentage points higher than the best text-based model among those tested, and 35.99 points higher than the best purely image-based model.

We observe that while text and image-based features both individually contain relevant information about mental health disorders, significantly higher predictive accuracy is achieved when they are used in combination. It can be deduced that the information contained in both modalities is complementary. Such fusion models are not frequently applied for classification problems; studying them may lead to much improved models in other areas as well.

Ethical Considerations

A person's mental health is a very private matter, and needs to be treated as such. For the purpose of this study, we have collected large-scale public information pertaining to people's mental health. While all of the information collected was made publicly available by the authors themselves, we nevertheless did not obtain consent for its use in such a study. With this in mind, we made sure to anonymize the data (replacing usernames and identifiers with random strings) before using it. We also made sure that the dataset was not made publicly available or circulated.

Another major concern is regarding the nature of the data itself. In order to collect a large dataset, we have collected tweets based on self-reported diagnosis of depression. As is the case with any self-diagnosis on social media, such reports are prone to falsification or simple exaggeration. It is quite likely that a number of these reports would not be clinically verified to be cases of depression. We have tried to allay this concern as much as possible by suitably filtering the set of users based on their other tweets, but the probability of misreporting is still non-trivial. Unfortunately, this is a necessary trade-off if we wish to acquire a large quantity of data without taking on the cost of manual labelling.

Conclusions and Future Work

In this work, we have studied and evaluated various models for detecting depression using publicly available tweets and profile and background images. We compared classical methods against sequential deep-learning based models, and examined the viability of purely image-based models as reliable detectors. Additionally, models based on modality fusion (i.e., the fusion of text-based and image-based features) were studied, and found to significantly outperform all preceding models.

Detection of mental health from social media posts is a challenging, not least due to the lack of reliable data. Most methods, including ours, are based on self-diagnosis; while these are found to perform reasonably well in practice, there are some concerns as to the legitimacy of such self-diagnosed reports on the internet.

Despite these issues, our work provides reliable models for practical applications that achieve great accuracy. Modality fusion, especially, is an area of research that could lead to great advances in applications of classification models.

Table 1: Collective Results For All Models and Metrics

Type	Model	Feature	Accuracy	F1 Score	AUC
Classical	Logistic Regression	Character 2-gram	60.42%	60.97%	64.61%
		Character 4-gram	62.9%	63.13%	68.24%
		Word 1-gram	62.4%	61.82%	67.53%
		Word 2-gram	58.22%	60.87%	62.15%
		Word 3-gram	54.81%	45.6%	58.22%
	Ridge Regression	Character 2-gram	60.37%	61.31%	64.43%
		Character 4-gram	62.87%	63.2%	68.2%
		Word 1-gram	62.98%	62.99%	68.54%
		Word 2-gram	57.82%	54.33%	61.92%
		Word 3-gram	54.79%	52.5%	58.2%
	Gradient Boosted Trees	Character 2-gram	65.06%	64.75%	67.18%
		Character 4-gram	68.44%	67.2%	66.46%
		Word 1-gram	70.46%	68.4%	72.83%
		Word 2-gram	61.8%	58.44%	60.89%
		Word 3-gram	52.38%	50.2%	51.6%
	Random Forest	Character 2-gram	65.18%	65.22%	67.14%
		Character 4-gram	71.14%	69.03%	72.4%
		Word 1-gram	69.81%	66.37%	70.14%
		Word 2-gram	63.2%	59.14%	60.46%
		Word 3-gram	58.73%	56.72%	61.8%
	Artificial Neural Network	Character 2-gram	81.3%	90.8%	82.14%
		Character 4-gram	81.94%	92.1%	83.43%
		Word 1-gram	79.91%	82.3%	80.42%
		Word 2-gram	68.4%	71.26%	70.61%
		Word 3-gram	62.65%	66.1%	65.76%
		Doc2Vec Vectors	64.3%	63.32%	70.47%
Sequential	LSTM	Word2Vec Vectors	64.93%	65.04%	71.04%
Image-based (CNN)	EfficientNetV2B0	Images	59.88%	62.12%	57.76%
	EfficientNetV2S	Images	62.1%	67.71%	62.81%
	EfficientNetV2B3	Images	62.5%	65.46%	58.58%
	EfficientNetV2M	Images	63.31%	68.5%	63.1%
Modality Fusion	CNN + ANN	Doc2Vec Vectors + EfficientNetV2B0 feature map (flattened)	99.3%	95.61%	93.2%

References

Chiu, C. Y.; Lane, H. Y.; Koh, J. L.; and Chen, A. L. P. 2021. Multimodal Depression Detection on Instagram Considering Time Interval of Posts. *J. Intell. Inf. Syst.*, 56(1): 25–47.

Coppersmith, G.; Harman, C.; and Dredze, M. 2014. Measuring Post Traumatic Stress Disorder in Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1): 579–582.

Coppersmith, G.; Leary, R.; Crutchley, P.; and Fine, A. 2018. Natural Language Processing of Social Media as Screening for Suicide Risk. *Biomedical informatics insights*, 10: 1178222618792860–1178222618792860. 30158822[pmid].

Gallo, I.; Ria, G.; Landro, N.; and Grassa, R. L. 2020. Image and Text fusion for UPMC Food-101 using BERT and CNNs. In *2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ)*, 1–6.

Guntuku, S. C.; Preotiuc-Pietro, D.; Eichstaedt, J. C.; and Ungar, L. H. 2019. What Twitter Profile and Posted Images Reveal About Depression and Anxiety. In *ICWSM*.

Hochreiter, S.; and Schmidhuber, J. 1997. Long Short-Term Memory. *Neural Comput.*, 9(8): 1735–1780.

Kingma, D. P.; and Ba, J. 2014. Adam: A Method for Stochastic Optimization.

Kumar, A.; and Garg, G. 2019. Sentiment analysis of multimodal twitter data. *Multimedia Tools and Applications*, 78(17): 24103–24119.

Le, Q.; and Mikolov, T. 2014. Distributed Representations of Sentences and Documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, II–1188–II–1196. JMLR.org.

Martínez-Castaño, R.; Pichel, J. C.; and Losada, D. E. 2020. A Big Data Platform for Real Time Analysis of Signs of Depression in Social Media. *International journal of environmental research and public health*, 17(13): 4752. PMC7370096[pmcid].

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient Estimation of Word Representations in Vector Space. *Proceedings of Workshop at ICLR*, 2013.

Pantic, I. 2014. Online social networking and mental health. *Cyberpsychology, behavior and social networking*, 17(10): 652–657. 25192305[pmid].

Rajaraman, A.; and Ullman, J. D. 2011. *Data Mining*, 1–17. Cambridge University Press.

Safa, R.; Bayat, P.; and Moghtader, L. 2022. Automatic detection of depression symptoms in twitter using multimodal analysis. *The Journal of Supercomputing*, 78(4): 4709–4744.

Stephen, J.; and P., P. 2019. Detecting the magnitude of depression in Twitter users using sentiment analysis. *International Journal of Electrical and Computer Engineering (IJECE)*, 9: 3247.

Wang, Y.; Wang, Z.; Li, C.; Zhang, Y.; and Wang, H. 2020. A Multitask Deep Learning Approach for User Depression Detection on Sina Weibo.