Augmenting Semantic Representation of Depressive Language: From Forums to Microblogs

Nawshad Farruque(✉), Osmar Zaiane, and Randy Goebel

Department of Computing Science, University of Alberta,
Edmonton, AB T6G 2R3, Canada
{nawshad,zaiane,rgoebel}@ualberta.ca

Abstract. We discuss and analyze the process of creating word embedding feature representations specifically designed for a learning task when annotated data is scarce, like depressive language detection from Tweets. We start from rich word embedding pre-trained from a general dataset, then enhance it with embedding learned from a domain specific but relatively much smaller dataset. Our strengthened representation portrays better the domain of depression we are interested in as it combines the semantics learned from the specific domain and word coverage from the general language. We present a comparative analyses of our word embedding representations with a simple bag-of-words model, a well known sentiment lexicon, a psycholinguistic lexicon, and a general pre-trained word embedding, based on their efficacy in accurately identifying depressive Tweets. We show that our representations achieve a significantly better F1 score than the others when applied to a high quality dataset.

Keywords: Machine learning · Natural language processing · Distributional semantics · Major Depressive Disorder · Social media

1 Introduction

Depression or Major Depressive Disorder (MDD) is regarded as one of the most commonly identified mental health problems among young adults in developed countries, accounting for 75% of all psychiatric admissions [3]. Most people who suffer from depression do not acknowledge it, for various reasons, ranging from social stigma to just ignorance; this means that a vast majority of depressed people remain undiagnosed. Lack of proper diagnosis eventually results in suicide, drug abuse, crime and many other societal problems. For example, depression has been found to be a major cause behind 800,000 deaths committed through suicide each year worldwide[1]. Moreover, the economic burden created by depression is estimated to have been 210 billion USD in 2010 [14] in the USA alone. Hence, detecting, monitoring and treating depression is very important and there

[1] https://who.int/mental_health/prevention/suicide/suicideprevent/en/.

U. Brefeld et al. (Eds.): ECML PKDD 2019, LNAI 11908, pp. 359–375, 2020.
https://doi.org/10.1007/978-3-030-46133-1_22

is a huge need for effective, inexpensive and almost real-time interventions. In such a scenario, social media provide the foundation of a remedy. Social media are very popular among young adults where depression is prevalent [15]. In addition, it has been found that people who are otherwise socially aloof (and more prone to having depression) can be very active in the social media platforms [9]. As a consequence, there has been significant depression detection research conducted already, based on various social media components, such as social network size, social media behavior, and language used in social media posts. It is found that, among these multi-modalities, human language alone can be a very good predictor of depression [9]. In the next sections we provide a brief summary of earlier research together with some background supporting our formulation of our proposed methods identifying depression from Tweets.

2 Background and Motivation

Previous studies suggest that the words we use in our daily life can express our mental state, mood and emotion [29]. Therefore analyzing language to identify and monitor human mental health problems has been regarded as an appropriate avenue of mental health modeling. With the advent of social media platforms, researchers have found that social media posts can be used as a good proxy for our day to day language usage [9]. There have been many studies that identify and monitor depression through social media posts in various social media, such as, Twitter [7,9,30], Facebook [24,33] and online forums [39].

Depression detection from social media posts can be specified as a low resource supervised classification task because of the paucity of valid data. Although there is no concrete precise definition of valid data, previous research emphasizes collecting social media posts, which are either validated by annotators as carrying clues of depression, or coming from the people who are clinically diagnosed as depressed, or both. Based on the methods of depression intervention using these data, earlier research can be mostly divided into two general categories: (1) post-specific depression detection (or depressive language detection) [7,16,38], and (2) user-specific depression detection, which considers all the posts made by a depressed user in a specific time window [31,32]. The goal of (1) is to identify depression in a more fine grained level, i.e., in social media posts, which further helps in identifying depression inclination of individuals when analyzed by method (2).

For the post specific depression detection task, previous research concentrate on the extraction of depression specific features used to train machine learning models, e.g., building depression lexicons based on unigrams present in posts from depressed individuals [9], depression symptom related unigrams curated from depression questionnaires [4], metaphors used in depressive language [25], or psycholinguistic features in LIWC [37]. For user specific depression identification, variations of topic modeling have been popular to identify depressive topics and use them as features [31,32]. But recently, some research has used convolutional neural network (CNN) based deep learning models to learn feature representations [28,39]. Most deep learning approaches require a significant

volume of labelled data to learn the depression specific embedding from scratch, or from a pre-trained word embedding in a supervised manner. So, in general, both post level and user level depression identification research emphasize the curation of labelled social media posts indicative of depression, which is a very expensive process in terms of time, human effort, and cost. Moreover, previous research showed that a robust post level depression identification system is an important prerequisite for accurately identifying depression at the user level [16]. In addition, most of this earlier research leveraged Twitter posts to identify depression because a huge volume of Twitter posts are publicly available.

Therefore the motivation of our research comes from the need for a better feature representation specific to depressive language, and reduced dependency on a large set of (human annotated) labelled data for depressive Tweet detection task. We proceed as follows:

1. We create a word embedding space that encodes the semantics of depressive language from a small but high quality depression corpus curated from depression related public forums.
2. We use that word embedding to create feature representations for our Tweets and feed them to our machine learning models to identify depressive Tweets; this achieves good accuracy, even with very small amount of labelled Tweets.
3. Furthermore, we adjust a pre-trained Twitter word embedding based on our depression specific word embedding, using a non-linear mapping between the embeddings (motivated by the work of [21] and [35] on bilingual dictionary induction for machine translation), and use it to create feature representation for our Tweets and feed them to our machine learning models. This helps us achieve around 3% higher F1-score than our strongest baseline in depressive Tweets detection.

Accuracy improvements mentioned in points 2 and 3 above are true for a high quality dataset curated through rigorous human annotation, as opposed to the low quality dataset with less rigorous human annotation; this indicates the effectiveness of our proposed feature representations for depressive Tweets detection. To the best of our knowledge, ours is the first effort to build a depression specific word embedding for identifying depressive Tweets, and to formulate a method to gain further improvements on top of it, then to present a comprehensive analysis on the quantitative and qualitative performance of our embeddings. Throughout our paper, we use the phrase "word embedding" as an object that consists of word vectors. So by "word embeddings" we mean multiple instances of that object.

3 Datasets

Here we provide the details of our two datasets that we use for our experiments and their annotation procedure, the corpus they are curated from and their quality comparisons.

3.1 Dataset1

Dataset1 is curated by the ADVanced ANalytics for data SciencE (ADVANSE) research team at the University of Montpellier, France [38]. This dataset contains Tweets having key-phrases generated from the American Psychiatric Association (APA)'s list of risk factors and the American Association of Suicidology (AAS)'s list of warning signs related to suicide. Furthermore, they randomly investigated the authors of these Tweets to identify 60 distressed users who frequently write about depression, suicide and self mutilation. They also randomly collected 60 control users. Finally, they curated a balanced and human annotated dataset of a total of around 500 Tweets, of which 50% Tweets are from distressed and 50% are from control users, with the help of seven annotators and one professional psychologist. The goal of their annotation was to provide a distress score (0–3) for each Tweet. They reported a Cohen's kappa agreement score of 69.1% for their annotation task. Finally, they merged Tweets showing distress level 0, 1 as control Tweets and 2, 3 as distressed Tweets. *Distressed Tweets* carry signs of suicidal ideation, self-harm and depression while control Tweets are about daily life occurrences, such as weekend plans, trips and common distress such as exams, deadlines, etc. We believe this dataset is perfectly suited for our task, and we use their distressed Tweets as our depressive Tweets and their control as our control.

3.2 Dataset2

Dataset2 is collected by a research group at the University of Ottawa [16]. They first filtered depressive Tweets from #BellLetsTalk2015 (a Twitter campaign) based on keywords such as, suffer, attempt, suicide, battle, struggle and first person pronouns. Using topic modeling, they removed Tweets under the topics of public campaign, mental health awareness, and raising money. They further removed Tweets which contain mostly URLs and are very short. Finally, from these Tweets they identified 30 users who self-disclosed their own depression, and 30 control users who did not. They employed two annotators to label Tweets from 10 users as either depressed or non-depressed. They found that their annotators labelled most Tweets as non-depressed. To reduce the number of non-depressive Tweets, they further removed neutral Tweets from their dataset, as they believe neutral Tweets surely do not carry any signs of depression. After that, they annotated Tweets from the remaining 50 users with the help of two annotators with a Cohen's kappa agreement score of 67%. Finally, they labelled a Tweet as depressive if any one of their two annotators agree, to gather more depressive Tweets. This left them with 8,753 Tweets with 706 depressive Tweets.

3.3 Quality of Datasets

Here we present a comparative analysis of our datasets based on the linguistic components present in them relevant to depressive language detection and their curation process as follows:

Analysis Based on Linguistic Components Present in the Datasets: For this analysis, we use Linguistic Inquiry and Word Count (LIWC) [37]. LIWC is a tool widely used in psycholinguistic analysis of language. It extracts the percentage of words in a text, across 93 pre-defined categories, e.g., affect, social process, cognitive processes, etc. To analyse the quality of our datasets, we provide scores of few dimensions of LIWC lexicon relevant for depressive language detection [9,18,26], such as, 1st person pronouns, anger, sadness, negative emotions, etc (see Table 1 for the complete list) for the depressive Tweets present both in our datasets. The bold items in that table shows significant score differences in those dimensions for both datasets and endorses the fact that Dataset1 indeed carries more linguistic clues of depression than Dataset2 (the higher the score, the more is the percentage of words from that dimension is present in the text). Moreover, Tweets labelled as depressive in Dataset2 are mostly about common distress of everyday life unlike those of Dataset1, which are indicative of severe depression. We provide few random samples of Tweets from Dataset1 and Dataset2 depressive Tweets at Table 2 and their corresponding word clouds at Fig. 1 as well.

Table 1. Scores of Dataset1 and Dataset2 in few LIWC dimensions relevant to depressive language detection (bold categories have significant score differences).

LIWC category	Example words	Dataset1 depressive Tweets score	Dataset2 depressive Tweets score
1st person pronouns	I, me, mine	**12.74**	7.06
Negations	No, not, never	**3.94**	2.63
Positive emotion	Love, nice, sweet	**2.79**	2.65
Negative emotion	Hurt, ugly, nasty	**8.59**	6.99
Anxiety	Worried, fearful	0.72	**1.05**
Anger	Hate, kill, annoyed	**2.86**	2.51
Sadness	Crying, grief, sad	**3.29**	1.97
Past focus	Ago, did, talked	2.65	**3**
Death	Suicide, die, overdosed	**1.43**	0.44
Swear	Fuck, damn, shit	**1.97**	1.39

Table 2. Sample Tweets from Dataset1 and Dataset2

Datasets	Depressive Tweets
Dataset1	"I wish I could be normal and be happy and feel things like other people"
	"I feel alone even when I'm not"
	"Yesterday was difficult...and so is today and tomorrow and the days after..."
Dataset2	"Last night was not a good night for sleep... so tired And I have a gig tonight... yawnnn"
	"So tired of my @NetflixCA app not working, I hate Android 5"
	"I have been so bad at reading Twitter lately, I don't know how people keep up, maybe today I'll do better"

Fig. 1. Dataset1 depressive Tweets word cloud (left) and Dataset2 depressive Tweets word cloud (right)

Analysis Based on Data Curation Process: We think Dataset2 is of lower quality compared to Dataset1 for the following reasons: (1) this dataset is collected from the pool of Tweets which is a part of a mental health campaign, and thus compromises the authenticity of the Tweets; (2) the words used for searching depressive Tweets are not validated by any depression or suicide lexicons; (3) although two annotators were employed (none of them are domain experts) to label the Tweets, a Tweet was considered as depressive if at least one annotator labelled it as depressive, which introduced more noise in the data; (4) it is not confirmed how neutral Tweets were identified, since the neutral Tweets may convey depression as well; (5) a person was identified as depressed if s/he disclose their depression, but it was not mentioned how these disclosures were determined. Simple regular expression based methods to identify these self disclosures can introduce a lot of noise in the data. In addition, these self disclosures may not be true.

3.4 Depression Corpus

To build depression specific word embedding we curate our own depression corpus. For this, we collect all the posts from the Reddit depression forum: r/depression[2] from the year 2006 to 2017 and all those from Suicide Forum[3] up through the year 2017 and concatenated to total of 856,897 posts. We choose to use these forums because people who post anonymously in these forums usually suffer from severe depression and share their struggle with depression and its impact in their personal lives [8]. We believe these forums contain useful semantic components indicative of depressive language. Technical and ethical aspects of building word embedding representation on this corpora are presented in Sects. 4.3 and 7 respectively.

4 Feature Extraction Methods

4.1 Bag-of-Words (BOW)

We represent each Tweet as a vector of vocabulary terms and their frequency counts in that Tweet, also known as bag-of-words. The vocabulary terms refer to the most frequent 400 terms existing in the training set. Before creating the vocabulary and the vector representation of the Tweets, we perform the following preprocessing: (1) we make the Tweets all lowercase; (2) tokenize them using NLTK Tweet tokenizer[4]; (3) remove all stop words except the first person pronouns such as, I, me and my (because they are useful for depression detection). The reason for using Tweet tokenizer is to consider Tweet emoticons (:-)), hashtags (#Depression) and mentions (@user) as single tokens.

4.2 Lexicons

We have tried several emotion and sentiment lexicons, such as, LabMT [10], Emolex [23], AFINN [27], LIWC [37], VADER [12], NRC-Hashtag-Sentiment-Lexicon (HSL) [17] and CBET [34]. Among these lexicons we find LIWC and HSL perform the best and hence we report the results of these two lexicons. The following subsections provide a brief description of LIWC, HSL and lexicon-based representation of Tweets.

Linguistic Inquiry and Word Count (LIWC): LIWC is a tool widely used in psycholinguistic analysis of language. It extracts the percentage of words in a text, across 93 pre-defined categories, e.g., affect, social process, cognitive processes, etc. A text input is converted into a 93 length vector representation of that text (in our case Tweets), that are input for our machine learning models. Note that LIWC has been widely used as a good baseline for depressive Tweet detection in earlier research [5,26].

[2] https://reddit.com/r/depression/.
[3] https://suicideforum.com/.
[4] www.nltk.org/api/nltk.tokenize.html.

NRC Hashtag Sentiment Lexicon (HSL): This lexicon consists of 54,129 unigrams, each of which has a score that shows the difference between the PMI score of that unigram being associated with positive Tweets and negative Tweets (Tweets having positive/negative hashtags, respectively). The polarity of the score represents the polarity of the sentiment and the magnitude represents the degree of associativity with the sentiments. In our experiments, we tokenize each Tweet as described in Sect. 4.1, then use the lexicon to determine a score for each token in the Tweet, then sum them to provide a single value for each Tweet, which represents the sentiment and magnitude. We use that value as a feature for our machine learning models.

4.3 Distributed Representation of Words

The use of word embedding has been crucial in many downstream NLP tasks; domain specific embedding perform better than generic ones for domain specific tasks [1,2,36], and there have been many attempts till-to-date to make generic embedding useful for particular domain, for example, lexicon based retrofitting [11,40], and supervised retrofitting [28]. Lexicon-based retrofitting algorithms have an inherent problem of limited vocabulary coverage, where supervised retrofitting requires huge amount of labelled data. In contrast, we retrofitted a general pre-trained embedding based on the semantics present in depression specific embedding through a non-linear mapping between them. Our depression-specific embedding is created in an unsupervised manner from depression forum posts. Moreover, through our mapping process we learn a transformation matrix (see Eq. 3), which can be used to predict embedding for Out of Vocabulary (OOV) words, and helps achieve better accuracy.

Word Embedding Representation of Tweets: To represent a Tweet using word embedding, we take the average of the word vectors of the individual words in that Tweet, ignoring the ones that are out of vocabulary (OOV), i.e. absent in the word embedding vocabulary.

General Twitter Word Embedding (GTE): We use a pre-trained skip-gram word embedding having 400 dimensions learned from 400 million Tweets with vocabulary size of $3,039,345$ words [13] as a representative of word embedding learned from general dataset (in our case, Tweets), because we believe this has the most relevant vocabulary for our task. The creator of this word embedding use negative sampling ($k = 5$) with context window size $= 1$ and mincount $= 5$. Since it is pre-trained, we do not have control over the parameters it uses and simply use it as is. We use pre-trained embedding to avoid difficulties arising from creating our own from a huge dataset.

Depression Specific Word Embedding (DSE): We create a 400 dimensional depression specific word embedding (DSE) on our curated depression corpus as mentioned in Subsect. 3.4. First, we identify sentence boundaries in our

corpora based on punctuations such as (period, question mark and exclamation). Then we feed each sentence in skip-gram based word2vec implementation in gensim[5]. We use negative sampling ($k = 5$) with the context window size = 5 and mincount = 10 for the training of this word embedding. DSE has a vocabulary size of 29,930 words. We choose skip-gram for this training because skip-gram learns good embedding from small corpus [20].

Adjusted Twitter Word Embedding (ATE): A Non-linear Mapping Between GTE and DSE: In this step, we find a non-linear mapping between GTE and DSE. The goal of this mapping is to adjust GTE, such that it reflects the semantics of DSE. To do this, we use a Multilayer Perceptron Regressor (MLPR) having a single hidden layer with 400 hidden units and Rectified Linear Unit (ReLU) activations, that tries to minimize the Mean Squared Error (MSE) loss function, $\mathcal{F}(\theta)$ in Eq. 1, using stochastic gradient descent:

$$\mathcal{F}(\theta) = \arg\min_{\theta}(\mathcal{L}(\theta) + \alpha||\theta||_2^2) \tag{1}$$

where

$$\mathcal{L}(\theta) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{n}\sum_{j=1}^{n}(g_j(x) - y_j)^2 \tag{2}$$

and

$$g(x) = b_1 + (W_1(ReLU(b_2 + W_2 x))) \tag{3}$$

here, $g(x)$ is the non-linear mapping function between the embedding x (from GTE) and y (from DSE) of a word $w \in V$, where, V is a common vocabulary between GTE and DSE; W_1 and W_2 are the hidden-to-output and input-to-hidden layer weight matrices respectively, b_1 is the output layer bias vector and b_2 is the hidden layer bias vector (all these weights are indicated as θ in Eq. 1) and α is the $l2$ regularization parameter. In Eq. 2, m and n are respectively the length of V (in our case it is 28,977) and dimension of word vectors (in our case it is 400). Once the MLPR learns the θ that minimizes $\mathcal{F}(\theta)$, it is used to predict the vectors for all the words in GTE. After this step, we finally get adjusted Twitter word embedding representation which encodes the semantics of depression forums as well as word coverage from Tweets, we name it Adjusted Twitter word Embedding (ATE). We use scikit-learn MLPR implementation[6] with default parameter settings for our non-linear mapping, except random state, which is set to 1. This entire process of mapping between embeddings is depicted in Fig. 2.

Conditions for Embedding Mapping/Adjustment: Our non-linear mapping between two embeddings works better given that those two embeddings

[5] https://radimrehurek.com/gensim/models/word2vec.html.

[6] https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html.

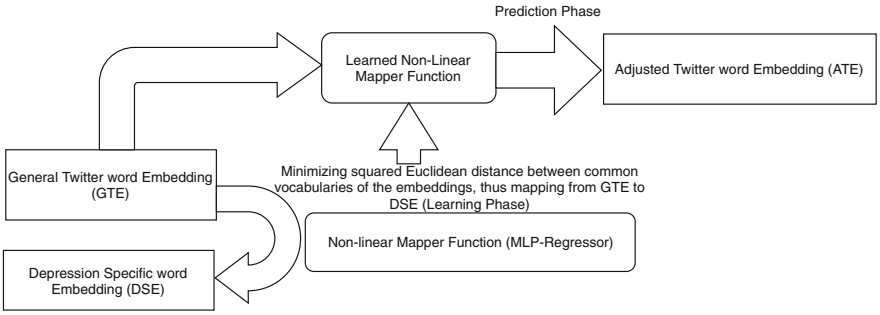

Fig. 2. Non-linear mapping of GTE to DSE (creation of ATE)

are created from the same word embedding creation algorithm (in our case skip-gram) and have same number of dimensions (i.e. 400). We also find that a non-linear mapping between our GTE and DSE produces slightly better ATE than a linear mapping for our task, although the former is a bit slower.

5 Experimental Setup

We report the results of best performing combinations out of all the 24 combinations from six feature extraction methods, such as, BOW, HSL, LIWC, GTE, DSE and ATE (described in Sect. 4) and four machine learning models, including Multinomial Naïve Bayes (MNB), Logistic Regression (LR), Linear Support Vector Machine (LSVM), and Support Vector Machine with radial basis kernel function (RSVM), for both datasets.

We also report the results of two experiments, one by [38] for Dataset1 and another by [16] for Dataset2, where they use their own depression lexicon as a feature representation for their machine learning models. For a single experiment, we split all our data into a disjoint set of training (70% of all the data) and testing (30% of all the data) (see Table 3). We use stratified sampling so that the original distribution of labels is retained in our splits. Furthermore, with the help of 10-fold cross validation in our training set, we learn the best parameter settings for all our model-feature extraction combinations except MNB that requires no such parameter tuning. For the SVMs and LR, we tune the parameter, $C \in \{2^{-9}, 2^{-7}, \ldots, 2^{5}\}$ and additionally, $\gamma \in \{2^{-11}, 2^{-9}, \ldots, 2^{2}\}$ for the RSVM. We use min-max feature scaling for all our features. For our imbalanced dataset we use cost sensitive LR and SVMs (as listed above) with class weights inversely proportional to the class frequencies in our input data.

We then find the performance of the best model on our test set. We have run 30 such experiments on 30 random train-test splits. Finally, we report the performance of our model-feature extraction combinations based on the Precision (Prec.), Recall (Rec.) and F1 score averaged over the test sets of those

30 experiments. See Tables 4 and 5 and Fig. 3 depicting the experiment results. We use scikit-learn library[7] for all our experiments.

Table 3. Number of Tweets in the train and test splits for the two datasets. The number of depressive Tweets is in parenthesis.

Datasets	Train	Test
Dataset1	355(178)	152(76)
Dataset2	6127(613)	2626(263)

6 Results Analysis

In this section we report quantitative and qualitative performance analysis of our embeddings in detecting depressive Tweets.

6.1 Quantitative Performance Analysis

In general, Tweet level depression detection is a tough problem and a good F1 score is hard to achieve [16]. Still, our LR-ATE achieves an F1 score of 0.81 which is around 3% better than our strongest baseline (GTE) and 10% better than [38] with F1 score of 0.71 in Dataset1. All the word embedding based models achieve on avg. 0.7926 F1 score which is 8% better than [38]. See Table 4 and Fig. 3.

Table 4. Average Prec., Rec. and F1 scores on Dataset1 best model-feat combination experiments

Category	Model-Feat.	Prec.	Rec.	F1
Baselines	LR-BOW	0.6967	0.8264	0.7548
	LR-HSL	0.6238	**0.9114**	0.7400
	LR-LIWC	0.7409	0.7772	0.7574
	LR-GTE	0.7694	0.7976	0.7822
Our Models	LR-DSE	0.7392	0.8411	0.7852
	LR-ATE	**0.7846**	0.8394	**0.8104**
Prev. Res.	[38]	0.71	0.71	0.71

In Dataset2, which is imbalanced (only 10% samples are depressive Tweets), all the word embedding based models achieve on avg. 0.4284 F1 score which is around 16% better than the best F1 achieved by [16] in the same dataset. However in that dataset, GTE is 4% better than DSE and 0.97% better than ATE, see Table 5 and Fig. 3.

[7] https://scikit-learn.org/stable/.

Table 5. Average Prec., Rec. and F1 scores on Dataset2 best model-feat combination experiments

Category	Model-Feat.	Prec.	Rec.	F1
Baselines	RSVM-BOW	0.2374	0.5296	0.3260
	RSVM-HSL	0.1168	0.6513	0.1980
	RSVM-LIWC	0.2635	**0.6750**	0.3778
	RSVM-GTE	0.3485	0.6305	**0.4448**
Our Models	RSVM-DSE	0.3437	0.5198	0.4053
	RSVM-ATE	**0.3497**	0.5821	0.4351
Prev. Res.	[16]	0.1706	0.5939	0.265

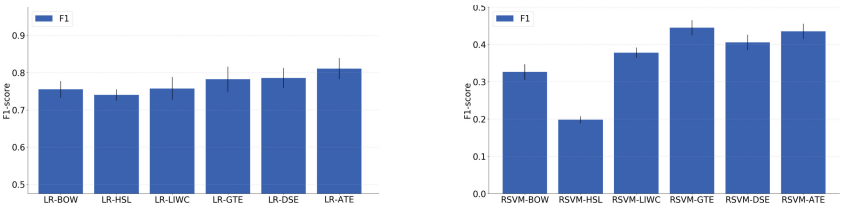

Fig. 3. Error bars for F1 scores on Dataset1 experiments (left) and Dataset2 experiments (right)

In Dataset1, HSL has the best recall, while LIWC has the best recall in Dataset2. In both datasets, HSL has the worst precision, while, LIWC and word embedding based methods have acceptable precision and recall.

6.2 Qualitative Performance Analysis

Here we report correctly predicted Tweets in Table 6 by LR-ATE (our overall best model) and LSVM-ATE (second best model), which are mistakenly predicted as control Tweets (i.e. false negatives) when LR-GTE and LSVM-GTE are used respectively in a test set from Dataset1. The first example from Table 6, *"Tonight may definitely be the night"*, may be indicative of suicidal ideation and should not be taken lightly, also, the second one *"0 days clean"* is the trade mark indication of continued self-harm, although many depression detection models will predict these as normal Tweets.

Additionally, we plot 2-dimensional Principal Component Analysis (PCA) projection of the embedding for LIWC 'POSEMO' and 'NEGEMO' words conveying positive and negative emotions respectively occurred most frequently in our datasets. Also, we use a word *sleepless*, indicating the common sleep problem encountered by many of the depressed people (see Fig. 4). We show that these

Table 6. False negative depressive Tweets when GTE is used, correctly predicted when ATE is used in a test set from Dataset1.

Tweets
"Tonight may definitely be the night"
"0 days clean"
"I'm a failure"
"I understand you're 'busy', but fuck that ... people make time for what they want"
"'Worthless' repeats in her mind as she holds on to what's left of her..."

words clearly form defined clusters, C1 (contains words carrying depressive sentiment) and C2 (contains words carrying non-depressive sentiment) in ATE where in GTE, these clusters overlap. Also, under each of these clusters we notice there are sub-clusters of closely related emotions. Although these sub-clusters are easily identifiable in ATE, they are almost absent in GTE, for example, *fuck* and *hate* are the words mostly used by the depressed people and should belong to C1 but they belong to C2 for GTE, overlapped with *thankful* and *love*. So by adjusting the embedding space of GTE based on DSE, we basically make the clear distinction among the words carrying depressive sentiment and the ones which do not, in their vector space.

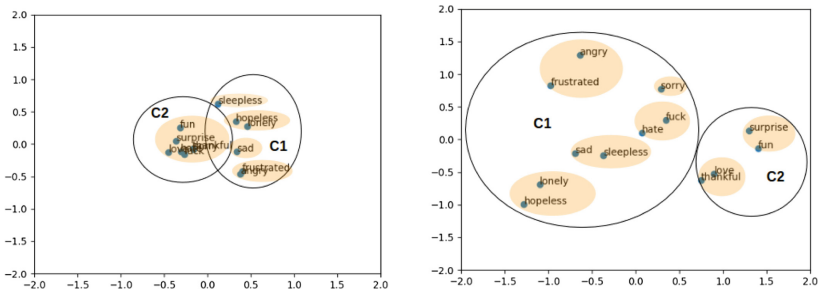

Fig. 4. 2-dimensional PCA projection of emotion carrying words in General Twitter word Embedding (GTE) (left) and Adjusted Twitter word Embedding (ATE) (right)

Overall, in both datasets, word embedding based methods perform much better than BOW and lexicons. The reason is, both GTE and ATE have bigger vocabulary and better feature representation than BOW and lexicons. Among non word embedding methods, BOW and LIWC perform better than HSL, because the former provide better discriminating features than the latter. However, in Dataset1, ATE is better than both GTE and DSE with DSE performing close enough. This confirms that DSE can capture the semantics of depressive

language very well. ATE is superior in performance because it leverages both the vocabulary coverage and semantics of a depressive language. In Dataset2, GTE turns out to be better than DSE with ATE performing closely, indicating that Tweet samples in Dataset2 are more about general distress than depression. In this case, the performance is affected mostly by the vocabulary size than the depressive language semantics.

Another important observation is that, in Dataset1 we see the LR classifier performs better where in Dataset2, RSVM works better than all others. We think it is because, both LR and RSVM consider dependency among features unlike feature independence assumption in MNB. LR performing better in Dataset1 confirms that in Dataset1 depressive and non-depressive Tweets are distinct to each other and linearly separable, while Dataset2 Tweets are not and a non-linear classifier such as RSVM is needed.

7 Ethical Concerns

We use *Suicide Forum* posts where users are strictly required to stay anonymous. Moreover, we use Reddit and Twitter *public* posts which incur minimal risk of user privacy violation as established by earlier research [6,19,22] utilizing same kind of data. Our word embeddings are built solely on the text and not on user data of the forums. No user identifiers or user profiles are stored by us as these are not required for our research. Moreover, we have our university research ethics office approval to use datasets released by other organization to us (like our Dataset1 and Dataset2) for conducting our research.

8 Conclusion

In this paper we empirically present the following observations for a high quality dataset:

- For depressive Tweets detection, we can use word embedding trained in an unsupervised manner on a corpus of depression forum posts, which we call Depression Specific word Embedding (DSE) and use it as a feature representation for our machine learning models and can achieve very good accuracy.
- Further, we can use DSE to adjust the general Twitter pre-trained word embedding (available off the shelf) through non-linear mapping between them. This Adjusted Twitter word Embedding (ATE) helps us achieve even better results for our task.
- We need not to depend on human annotated data or labeled data for any of our word embedding representation creation.
- Depression forum posts have specific distributed representation of words and it is different than that of general twitter posts and this is reflected in ATE, see Fig. 4.
- Our DSE and ATE embeddings are publicly available[8].

[8] https://doi.org/10.5281/zenodo.3361838.

9 Future Work

In the future we would like to analyze DSE more exhaustively to find any patterns in semantic clusters that specifically identify depressive language. We would also like to use ATE for Twitter depression lexicon induction and discover depressive Tweets. Thus, we can see a lot of promise in its use in creating semi-supervised learning based automated depression data annotation task later on.

Acknowledgements. We thank Natural Sciences and Engineering Research Council of Canada (NSERC) and Alberta Machine Intelligence Institute (AMII) for their generous support to pursue this research. We thank Prof. Greg Kondrak for his valuable advice and Bradley Hauer for his helpful suggestions. We also thank Roberto Vega and Shiva Zamani for their contribution in implementing standard text classification pipeline and initial baseline experiments.

References

1. Asgari, E., Mofrad, M.R.: Continuous distributed representation of biological sequences for deep proteomics and genomics. PLoS ONE **10**(11), e0141287 (2015)
2. Bengio, S., Heigold, G.: Word embeddings for speech recognition. In: Fifteenth Annual Conference of the International Speech Communication Association (2014)
3. Boyd, J.H., Weissman, M.M., Thompson, W.D., Myers, J.K.: Screening for depression in a community sample: understanding the discrepancies between depression symptom and diagnostic scales. Arch. Gen. Psychiatry **39**(10), 1195–1200 (1982)
4. Cheng, P.G.F., et al.: Psychologist in a pocket: lexicon development and content validation of a mobile-based app for depression screening. JMIR mHealth uHealth **4**(3), e88 (2016)
5. Coppersmith, G., Dredze, M., Harman, C.: Quantifying mental health signals in Twitter. In: Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pp. 51–60 (2014)
6. Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K.: 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, pp. 1–10 (2015)
7. De Choudhury, M.: Role of social media in tackling challenges in mental health. In: Proceedings of the 2nd International Workshop on Socially-Aware Multimedia, pp. 49–52. ACM (2013)
8. De Choudhury, M., De, S.: Mental health discourse on reddit: self-disclosure, social support, and anonymity. In: Eighth International AAAI Conference on Weblogs and Social Media (2014)
9. De Choudhury, M., Gamon, M., Counts, S., Horvitz, E.: Predicting depression via social media. In: Seventh International AAAI Conference on Weblogs and Social Media, p. 2 (2013)
10. Dodds, P.S., Harris, K.D., Kloumann, I.M., Bliss, C.A., Danforth, C.M.: Temporal patterns of happiness and information in a global social network: hedonometrics and Twitter. PLoS ONE **6**(12), e26752 (2011)
11. Faruqui, M., Dodge, J., Jauhar, S.K., Dyer, C., Hovy, E., Smith, N.A.: Retrofitting word vectors to semantic lexicons. In: Proceedings of NAACL (2015)

12. Hutto, C.J., Gilbert, E.: VADER: a parsimonious rule-based model for sentiment analysis of social media text. In: Eighth International AAAI Conference on Weblogs and Social Media (2014)
13. Godin, F., Vandersmissen, B., De Neve, W., Van de Walle, R.: Multimedia lab @ ACL WNUT NER shared task: named entity recognition for Twitter microposts using distributed word representations. In: Proceedings of the Workshop on Noisy User-Generated Text, pp. 146–153 (2015)
14. Greenberg, P.E., Fournier, A.A., Sisitsky, T., Pike, C.T., Kessler, R.C.: The economic burden of adults with major depressive disorder in the United States (2005 and 2010). J. Clin. Psychiatry **76**(2), 155–162 (2015)
15. Gustavson, K., Knudsen, A.K., Nesvåg, R., Knudsen, G.P., Vollset, S.E., Reichborn-Kjennerud, T.: Prevalence and stability of mental disorders among young adults: findings from a longitudinal study. BMC Psychiatry **18**(1), 65 (2018)
16. Jamil, Z., Inkpen, D., Buddhitha, P., White, K.: 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, pp. 32–40 (2017)
17. Kiritchenko, S., Zhu, X., Mohammad, S.M.: Sentiment analysis of short informal texts. J. Artif. Intell. Res. **50**, 723–762 (2014)
18. Kuppens, P., Sheeber, L.B., Yap, M.B., Whittle, S., Simmons, J.G., Allen, N.B.: Emotional inertia prospectively predicts the onset of depressive disorder in adolescence. Emotion **12**(2), 283 (2012)
19. Losada, D.E., Crestani, F.: A test collection for research on depression and language use. In: Fuhr, N., et al. (eds.) CLEF 2016. LNCS, vol. 9822, pp. 28–39. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-44564-9_3
20. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
21. Mikolov, T., Le, Q.V., Sutskever, I.: Exploiting similarities among languages for machine translation. arXiv preprint arXiv:1309.4168 (2013)
22. Milne, D.N., Pink, G., Hachey, B., Calvo, R.A.: CLPsych 2016 shared task: triaging content in online peer-support forums. In: Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology, pp. 118–127 (2016)
23. Mohammad, S.M., Turney, P.D.: NRC emotion lexicon. NRC Technical report (2013)
24. Moreno, M.A., et al.: Feeling bad on Facebook: depression disclosures by college students on a social networking site. Depress. Anxiety **28**(6), 447–455 (2011)
25. Neuman, Y., Cohen, Y., Assaf, D., Kedma, G.: Proactive screening for depression through metaphorical and automatic text analysis. Artif. Intell. Med. **56**(1), 19–25 (2012)
26. Nguyen, T., Phung, D., Dao, B., Venkatesh, S., Berk, M.: Affective and content analysis of online depression communities. IEEE Trans. Affect. Comput. **5**(3), 217–226 (2014)
27. Nielsen, F.Å.: A new ANEW: evaluation of a word list for sentiment analysis in microblogs. arXiv preprint arXiv:1103.2903 (2011)
28. Orabi, A.H., Buddhitha, P., Orabi, M.H., Inkpen, D.: Deep learning for depression detection of Twitter users. In: Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, pp. 88–97 (2018)
29. Pennebaker, J., Mehl, M., Niederhoffer, K.: Psychological aspects of natural language use: our words, our selves. Annu. Rev. Psychol. **54**, 547–577 (2003)

30. Reece, A.G., Reagan, A.J., Lix, K.L., Dodds, P.S., Danforth, C.M., Langer, E.J.: Forecasting the onset and course of mental illness with Twitter data. Sci. Rep. **7**(1), 13006 (2017)
31. Resnik, P., Armstrong, W., Claudino, L., Nguyen, T., Nguyen, V.A., Boyd-Graber, J.: Beyond LDA: exploring supervised topic modeling for depression-related language in Twitter. In: Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pp. 99–107 (2015)
32. Resnik, P., Garron, A., Resnik, R.: Using topic modeling to improve prediction of neuroticism and depression. In: Proceedings of the 2013 Conference on Empirical Methods in Natural, pp. 1348–1353. Association for Computational Linguistics (2013)
33. Schwartz, H.A., et al.: Towards assessing changes in degree of depression through Facebook. In: Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pp. 118–125 (2014)
34. Shahraki, A.G., Zaïane, O.R.: Lexical and learning-based emotion mining from text. In: International Conference on Computational Linguistics and Intelligent Text Processing (CICLing) (2017)
35. Smith, S.L., Turban, D.H., Hamblin, S., Hammerla, N.Y.: Offline bilingual word vectors, orthogonal transformations and the inverted softmax. arXiv preprint arXiv:1702.03859 (2017)
36. Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., Qin, B.: Learning sentiment-specific word embedding for Twitter sentiment classification. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), vol. 1, pp. 1555–1565 (2014)
37. Tausczik, Y.R., Pennebaker, J.W.: The psychological meaning of words: LIWC and computerized text analysis methods. J. Lang. Soc. Psychol. **29**(1), 24–54 (2010)
38. Vioulès, M.J., Moulahi, B., Azé, J., Bringay, S.: Detection of suicide-related posts in Twitter data streams. IBM J. Res. Dev. **62**(1), 7:1–7:12 (2018)
39. Yates, A., Cohan, A., Goharian, N.: Depression and self-harm risk assessment in online forums. arXiv preprint arXiv:1709.01848 (2017)
40. Yu, L.C., Wang, J., Lai, K.R., Zhang, X.: Refining word embeddings using intensity scores for sentiment analysis. IEEE/ACM Trans. Audio Speech Lang. Process. (TASLP) **26**(3), 671–681 (2018)