Learning Depression Patterns from MyPersonality and Reddit

Weiwei Yang, Xuetong Sun, Ruofei Du
UMIACS, Computer Science Department
University of Maryland
College Park, MD, 20742, USA
{wwyang,xtsun,ruofei}@cs.umd.edu

Abstract

Millions of Americans suffer from mental health problems caused by negative sentiment such as depression and neuroticism. Nevertheless, traditional clinical diagnosis is not preferred as a large-scale approach due to the high expenditures and subjective judgement. In this paper, we conduct exploratory data analysis on three datasets from MyPersonality and Reddit, train and compare different predictive models for depression and neuroticism with various features. Our experimental results indicate promising feasibility to identify such mental health problems using natural language processing techniques. Moreover, we discuss the differences among χ^2-Test, PMI, LDA, between single and multiple feature selections, and between SVM and SLDA. Thus, we accomplish and overfulfill all of our proposed goals. Eventually, we propose several directions for future research.

1 Introduction

It is widely acknowledged that negative emotion, such as *stress, anger, frustration, depression* and *neuroticism* could consume people's mental energy in life. Psychologists (Gross and Muñoz, 1995; John and Srivastava, 1999) have shown that failure in regulating such emotion could cause serious mental health problems. Just in the United States, solving such problems costs hundreds of billions of dollars every year. Unfortunately, traditional clinical diagnosis could hardly be conducted on a large scale due to the high expenditures, subjective report and demand on psychological experts. Hence, our research questions are: can we apply natural language processing techniques to tackle mental health problems automatically? More specifically, can we learn depression patterns from numerous messages in social media?

Nowadays, social media plays an increasingly significant role in people's daily life because of its interactivity, popularity and social relevance. (Kaplan and Haenlein, 2010) Everyday, billions of users create, share and exchange messages about their life and personal feelings on websites like *Facebook*, *Reddit* and *Twitter*. These data can be used for diagnosis of depression on a large scale. Prior research has revealed that people talk about their negative emotion and psychological treatment in social media. (Park et al., 2012) We are also inspired by recent work (De Choudhury et al., 2013b; Resnik et al., 2013; Coppersmith et al., 2014) that use various natural language processing techniques and datasets to investigate mental disorders.

In this project, we apply what we learnt from *CMSC 733 Computational Linguistic II* by *Prof. Philip Resnik* to find out who are potentially affected by depression on three datasets from MyPersonality[1] and Reddit[2].

First, we conduct exploratory data analysis on the datasets. To gain the overview of the data, we extract high-frequency unigrams and collocations in bigrams from people's posts in their social networks and visualize them in a tag cloud. To investigate the details, we seek for words and phrases that we believe are related to depression to see if there is any difference between the languages used by potentially depressed and non-depressed people.

Next, we use supervised classification to categorize each person with depressed or non-depressed label based on his or her post(s). In the feature selection stage, we filter stop words and extract linguistic features using χ^2-test, pointwise mutual information (PMI), Latent Dirichlet Allocation (LDA) and combinations of these. In the training stage, we apply both support vector ma-

[1] http://mypersonality.org
[2] https://www.reddit.com

chine (SVM) and supervised LDA (SLDA) and observe the differences. In the testing stage, we use a 5-fold cross validation on the datasets and report the average precision, recall and F1 value.

Eventually, we discuss what combination of features and models perform well and why.

2 Background and Related Work

Motivated by the growing concern for mental health problems, especially negative emotional signals (*e.g. depression and neuroticism*) from social media, our work builds upon a rich literature of prior research by psychologists and computer scientists on psychological diagnostic approaches, automated language analytics, as well as prior arts in computational linguistics and machine learning.

2.1 Challenges in Mental Health Problems

Solving mental health problems are becoming increasingly challenging in the United States. According to (U.S. Department of Health and Human Services, 2009), mental health accounted for 6.3% of all health expenditures. The total cost rose from $35.2B to $147B between 1996 and 2009. According to the Centers for Disease Control and Prevention in the US, the ratio of postpartum depression from new mothers, which typically begins in the first month after giving birth has risen to 1 in every 9 (Miller, 2002). For American children, diagnosis of autistic spectrum disorders has risen to 1 in every 68 (Wingate et al., 2014). For young people between 10 and 24 years old, suicide has become the third leading cause of death.[3]

2.2 Psychological Diagnostic Approaches

Manual coding of patient language has been widely applied by clinical psychologists to diagnose formal thought disorders (American Psychiatric Association, 2013). Past empirical research has modeled human personality into different categories to help identifying mental health problems. For example, the Big Five personality traits (Costa and MacCrae, 1992) are widely used in interviews, self-descriptions and observation. It classifies personal traits into five broad domains including *openness, conscientiousness, extraversion, agreeableness, and neuroticism*. People with higher levels of neuroticism have been proved to be at higher risk of depression and anxiety. (American Psychiatric Association, 2013) Nevertheless, traditional psychological diagnosis suffers from the following issues:

1. **Accessibility** According to (American Psychological Association, 2002), many patients in rural areas lack accessibility to *qualified* clinicians for psychological evaluations.

2. **Accuracy** Subjective norm-referenced self-reports usually depend on patients' awareness and motivation, which leads to inaccuracy in clinical assessment. (Kessler and Üstün, 2004)

3. **Scalability** Diagnosis of mental health by interview usually costs much time, money and human endeavor from psychological clinicians. (Clark and Drake, 1994) Worse still, some patients without adequate insurance may not be able to afford clinical diagnosis.

2.3 Population-level Analysis

Survey is the main vehicle to investigate mental health problems for population-level analysis. In the old days, the Behavioral Risk Factor Surveillance System (BRFSS) (Centers for Disease Control and Prevention, 2001) has been widely used to identify behavioral risk factors. Another example is Postpartum Depression Predictors Inventory (PPDI), which reflects prenatal depression, life stress, lack of social support, maternity blues and so on (Beck, 1998). Nonetheless, such surveys are usually conducted via phone interviews, which suffer from significant cost and long delays between data collection and results or insights from the data.

Nowadays, researchers turn to social media to tackle challenges in mental health. (Park et al., 2012) found the initial evidence that people share their depression and even their treatment on Twitter. (De Choudhury et al., 2013b) solicit Twitter participants to post their self-report depression scale via CES-D (Radloff, 1977) and analyze linguistic and behavioral patterns from the data. Previous research also uses personality test (Schwartz and Eichstaedt, 2013) and depression battery (De Choudhury, 2013). In addition to depression, (Coppersmith et al., 2014) gather Twitter data to identify post-traumatic stress disorder (PTSD), bipolar disorder, and seasonal affective disorder (SAD).

[3] `http://goo.gl/GMFUm4`

2.4 Automated Language Analytics for Sentiment Analysis

As for automated language processing, linguistic inquiry word count (LIWC) (Pennebaker and King, 1999) has been widely applied to investigate linguistic signals for assorted mental health issues (Pennebaker et al., 2003; Golder and Macy, 2011; De Choudhury et al., 2013a). (Greene and Resnik, 2010) built a strong supervised predictive model between implicit sentiment and linguistic features such as observable proxies for underlying semantics (OPUS). An annotated suicide note corpus was collected by (Pestian et al., 2012) to help psychiatrists understand the emotions of people who have suicidal thoughts. (Resnik et al., 2013) illustrate how topic modeling using LDA (Blei et al., 2003) can be applied in the prediction of depression and neuroticism for clinical assessments.

Specifically in social media, researchers have observed differences between depressed and control groups on Twitter via LIWC: depressed users more frequently use first person pronouns (Pennebaker et al., 2007) and negative emotional (*e.g. angry, frustrated*) words, but show no differences in positive emotion word usage. (Kramer, 2010) mined Facebook to create a happiness index for sentiment analysis. (Brubaker et al., 2012) discovered sentiment features related to grief and distress of MySpace posts. Recently, by analysing posts on Tumblr, (De Choudhury, 2015) discovers two online communities related to anorexia, pro-anorexia and pro-recovery. The two communities display different social and cognitive activities. It is even found out that one group tens to *infiltrate* the other group to spread their opinions.

3 Data

We use three datasets, two from MyPersonality[4] and one from Reddit[5], to investigate the linguistic patterns related to depression and neruoticism.

MyPersonality datasets are composed of the depression dataset and the neuroticism dataset. They are collected Facebook user status posts from those who use a personality test app[6]. The depression dataset provides the CES-D (Radloff, 1977) depression index of each participant, which is the sum of the score from a 20-question survey, ranging from 0 to 60. The higher the CES-D depression index is, the more likely the individual is affected by depression. To maintain a balanced dataset, we designate the one third with highest CES-D scores as positive (depressed), and another one third with the lowest CES-D scores as negative (non-depressed). We name the two sub-datasets $\mathrm{MD_p}$ and $\mathrm{MD_n}$ respectively. The boundary CES-D index of this division is 31. The neuroticism dataset does not have a score that directly reflects depression. We use the neuroticism score to divide this dataset into positive-neurotic and negative-neurotic datasets. We name the two $\mathrm{MN_p}$ and $\mathrm{MN_n}$ respectively.

The Reddit dataset collects posts from the popular anonymous online forum Reddit. We use those from the depression "subreddit" as the positive dataset, namely $\mathrm{RD_p}$. Posts from other subreddits collected in the negative dataset, which we call $\mathrm{RD_n}$.

In the MyPersonality dataset, each user has multiple posts. All the posts from one user is gathered in one file. In the classification stage, we determine whether a user is depressed by extracting features from his or her entire file. In the Reddit dataset, each post is associated with a arbitrary user ID because of the anonymous nature of Reddit. Posts from a single user will display different user IDs. Therefore, we determine whether a post is depressed in classification.

4 Data Linguistic Analysis

We take a look at the languages used in the data set in this section. We mainly analyse the unigrams and bigrams to see if there are any distinct differences between the languages used by people with and without depression.

4.1 Metrics

We use the following metrics to analyse the languages used.

- $\Pr(\cdot)$: the probability of unigrams or bigrams
- PMI: pointwise mutual informtaion of bigrams
- χ^2: χ^2 value of bigrams
- $-2\log\lambda$: value based on log-likelihood of bigrams
- $D(c_1 \,||\, c_2)$: KL-divergences of corpus c_1 and c_2

[4] http://mypersonality.org
[5] https://reddit.com
[6] https://goo.gl/VxKBdW

Detailed descriptions of the metrics can be seen in the following subsections. Here we define some notations:

- w: a word w, also a unigram
- $\neg w$: any word that is not w
- $*$: any word (unigram)
- w_1w_2: a bigram composed of words w_1 and w_2 in this order
- $w_1\neg w_2$: any bigram composed of words w_1 and a word that is not w_2
- $\neg w_1w_2$: any bigram composed of of a word that is not w_1 together with word w_2
- $\neg w_1\neg w_2$: any bigram that is not w_1w_2
- w_1*: any bigram that starts with w_1
- $*w_2$: any bigram that ends with w_2
- c_w: count of occurrences of unigram w in the corpus
- $c_{w_1w_2}$: count of occurrences of bigram w_1w_2 in the corpus
- U: the set of unigrams in the corpus
- B: the set of bigrams in the corpus
- C_U: the total count of all occurrences of unigrams in the corpus
- C_B: the total count of all occurrences of bigrams in the corpus

4.1.1 Probability of Unigrams

The probability of a unigram w is defined as

$$\Pr(w) = \frac{c_w}{C_U} \tag{1}$$

This is the maximum likelihood estimation of the distribution of the unigrams. It indicates how often (frequently) the unigram w shows up in the corpus.

4.1.2 Probability of Bigrams

The frequency of a bigram w_1w_2 is

$$\Pr(w_1w_2) = \frac{c_{w_1w_2}}{C_B} \tag{2}$$

This is the maximum likelihood estimation of the distribution of the bigrams. It indicates how often the bigram w_1w_2 shows up in the corpus. The greater the frequency, the more likely that there is some kind of collocation between w_1 and w_2.

4.1.3 PMI of Bigrams

The pointwise mutual information of a bigram w_1w_2 is

$$\mathrm{PMI}(w_1w_2) = \log_2 \frac{\Pr(w_1w_2)}{\Pr(w_1)\Pr(w_2)} \tag{3}$$

Mutual information can be used to model dependence between words in a bigram.

Assuming w_1 and w_2 are independent,

$$\Pr(w_1w_2) = \Pr(w_1)\Pr(w_2) \tag{4}$$

Then

$$\mathrm{PMI}(w_1w_2) = \log_2 1 = 0 \tag{5}$$

Assuming w_2 is completely dependent on w_1,

$$\Pr(w_1w_2) = \Pr(w_1) \tag{6}$$

Then

$$\mathrm{PMI}(w_1w_2) = \log_2 \frac{1}{\Pr(w_2)} \tag{7}$$

The greater the pointwise mutual information is, the greater the dependence is.

4.1.4 χ^2 of Bigrams

χ^2 value of a bigram w_1w_2 is

$$\chi^2(w_1w_2) = C_B(c_{w_1w_2}c_{\neg w_1\neg w_2} - c_{w_1\neg w_2}c_{\neg w_1w_2}) \frac{1}{(c_{w_1w_2} + c_{w_1\neg w_2})(c_{w_1w_2} + c_{\neg w_1w_2})} \frac{1}{(c_{w_1\neg w_2} + c_{\neg w_1\neg w_2})(c_{\neg w_1w_2} + c_{\neg w_1\neg w_2})} \tag{8}$$

χ^2 value reflects how much we can reject the hypothesis that the words in the bigram are independent. The greater the χ^2 value, the greater confidence with which we can reject the independence hyphothesis.

4.1.5 Log-likelihood of Bigrams

Log likelihood comes from examining two hypotheses. H_1 assumes that the two words in the bigram are independent. H_2 assumes the complete opposite of H_1. The likelihood of the two hypotheses are calculated, the ratio λ is used as the metric to model collocation. Let $L(H)$ be the

likelihood of hypothesis H based on the observation, the log of λ is

$$\begin{aligned}\log \lambda(w_1 w_2) =& \log \frac{L(H_1)}{L(H_2)} \\ =& \log L(c_{w_1 w_2}, c_{w_1 *}, p) \\ &+ \log L(c_{* w_2} - c_{w_1 w_2}, C_B - c_{w_1 *}, p) \\ &- \log L(c_{w_1 w_2}, c_{w_1 *}, p_1) \\ &- \log L(c_{* w_2} - c_{w_1 w_2}, C_B - c_{w_1 *}, p_2)\end{aligned} \tag{9}$$

where

$$p = \frac{c_{* w_2}}{C_B} = \Pr(c_{* w_2}) \tag{10}$$

$$p_1 = \frac{c_{w_1 w_2}}{c_{w_1 *}} = \Pr(c_{w_1 w_2} | c_{w_1 *}) \tag{11}$$

$$p_2 = \frac{c_{* w_2} - c_{w_1 w_2}}{C_B - c_{w_1 *}} = \Pr(\neg w_1 w_2 | \neg w_1 *) \tag{12}$$

and L is a binomial distribution

$$L(k, n, x) = x^k (1 - x)^{n-k} \tag{13}$$

The greater $\log \lambda$, the more probable independence hypothesis is relative to dependence hypothesis. To be uniform with other metrics modelling collocation where greater value means greater probability of collocation, the metric we use (Manning and Schütze, 1999) is $-2 \log \lambda$.

4.1.6 KL-divergence of Unigram Distribution

KL-divergence can model the difference between two distributions. Here, we use KL-divergence to model difference between two corpora's unigram distributions.

$$D(c_1 \,||\, c_2) = \sum_{w \in U} \Pr(w|c_1) \log \frac{\Pr(w|c_1)}{\Pr(w|c_2)} \tag{14}$$

The greater the KL-divergence is, the greater the difference between the unigram distributions of the corpora are.

4.2 Implementation

We divide both the MyPersonality and Reddit corpora further into positive and negative corpora, where positive is composed of posts from self-reported depressed users, and negative from non-depressed or non-depression-related users. Details about the data can be found in the previous section.

We implemented the analysis program in Java. Stopwords are ignored. All punctuations and apostrophes are taken out. A string composed of multiple words that are connected by hyphen(s) is considered one word.

Because data are grabbed from the Internet, there are some gibberish. We output the 1000 unigrams or bigrams with the highest metric values, as well as their metric values for manual comparison. Then we manually compare the output.

We also calculate the KL-divergence between every ordered pair of unigram distributions.

4.3 Insights

For comparison, we use the MyPersonality depression dataset and the Reddit dataset. We present what difference we find in comparing positive and negative corpora. We also show the similarities and differences between any two corpora with KL-divergence.

4.3.1 Depression

Table 1 lists some of the top 100 high-frequency unigrams appeared in the positive dataset that do not have high frequency in the negative dataset. We only select and show those words that we think are related to depression.

crying	awkward	ugly
harder	depressed	mistake
stress	fucked	nervous

Table 1: Depression-related high-frequency unigrams that are only in positive depression corpus

Table 2 presents high-frequency unigrams in the negative dataset that do not have high frequency in the positive dataset. Intuitively, we think that

quiet	jeopardy	terrible
exhausted	low	nor

Table 2: Depression-related high-frequency unigrams that are only in the negative depression corpus

in terms of unigrams, language used in the positive corpus (depressed) is more depression-related, depression-indicative and intense than that used in the negative corpus (non-depressed).

We also collect depression-related bigrams unique in the positive and negative corpora respectively (Table 3 and 4). Here we present the bi-

grams with high $-2\log\lambda$ values. The reason for using this particular metric will be discussed later.

dumb spoiled	mood swings
crying crying	havent slept
im scared	makes cry

Table 3: Depression-related bigrams that has high $-2\log\lambda$ only in positive depression corpus

hate hate	paper due
called ugly	grudges understand

Table 4: Depression-related bigrams that has high $-2\log\lambda$ only in negative depression corpus

Without context, we cannot tell whether bigrams used in the positive corpus are more indicative of depression. However, we have the intuition that the positive (depressed) corpus tends to talk more about the users themselves, with high-frequency bigrams like "cant wait", "im tired", "dont wanna", "im sorry", "im getting", "wish (me) luck" and so on. On the other hand, the negative (non-depressed) corpus tend to talk more about surrounding things, like movies, TV shows, video games, celebrities and so on (eg. "star trek", "harry potter", "jeopardy answer", "sheldon cooper").

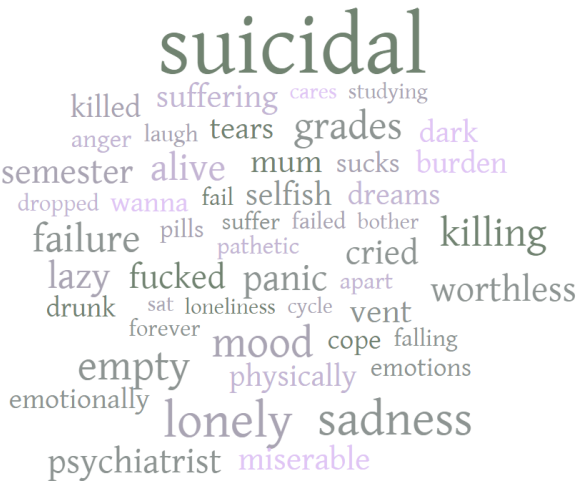

Figure 1: The tag cloud generated with the top 50 most-frequent unigrams from the positive Reddit dataset by PyTagCloud (`https://github.com/atizo/PyTagCloud`). Larger font size indicates greater probability of unigrams.

4.3.2 Reddit

Because of how this corpus is composed, the positive and negative datasets are written by completely different users and cover completely different topics. Figure 1 demonstrates the tag cloud generated with the top 50 most-frequent unigrams from the positive Reddit dataset. Table 5 shows the top 10 high-frequency unigrams that are unique in the positive and negative corpora respectively.

positive	**negative**
suicidal	gt
lonely	cmv
sadness	blood
mood	comment
empty	business
killing	summary
grades	yoga
failure	boat
alive	pulling
panic	thread

Table 5: The top 10 unique unigrams in positive and negative Reddit corpora respectively

We achieve similar results by extracting bigrams from the Reddit corpus. The positive (depressed) dataset is full of bigrams like *"seek help"*, *"cognitive distortions"*, *"roller coaster"*, *"suicide note"* and so on, while the negative (non-depressed) dataset is composed of bigrams of all kinds of topics, like *"blood sugar"*, *"lose weight"*, *"entry level"*, *"lymph node"*, *"credit card"*.

4.4 Discussion

Although we use four metrics to model bigram collocation, we end up relying mainly on the log-likelihood value. The reason is that, because the data is grabbed from the Internet, there are some meaningless bigrams such as *"httpwwwfacebookcomappsapplication-phpid gwiazdek"* and *"uooooooooooooooooo bwhuauhahuhua"*, which only happen once or twice in the entire corpus. There are also misspelled words and what appear to be Spanish or other foreign languages. The individual words in these kinds of bigrams almost never appear in any other bigrams. This tends to blow up the PMI or χ^2 values of these bigrams.

Frequency and log-likelihood perform better in the sense that they both give bigrams that make

relatively more sense. We rely on log-likelihood because it is more robust in measuring collocation and can sift out bigrams like *"oh god"* and *"im getting"* which have high-frequency only because people use such languages a lot.

The KL-divergence of each ordered pair of corpora can be seen in Table 6. It is reasonable the $\mathrm{MN_p}$ and $\mathrm{MN_n}$ are similar in terms of unigram distribution, because they come from the same dataset after all. $\mathrm{MN_p}$ is more similar to $\mathrm{RD_n}$ instead of $\mathrm{RD_p}$, probably because $\mathrm{RD_n}$ is larger than $\mathrm{RD_p}$, and more evenly distributed over the unigram vocabulary. However, it is hard to imagine why $\mathrm{RD_n}$ is more similar to $\mathrm{MN_p}$ and $\mathrm{MN_n}$ than $\mathrm{RD_p}$.

$D(c_{row} \mid\mid c_{col})$	$\mathrm{MN_p}$	$\mathrm{MN_n}$	$\mathrm{Rd_p}$	$\mathrm{Rd_n}$
$\mathrm{MN_p}$		1.249	3.366	2.269
$\mathrm{MN_n}$	1.678		3.714	2.471
$\mathrm{Rd_p}$	1.367	1.321		0.734
$\mathrm{Rd_n}$	1.863	1.680	2.027	

Table 6: The KL-divergence between all ordered pairs of corpora.

5 Classification

In this section, we apply binary classification on all corpora, so that depressed/neurotic users can be automatically distinguished. In order to improve the classification results, we explore and compare multiple feature selection and classification algorithms.

5.1 Dataset Details

We choose positive and negative examples using the methods in Section 3. Table 7 presents details of datasets we use. The value range denotes the range of CES-D score and Neuroticism score in *Depression in MyPersonality* and *Neuroticism in MyPersonality* respectively.

Dataset	#Examples	Value Range
$\mathrm{MD_p}$	313	31 – 48
$\mathrm{MD_n}$	313	10 – 25
$\mathrm{MN_p}$	4250	3.00 – 5.00
$\mathrm{MN_n}$	4250	1.00 – 2.25
$\mathrm{RD_p}$	1603	N/A
$\mathrm{RD_n}$	18318	N/A

Table 7: Dataset Details

5.2 Preprocessing

We employ OpenNLP (Baldridge, 2005) to preprocess the corpora. The first step is tokenization, in which the last word of each sentence is separated with punctuations and the same words are aggregated. The second step is stemming. English words have different forms under different tenses (*e.g. go, went and gone*). By stemming these words, each word's spelling is unified. We also applied part-of-speech (POS) tagging, so that we can select words with certain POS for classification. Another important step is to remove stopwords like *"a"*, *"the"* and *"of"*. These words appear frequently but barely have any meanings. Removing these words significantly benefit the classification task, as we will show later.

5.3 Feature Selection

We investigate three feature selection algorithms and their combinations. First, we employ χ^2-test to select words which are highly correlated with depression / neuroticism class and of high frequency. In order to select correlated words with low frequency, we adopt PMI (Equation 3) for this task. As PMI focuses on selecting low frequency words, the word set it returns usually contains many words in the "long tail" part of word distribution. In addition to word-level algorithms above, we also employ LDA to extract topic-level features of documents. With the help of LDA, each document's topic distribution is estimated and used as additional features besides lexicons.

5.4 Classification Algorithms

Classification algorithm plays a crucial rold in this task. First, we employ Support Vector Machine (SVM) for classification. SVM directly deals with features and determines a hyperplane to separate positive and negative data points. When the data is not linearly separable, it projects data points to high dimensional space using kernel functions and then tries to separate them linearly. The advantage is that it makes full use of all the features to decide the hyperplane and runs rather fast due to kernel methods, in spite of dealing with high dimensional data. However, SVM is not capable of extracting additional features from the data. Therefore, its performance may be limited.

In order to capture more features in classification, we also employ Supervised LDA (SLDA) (Mcauliffe and Blei, 2008) to classify

documents. The generative process of SLDA is similar to LDA. The only difference is that SLDA draws a response value for each document. The response value is drawn from a Gaussian distribution where the mean equals to a linear combination of document's topic proportions. In this way, SLDA extracts topic information from documents, as well as based on documents' labels. In our implementation of SLDA, the linear weights are given a standard Gaussian prior and optimized using L-BFGS algorithm (Liu and Nocedal, 1989) in Mallet package (McCallum, 2002).

Model	Feature	Precision	Recall
SVM	Baseline	0.5772	0.1496
	Stopwords	0.6350	0.2328
	χ^2	0.6186	0.2333
	PMI	0.6396	0.2296
	LDA	0.6350	0.2328
	χ^2+PMI	0.6433	0.2296
	χ^2+LDA	0.6222	0.2365
	PMI+LDA	0.6433	0.2296
	χ^2+PMI+LDA	0.6433	0.2296
SLDA	Baseline	0.5813	0.6230
	Stopwords	0.5704	0.5782
	χ^2	0.5328	0.6008
	PMI	0.5560	0.5142
	χ^2+PMI	0.5643	0.4917

Table 8: The average precision and recall of different predictive models for *Depression in MyPersonality* based on 5-fold cross validation.

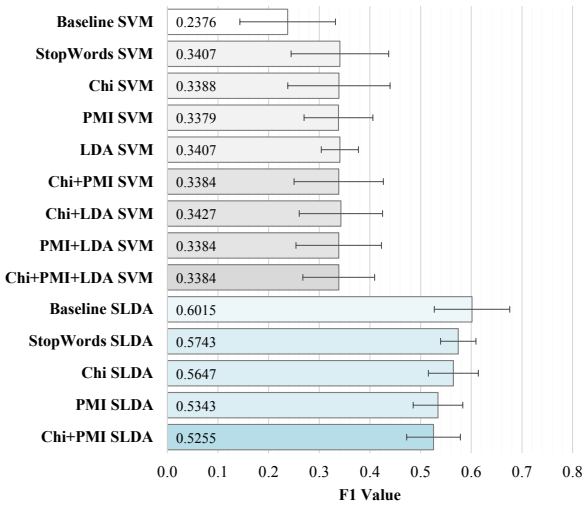

Figure 2: The average F1 values of different predictive models for *Depression in MyPersonality*. Gray and blue colors indicate performance of SVM and SLDA respectively. The darker color indicates more feature selection algorithms.

6 Evaluation

As the most basic supervised learning paradigm, we treat each dataset (*i.e. Depression in MyPersonality, Neuroticism in MyPersonality and Depression in Reddit*) as a separate task. We evaluate the classification performance with precision, recall and F1 values.

Model	Feature	Precision	Recall
SVM	Baseline	0.6994	0.3027
	Stopword	0.7034	0.3981
	χ^2	0.6969	0.3835
	PMI	0.7045	0.3986
	LDA	0.7027	0.3979
	χ^2+PMI	0.7048	0.3986
	χ^2+LDA	0.6970	0.3843
	PMI+LDA	0.7044	0.3984
	χ^2+PMI+LDA	0.7040	0.3993
SLDA	Baseline	0.5995	0.6078
	StopWords	0.5868	0.6480
	χ^2	0.5825	0.6636
	PMI	0.5892	0.6421
	χ^2+PMI	0.5940	0.6426

Table 9: The average precision and recall of different models for *Neuroticism in MyPersonality*.

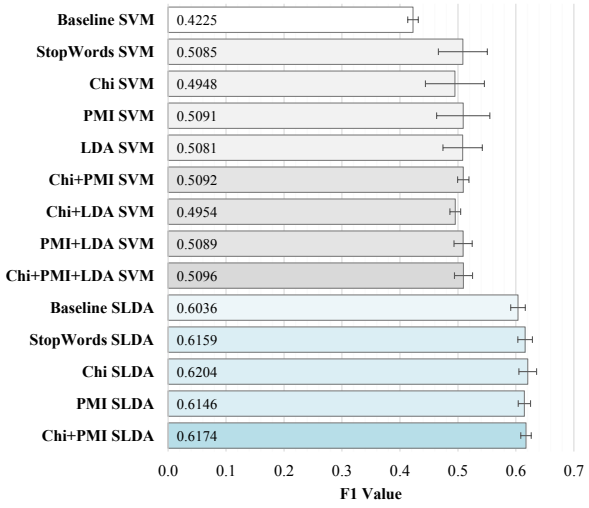

Figure 3: The average F1 values of different predictive models for *Neuroticism in MyPersonality*. The color coding strategy is the same as Figure 2.

6.1 Result

The results of precision and recall for MyPersonality and Reddit corpora in Table 8, 9 and 10. Besides, Figure 2, 3 and 4 show F1 values on three datasets with error bars (standard deviation).

Note that in each task, only the "Baseline" setting contains stopwords. The thresholds of χ^2-test and PMI are selected by iterations, when the performance reaches the best. When applying χ^2-test and PMI together, a word is selected as feature as long as either its χ^2-test value or PMI value is above the corresponding threshold. While running LDA and SLDA, we set the number of topics to 10 for MyPersonality corpora and 33 for Reddit corpus (because Reddit corpus comes from 33 subreddits). All the results are based on 5-fold cross validation.

Model	Feature	Precision	Recall
SVM	Baseline	0.8092	0.3812
	Stopword	0.8045	0.5047
	χ^2	0.8091	0.5065
	PMI	0.7959	0.4847
	LDA	0.8185	0.5365
	χ^2+PMI	0.8037	0.5047
	χ^2+LDA	0.8114	0.5334
	PMI+LDA	0.7978	0.4947
	χ^2+PMI+LDA	0.8118	0.5396
SLDA	Baseline	0.6753	0.3069
	StopWords	0.6990	0.3956
	χ^2	0.7074	0.3338
	PMI	0.6722	0.3095
	χ^2+PMI	0.6872	0.4044

Table 10: The average precision and recall of different models for *Depression in Reddit*.

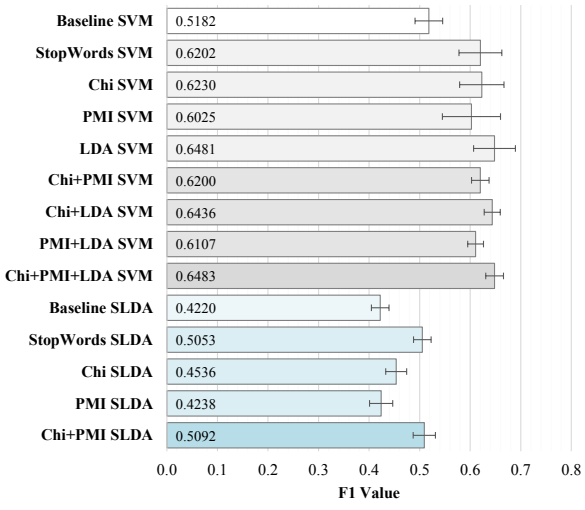

Figure 4: The average F1 values of different predictive models for *Depression in Reddit*. The color coding strategy is the same as Figure 2.

6.2 Discussion

We discuss the differences regarding using stopwords or not, among χ^2-Test, PMI, LDA, between single and multiple feature selection, and between SVM and SLDA.

6.2.1 Keeping Stopwords vs Removing Stopwords

It turns out that by removing stopwords, the classification performance is improved, except for SLDA on the *Depression in MyPersonality* corpus. Stopwords barely have any meanings, but usually have high frequencies in documents. However, classification algorithms can not identify these words and will assign these features with weights which are difficult to tune because the features are almost irrelevant to classes.

Although removing stopwords generally improves classification performance, the improvements for SVM and SLDA are different. It appears that SVM (averagely improved 0.0970 in F1 value) benefits more than SLDA (averagely improved 0.0146 in F1 value). This is probably because when SLDA is sampling topic proportions, it extracts principal components and discards irrelevant components, which makes the room for improvement smaller.

6.2.2 χ^2-Test vs PMI vs LDA

By comparing the performance of χ^2-test and PMI, we find that χ^2 outperforms PMI by 0.1218 in F1 score averagely over corpora and classification algorithms. We assume that χ^2-test and PMI are good at selecting highly correlated words in high and low frequency respectively. It turns out that high frequency words matter more for classification task.

Nevertheless, both feature selection algorithms fail to bring improvement on F1 value, compared to removing stopwords only. Even though, χ^2-test and PMI are still useful, because they significantly accelerate the classification process by removing irrelevant features. χ^2-Test only keeps around 1% of original vocabulary while PMI keeps approximately 50%.

LDA topic vectors, on the other hand, are extra features for documents. It serves as a summary of documents and appears to work better (0.0917) than χ^2-test (-0.0427) and PMI (-0.0663) in improving F1 value. The disadvantage of topic vectors is also obvious – it takes too much time to get them, as we usually need to run hundreds, even

thousands of iterations in order to get high-quality topic vectors.

6.2.3 Single Feature Selection vs Multiple Feature Selections

We expect the performance to be better when multiple feature selection algorithms are applied, compared to the situation that only a single one is applied. However, it is only true on the *Depression in Reddit* dataset, not on MyPersonality corpora. It indicates that multiple feature selections does not necessarily outperform single feature selection. The performance is also related to the corpus we use.

On the *MyPersonality* corpora, although the performance is not improved when multiple feature selection algorithms are applied, it does help raise the F1 value compared to the one that works worse. For example, when we apply χ^2-test and PMI on the *Depression in MyPersonality* corpus, the F1 value is 0.6200. It is better than the result where only PMI is used (0.6025). This phenomenon suggests that if we employ multiple feature selection algorithms, we can avoid disadvantages of a single algorithm, which makes the aggregated algorithm more robust.

6.2.4 SVM vs SLDA

As we can see from the tables above, SLDA outperforms SVM on *MyPersonality* corpora, while SVM performs better on the *Reddit* corpus. By looking at corpora, we find that the average length of documents in *MyPersonality* corpora is much longer than that of *Reddit* documents. We can probably assume that SLDA is more sensitive to length of documents: if a document has more words, SLDA can better estimate the topic proportions of that document and therefore achieve a better classification performance. SVM, on the contrary, is less affected by the number of non-zero features. It only needs little information to compute the hyperplane and support vectors.

Although SLDA deals with long documents better than SVM, its disadvantage is not negligible: SLDA takes too much time to converge when documents are long. When we run SLDA for baseline on *Nueroticism in MyPersonality* corpus, it takes about 5 hours to finish the cross validation (each fold has 300 iterations), while SVM takes only several minutes on the same corpus. Hence, the running time is a crucial factor when deciding which algorithm to use on long documents.

7 Conclusion and Future Work

In this technical report, we present our work in learning depression patterns from Social Media using three datasets from *MyPersonality* and *Reddit*. Our accomplishments are as follows:

1. **Exploratory Data Analysis** We accomplish our first goal by extracting unigram and bigram models and compare them based on frequency, PMI, χ^2-test and log-likelihood. We also show visualization via tag cloud.

2. **Classification** We successfully train predictive models using SVM and conduct five-fold cross-validation. We report average precision, recall and F1-score on the datasets.

3. **Bonus Credits** Beyond our proposed goals, we employ SLDA for the binary classification. Additionally, we thoroughly compare and discuss the differences regarding using stopwords or not, among χ^2-Test, PMI, LDA, between single and multiple feature selection, and between SVM and SLDA. We also consider neuroticism beyond our proposal.

We also propose the following directions for future research:

1. **Temporal Analysis** It is of great interest for us to explore how depression or neuroticism level changes from time to time or before and after clinical treatment.

2. **Social Relevance** The interaction with friends, families and even strange people can have great impact in regulating negative emotions in social media. By taking comments, forward, replies, likes into account, future researchers may gain more insights in predicting and monitoring mental health problems

3. **Multimedia Features** Beyond linguistic features, current social media has more and more modalities including images, videos and geo-location information. Thus we believe that linguistic researchers could benefit from such features by collaborating with computer vision and geo-spatial experts.

Acknowledgments

The authors greatly appreciate Prof. Philip Resnik and Raul Guerra for teaching CMSC 773 and advising the team for this enjoyable project.

Credits

We had a very active group, every member contributes to the project extensively. We organized weekly group meeting every Wednesday night from 5pm to 9pm since the very beginning of April to 18 May, 2015. For the final write-up, we spent an entire weekend on it.

Specifically, *Weiwei* mainly works on the feature selection, classification algorithms and experiment conduction. *Xuetong* mainly works on the exploratory data analysis, linguistic metrics and initial insights. *Ruofei* mainly works on the tag cloud visualization, figures, tables, as well as surveys and write-up. We collaborated and helped each other whenever problems occurred.

References

[American Psychiatric Association2013] American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders*. 5 edition.

[American Psychological Association2002] American Psychological Association. 2002. The Critical Need for Psychologists in Rural America. *Retrieved March*, 29:2005.

[Baldridge2005] Jason Baldridge. 2005. The OpenNLP Project. *URL: `http://opennlp.apache.org/index.html`, (accessed on 18 April 2015)*.

[Beck1998] Cheryl Tatano Beck. 1998. A Checklist to Identify Women at Risk for Developing Postpartum Depression. *Journal of Obstetric, Gynecologic, & Neonatal Nursing*, 27(1):39–46.

[Blei et al.2003] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, 3:993–1022.

[Brubaker et al.2012] Jed R Brubaker, Funda Kivran-Swaine, Lee Taber, and Gillian R Hayes. 2012. Grief-stricken in a crowd: The language of bereavement and distress in social media. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media (ICWSM)*.

[Centers for Disease Control and Prevention2001] Centers for Disease Control and Prevention. 2001. Behavioral Risk Factor Surveillance System Survey Questionnaire. *Atlanta, Georgia: US Department of Health and Human Services, Centers for Disease Control and Prevention*, pages 22–23.

[Clark and Drake1994] Robin E Clark and Robert E Drake. 1994. Expenditures of Time and Money by Families of People with Severe Mental Illness and Substance Use Disorders. *Community Mental Health Journal*, 30(2):145–163.

[Coppersmith et al.2014] Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying Mental Health Signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60. Association for Computational Linguistics.

[Costa and MacCrae1992] Paul T Costa and Robert R MacCrae. 1992. *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO FFI): Professional Manual*. Psychological Assessment Resources.

[De Choudhury et al.2013a] Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013a. Major Life Changes and Behavioral Markers in Social Media: Case of Childbirth. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW)*, pages 1431–1442. ACM.

[De Choudhury et al.2013b] Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013b. Predicting Postpartum Changes in Emotion and Behavior via Social Media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*, pages 3267–3276. ACM.

[De Choudhury2013] Munmun De Choudhury. 2013. Role of Social Media in Tackling Challenges in Mental Health. In *Proceedings of the 2nd International Workshop on Socially-Aware Multimedia (SAM'13)*, pages 49–52.

[De Choudhury2015] Munmun De Choudhury. 2015. Anorexia on Tumblr : A Characterization Study Studies on Anorexia. In *Proceedings of DH'15: 5th ACM Digital Health Conference. DH'15.*

[Golder and Macy2011] Scott A Golder and Michael W Macy. 2011. Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures. *Science*, 333(6051):1878–1881.

[Greene and Resnik2010] Stephan Greene and Philip Resnik. 2010. More than words: syntactic packaging and implicit sentiment. In *NAACL '09 Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 503–511.

[Gross and Muñoz1995] James J Gross and Ricardo F Muñoz. 1995. Emotion Regulation and Mental Health. *Clinical Psychology: Science and Practice*, 2(2):151–164.

[John and Srivastava1999] Op P John and S Srivastava. 1999. The Big Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives. *Handbook of personality: Theory and research*, 2(510):102–138.

[Kaplan and Haenlein2010] Andreas M. Kaplan and Michael Haenlein. 2010. Users of the World, Unite! The Challenges and Opportunities of Social Media. *Business Horizons*, 53(1):59–68, January.

[Kessler and Üstün2004] Ronald C Kessler and T Bedirhan Üstün. 2004. The World Mental Health (WMH) Survey Initiative Version of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI). *International Journal of Methods in Psychiatric Research*, 13(2):93–121.

[Kramer2010] Adam DI Kramer. 2010. An Unobtrusive Behavioral Model of Gross National Happiness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 287–290. ACM.

[Liu and Nocedal1989] Dong C Liu and Jorge Nocedal. 1989. On the Limited Memory BFGS Method for Large Scale Optimization. *Mathematical Programming*, 45(1-3):503–528.

[Manning and Schütze1999] Christopher D Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT press.

[Mcauliffe and Blei2008] Jon D Mcauliffe and David M Blei. 2008. Supervised Topic Models. In *Advances in neural information processing systems*, pages 121–128.

[McCallum2002] Andrew K McCallum. 2002. {MALLET: A Machine Learning for Language Toolkit}.

[Miller2002] Laura J Miller. 2002. Postpartum Depression. *Journal of American Medicine Association (JAMA)*, 287(6):762–765.

[Park et al.2012] Minsu Park, Chiyoung Cha, and Meeyoung Cha. 2012. Depressive Moods of Users Portrayed in Twitter. In *Proceedings of the ACM SIGKDD Workshop on Healthcare Informatics (HI-KDD)*, pages 1–8.

[Pennebaker and King1999] James W Pennebaker and Laura A King. 1999. Linguistic Styles: Language Use as an Individual Difference. *Journal of Personality and Social Psychology*, 77(6):1296.

[Pennebaker et al.2003] James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. 2003. Psychological Aspects of Natural Language Use: Our Words, Our Selves. *Annual Review of Psychology*, 54(1):547–577.

[Pennebaker et al.2007] James W Pennebaker, Cindy K Chung, Molly Ireland, Amy Gonzales, and Roger J Booth. 2007. The Development and Psychometric Properties of LIWC2007.

[Pestian et al.2012] John Pestian, John Pestian, Pawel Matykiewicz, Brett South, Ozlem Uzuner, and John Hurdle. 2012. Sentiment Analysis of Suicide Notes: A Shared Task. *Biomedical Informatics Insights*, 5:3.

[Radloff1977] Lenore Sawyer Radloff. 1977. The CES-D Scale a Self-Report Depression Scale for Research in the General Population. *Applied Psychological Measurement*, 1(3):385–401.

[Resnik et al.2013] Philip Resnik, Anderson Garron, and Rebecca Resnik. 2013. Using Topic Modeling to Improve Prediction of Neuroticism and Depression in College Students. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, number October, pages 1348–1353.

[Schwartz and Eichstaedt2013] Ha Schwartz and Jc Eichstaedt. 2013. Characterizing Geographic Variation in Well-Being Using Tweets. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 583–591.

[U.S. Department of Health and Human Services2009] U.S. Department of Health and Human Services. 2009. *National Expenditures for Mental Health Services and Substance Abuse Treatment, 1986-2009*. US Department of Health and Human Services, Substance Abuse and Mental Health Services Administration.

[Wingate et al.2014] Martha Wingate, Russell S Kirby, Sydney Pettygrove, Chris Cunniff, Eldon Schulz, Tista Ghosh, Cordelia Robinson, Li-Ching Lee, Rebecca Landa, John Constantino, et al. 2014. Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years-Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2010. *MMWR Surveillance Summaries*, 63(2).