MONITORING TWEETS FOR DEPRESSION TO DETECT AT-RISK USERS

BY

ZUNAIRA JAMIL

Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements
for the M.Sc. in Computer Science

School of Electrical Engineering and Computer Science
Faculty of Engineering
University of Ottawa

ABSTRACT

According to the World Health Organization, mental health is an integral part of health and well-being. Mental illness can affect anyone, rich or poor, male or female. One such example of mental illness is depression. In Canada 5.3% of the population had presented a depressive episode in the past 12 months. Depression is difficult to diagnose, resulting in high under-diagnosis. Diagnosing depression is often based on self-reported experiences, behaviors reported by relatives, and a mental status examination. Currently, authorities use surveys and questionnaires to identify individuals who may be at risk of depression. This process is time-consuming and costly.

We propose an automated system that can identify at-risk users from their public social media activity. More specifically, we identify at-risk users from Twitter. To achieve this goal we trained a user-level classifier using Support Vector Machine (SVM) that can detect at-risk users with a recall of 0.8750 and a precision of 0.7778.

We also trained a tweet-level classifier that predicts if a tweet indicates distress. This task was much more difficult due to the imbalanced data. In the dataset that we labeled, we came across 5% distress tweets and 95% non-distress tweets. To handle this class imbalance, we used undersampling methods. The resulting classifier uses SVM and performs with a recall of 0.8020 and a precision of 0.1237.

Our system can be used by authorities to find a focused group of at-risk users. It is not a platform for labeling an individual as a patient with depression, but only a platform for raising an alarm so that the relevant authorities could take necessary interventions to further analyze the predicted user to confirm his/her state of mental health. We respect the ethical boundaries relating to the use of social media data and therefore do not use any user identification information in our research.

To my parents, for their love and support.

ACKNOWLEDGMENTS

First and foremost, I offer my sincerest and deepest gratitude to my supervisor Dr. Diana Inkpen for her eminently capable supervision, fantastic feedback, creative ideas, gracious patience, and endless support. I feel truly blessed to have had such a hard-working mentor.

Many thanks to Dr. Kenton White for providing us with the data for this research, helping us with the development challenges and for sharing insights into machine learning and social media.

Many thanks to my research team Prasadith Buddhitha, Saurav Dindyal, and Ruba Skaik for the discussions, sharing ideas, and helping me with experiments.

A special thanks to the annotators Bryan Paget, Sameen Salim, and Nimrah Rashid who spent hours and hours labeling the data for this research.

Thanks to my committee members, Dr. Yuhong Guo, Dr. Marina Sokolova, and Dr. Kenton White, for providing feedback to help me improve this manuscript and make my results more presentable.

Thanks to all participants of TAMALE seminars and NLP-chats for sharing insights into cutting-edge NLP research. Thank you to all members of NLP group namely Haifa Alharthy, Ehsan Amjadian, Vaibhav Kesarwani, Romualdo Alves, Parinaz Sobhani, Arya Rahgozar, and Hanqing Zhou for the support and encouragement. Thank you to Dr. Stan Szpakowicz, Dr. Amal Zouaq, Dr. Chris Tanasescu, and Dr. Marina Sokolova for the regular discussions and providing insights into different areas of NLP. Thank you to Dr. Xiaodan Zhu, and Dr. Saif Mohammad from NRC for introducing us to cutting edge research. Thank you again Professor Diana Inkpen for organizing TAMALE seminars and inviting esteemed guest speakers.

Thanks to Dr. Kathleen Pajer (CHEO, Department of Psychiatry), and Dr. Bill Gardner (CHEO, Department of Psychiatry), for explaining the challenges and risks associated with mental illness.

Thank you to our collaborators Dr. Sandra Bringay (Universite de Montpellier, France) for providing us with keyword/phrases associated with mental illness. Thank you Mike Donald for performing preliminary experiments on French data affirming that the developed system can easily be adapted to French tweets.

Thank you to the Natural Sciences and Engineering Research Council of Canada (NSERC) for providing funding for this research.

Thank you to the staff at Morisset library for maintaining a robust set of licenses to NLP research, without which this work would not have been possible.

Finally, I thank to my family and friends for their support, understanding, patience and encouragement throughout my research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AUC	Area Under the Curve
CLPsych	Clinical Psychology
FN	False Negative
FP	False Positive
LDA	Linear Discriminant Analysis
LIWC	Linguistic Inquiry and Word Count
LSA	Latent Semantic Analysis
ML	Machine Learning
NB	Naive Bayes
NLP	Natural Language Processing
POS	Part of Speech
PTSD	Post Traumatic Stress Disorder
RF	Random Forest
ROC curve	Receiver Operator Characteristics Curve
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machine
TF-IDF	Term Frequency- Inverse Document Frequency
TN	True Negative
TP	True Positive
WHO	World Health Organization

CHAPTER 1

INTRODUCTION

This dissertation describes a set of experiments performed on public tweets, to identify users who suffer from depression or are at-risk of depression, using text mining techniques. We also attempt to identify distressed tweets in a Twitter stream.

1.1 Motivation

According to a recent report of the World Health Organization (WHO), mental health is an integral part of health and well-being (WHO, 2004). Mental disorders can affect anyone, rich or poor, male or female, of any age or social group. The experience of mental illness is often described as difficult, especially when associated with demeaning prejudices and lack of understanding. Mental illness is also difficult to diagnose. There is no reliable laboratory test for most forms of mental illness and typically, diagnostic is based on the patient's self-reported experiences, behaviors reported by relatives, and a mental status examination. Unfortunately, mental disorder problems are increasing worldwide also because of population aging.

In the context of mental illness, depression is very common. In Canada, 5.3% of the population had presented a depressive episode in the past 12 months, while in France the percentage was 7.8%. The goal of our project is to exploit the massive data issued from social media and apply social media mining and sentiment analysis methods to detect at-risk people.

1.2 Thesis Goals

The goal of this thesis is to apply Natural Language Processing and Machine Learning techniques to build a system that given a set of tweets from a user

can identify at-risk tweets and hence at-risk users. For this task, we need to identify useful textual or community features. This system needs to take into account knowledge of social media posts, such as 1) tweets are short and may convey less emotion; 2) users tweet about a wide variety of topics; 3) the language does not conform to grammatical structure and may contain spelling errors/shorthand notation to fit into 140 characters.

This leads to a secondary goal of the thesis, which is to identify relevant tweets for analysis from the large amount of Twitter data.

1.3 Intended Contributions

The intended contributions of this thesis are as follows:

1. We introduce a new dataset based on users who participated in the BellLetsTalk campaign. This dataset is annotated at tweet-level and user-level. Our user-level annotations take distressed tweets into account and do not rely solely on user's self-disclosure.

2. We attempt to develop a tweet-level classification model that predicts if a tweet indicates distress. This is a difficult problem due to the large amount of off-topic tweets and a small amount of relevant tweets. For this we experiment with various sampling methods.

3. For the user-level classification model, we perform extensive experiments using text-derived features.

4. We attempt to develop a user-level classifier that takes distressed tweets into account.

1.4 Outline

The remainder of this thesis is organized as follows: In Chapter 2, we review the fundamentals of Natural Language Processing and Machine Learning focusing on concepts used in this thesis. In Chapter 3, we provide a literature review of studies related to mental illness in social media, including an overview of methodologies used by existing systems. Chapter 4 provides a

description of the dataset, and how the data was annotated. This is followed by a detailed description of experiments performed at tweet-level and user-level in Chapter 5. The results are presented and discussed in Chapter 6, followed by conclusion and ideas for future work in Chapter 7.

CHAPTER 2

BACKGROUND

2.1 Natural Language Processing and Machine Learning

This research makes use of Natural Language Processing (NLP) and Machine Learning (ML) techniques to extract sentiment from tweets. This section provides a brief overview of NLP, ML, and their common applications.

2.1.1 Brief introduction to Natural Language Processing

Natural Language processing (NLP) is the study of human language interaction with a computer. It allows a machine to understand natural language. It is part of Computer Science, in particular Artificial Intelligence. NLP techniques allow computers to analyze, understand and derive meaning from human language. The first task attributed to NLP was the translation from Russian to English back in 1950s. Since then, significant advancements have been made in the field. Today, by using NLP, developers can organize and structure knowledge to perform tasks such as automatic summarization, translation, named entity recognition, relationship extraction, sentiment analysis, and topic segmentation with a reasonable accuracy (Farzindar and Inkpen, 2015). Understanding human language is a difficult task because to understand language requires more than just understanding words. It requires understanding of how words are linked to create meaning. In many cases, a word-to-word or literal translation is not accurate. Many words have multiple meanings, which adds an additional challenge to understanding language.

NLP applications are widely used today, even when people are not aware of its use. Some of the most commonly used applications are machine transla-

tion, spam filters, autocomplete suggestions (on phones and search engines), etc.

Some other applications that not everyone uses include text summarization apps and sentiment analysis apps (that allow us to find out how people feel about a product, service or person). Text classification techniques are used for classifying SMS/emails as spam or not, or to classify books into genres. Conversational agents, like Eliza, or more recent, Microsoft's Tay bot are a few other of the many applications of NLP.

These applications fall into one or more of the following current research direction in NLP: [1]

- Sentiment Analysis: deriving sentiment from text, e.g., positive, negative, neutral, or even emotions, such as happiness, sadness, anger, disgust, fear.
- Text Summarization: summarizing a single article or multiple articles based on a single theme.
- Textual Entailment: deriving directional relations between text fragments in order to associate multiple word phrases/words with same meaning.
- Information extraction: finding structured information from unstructured data, like detecting entities and relations between them, and solving co-references.
- Topic Segmentation: extracting topics from text. The topics could overlap.
- Question Answering: answering closed (specific) and open (subjective) questions. These are the basis for virtual assistants, like iPhone's Siri.
- Part-of-Speech Tagging: tagging words with parts-of-speech such as nouns, verbs, adjectives, etc.
- Translation: translating from one language into another.

[1] https://www.quora.com/What-are-the-current-hot-topics-in-natural-language-processing

The aim of developing NLP techniques is to use the vast amounts of data available online (and offline) and extract useful information without human intervention. Human intervention is expensive and time consuming. Computers can complete the same tasks much faster, but have yet to achieve the same accuracy. There are two major schools of thought in NLP.

1. Rule based systems: Applications are designed as sets of rules to guide a system. An example is building a sentence parser using a nominally complete set of rules that define allowable words, their parts of speech, and allowable sequences of parts of speech; another example is building a machine translation system using rules-based finite state transducers that directly map input words to output words, perhaps with some rules for reordering words in the process. The problem with this approach is that it is a difficult task to model natural language by a finite set of rules (even when ignoring the disagreements among linguists). Additionally, these rules are based on a well-defined vocabulary and any new words may not be part of this vocabulary (this would prove to be a challenge with the evolving language of social media).

2. Statistical NLP: consists of quantitative (often probabilistic) approaches to dealing with language, modeling language implicitly (by counting words or short sequences of words or by using large sets of aligned parallel text to accomplish machine translation) rather than using explicit rules. These too can be subject to criticism, as the statistical assumptions that underlie these techniques may not match our intuition of how language works (or should work), and corpus based applications can be criticized for having insufficient data.

In this research, we use statistical NLP approaches, in particular NLP methods based on machine learning algorithms. Instead of hand-coding rules, NLP can rely on machine learning techniques to learn these rules by analyzing a set of examples and making statistical inferences.

2.1.2 Brief introduction to Machine Learning

Machine Learning (ML) is a subfield of Artificial Intelligence. It is best described by Arthur Samuel as **"[Machine learning is the] field of study**

that gives computers the ability to learn without being explicitly programmed." - (1959).

This means that computers can be taught to learn from examples. ML seeks to make decisions about unseen data without being explicitly instructed on how to do so. There are three main divisions of machine learning tasks, namely "supervised learning", "unsupervised learning" and "semi-supervised learning".

- Supervised machine learning: a model is "trained" on a pre-defined set of labeled "training examples"; then the model can be applied on new data in order to make predictions. Supervised learning problems are also referred to as classification problems.

- Unsupervised machine learning: a model is built on unlabeled data by finding patterns and relationships in the data. Unsupervised learning problems are also referred to as clustering problems.

- Semi-supervised machine learning: a model is trained on labeled examples and then unlabeled examples are used to refine the boundaries between classes. (Han et al., 2011)

Real life applications of Machine Learning include credit card fraud detection, character recognition, speech understanding, face detection, product recommendation, medical diagnosis, customer segmentation, shape detection, and sign language interpretation. [2]

Machine Learning works by finding patterns in data (associated or not with given classes). In all of the above applications, data is first converted to a representation (set of features) that can be understood by a computer. In NLP, the text has to be transformed into a numeric (or discrete) representation in order to be able to apply Machine Learning algorithms. Similarly, Computer Vision deals with images and converts them in a form (representation) that can be understood by a computer.

A typical machine learning application consists of Data collection and preprocessing, feature engineering, training a model and testing performance.

[2] http://machinelearningmastery.com/practical-machine-learning-problems/

2.1.3 Feature engineering

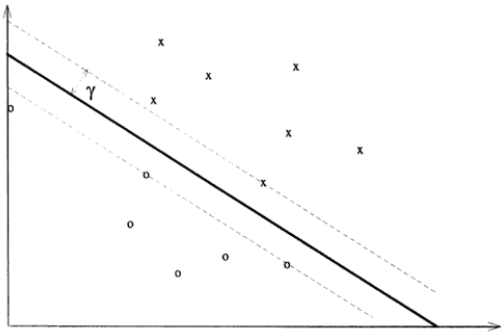
In Machine Learning, feature engineering is referred to as the process of using domain knowledge of the data to create features that can be used by machine learning algorithms to find patterns. A feature is a piece of information that might be useful for prediction. Attribute, variable are some other terms used synonymous to features, when defining a Machine Learning task.
A dataset given to ML algorithm consists of dependent and independent variables. Dependent variable is usually a nominal value i.e., the target outcome of a prediction. A variable is nominal when it has fixed number of categories. Independent variables on the other hand can be of type "numeric" (such as integer or real) or "nominal". A few Machine Learning algorithms may also accept variables of type "string", such as "NaiveBayesMultinomialText" classifier in WEKA. Features can be provided as part of the dataset by an expert or derived from data in case of text data. Next step is to find a suitable classifier for the task.

2.1.4 Classifier selection

There is a large variety of Machine Learning algorithms commonly used today. Some of the widely used algorithms in Natural Language processing tasks are Support Vector Machine (SVM), Naive Bayes (NB), Random Forest (or other tree-based algorithms), and ensemble methods. There is no one algorithm that is suited for all tasks. Researchers usually try a variety of algorithms and optimize them for problem of their interest.

For any Machine Learning task, the research involves finding useful features. Classification algorithms then find patterns within these features. In case of NLP tasks, these features are often extracted from text. Bags of Words and n-grams (section 2.2) are examples of features that rely on word frequencies. Other features are more problem specific, such as the sentiment value of a document, its tone, readability level, etc. The purpose of generating features is to extract information from text and translate it into a representation that can be understood by a machine learning algorithm.

Figure 2.1: Svm maximum separating hyperplane



Support Vector Machines

Support Vector Machines (SVM) is a classification algorithm. Given a set of labeled training examples for a binary class problem, SVM's training algorithm builds a model that finds an optimum hyperplane separating examples from the two classes. It maximizes the margin, or the distance between the separating hyperplane and the training examples nearest to the hyperplane. See Figure 2.1 (Cristianini and Shawe-Taylor, 2000).

When given a test sample, SVM makes a prediction based on which side of the hyperplane the sample falls on.

Naive Bayes classifiers

The Naive Bayes (NB) algorithm is widely used in Machine Learning due to its efficiency and ability to combine evidence from large number of features (Manning and Schütze, 1999). Naive Bayes works with the assumption that all attributes (features) are conditionally independent. Conditional probability is the probability that an event will occur given that something else has already happened. Often, we know how frequently some evidence is observed, given a known outcome. We have to use the known fact to compute the chance of that outcome happening, given the evidence. This can be done using Bayes Rule.

Here is an example of Bayes Rule:

$$\textit{Prob-of-a-disease-D-given-Test-positive} = \frac{Prob(TestIsPositive|Disease)*P(Disease)}{Prob(TestingPositive,WithOrWithoutTheDisease)}$$

$$P(outcome|evidence) = \frac{P(LikelihoodOfEvidence)*PriorProbOfOutcome}{P(evidence)}$$

During training, most of the parameters required for the function are computed and stored in the model. In order to make predictions on a test sam-

ple, we can just run the formula above for each possible outcome. The model classifies the sample as belonging to the class with highest probability [3]

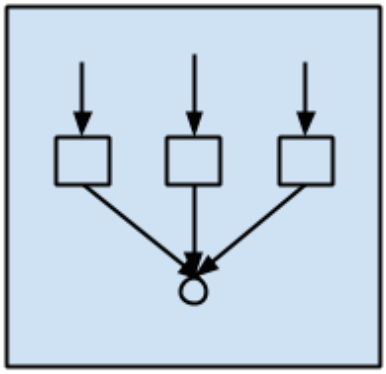
Tree-based classifiers

Decision trees, Random Forest, J48 are all examples of Tree-based algorithms. Webb (2003) describes decision trees as being capable of modeling complex nonlinear decision boundaries. Overly large trees are constructed and then pruned to minimize a cost-complexity criterion. The resulting tree is easily interpretable and can provide insight into the data structure. This is one of the major advantages of tree algorithms.

Ensemble methods

Dietterich (2000) describes ensemble methods as learning algorithms that construct a set of classifiers and then classify new data points by taking a weighted vote of their predictions. Examples of ensembles include bagging and boosting algorithms. A simplified representation of ensemble is shown in Fig.2.2[4]

Figure 2.2: Simplified representation of ensemble



[3]explained with an example: http://stackoverflow.com/a/20556654/1686651
[4]http://machinelearningmastery.com/a-tour-of-machine-learning-algorithms

2.1.5 Performance measures

In order to verify the correctness of a classifier and establish its suitability, it is required to measure its performance. Traditionally, the most commonly used metrics are "Accuracy" and "Error rate". Accuracy is the percentage of correctly classified instances, while Error Rate is the percentage of incorrectly classified instances.

Considering a basic two-class classification problem, let {Actual yes, Actual no} be a true positive and negative class label and {Predicted yes, Predicted no} be a predicted positive and negative class labels. Then, a representation of classification performance can be formulated by a confusion matrix (Table 2.1).
$Accuracy = (TP + TN)/(P_c + N_c)$
$ErrorRate = 1 - Accuracy$

Table 2.1: Confusion matrix

	Actual yes	**Actual no**
Predicted yes	True Postive (TP)	False Positive (FP)
Predicted no	False Negative (FN)	True Negative (TN)
column counts	P_c	N_c

Given that accuracy and error rate are inversely proportional, a classifier is considered suitable for a task if it has high accuracy and low error rate.

2.2 From Machine Learning to Text Understanding

A machine can not understand Text but it can simulate understanding. For that it must know, to some context, the rules of natural language. NLP deals with different aspects of language: phonology, morphology, syntax, semantics and pragmatics, while attempting to overcome ambiguity. The most common approaches when working with text data are bag of Words, n-grams, part-of-speech tags, and syntactic dependency relations. In addition, Topic modeling techniques are quickly gaining popularity especially for unsupervised learning tasks.

2.2.1 Bag-of-Words model

In this model, text (a sentence or a document) is represented as the bag (multiset) of its words, disregarding grammar and even word order, but keeping multiplicity. In other words, it uses the frequency of words as a feature for training a classifier.

As an example[5], here are two text documents:
(1) John likes to watch movies. Mary likes movies too.
(2) John also likes to watch football games.

Based on these documents, a list is constructed.

[
"John",
"likes",
"to",
"watch",
"movies",
"also",
"football",
"games",
"Mary",
"too"
]

The most common feature value used in the bag-of-words model is term-frequency. We can construct the following two lists to record the term frequencies of all the distinct words in the two documents:
(1) [1, 2, 1, 1, 2, 0, 0, 0, 1, 1]
(2) [1, 1, 1, 1, 0, 1, 1, 1, 0, 0]
Each entry of the lists refers to count of the corresponding entry in the list. Similarly there are 2 other feature values that can be used with Bag of Words model, namely "Binary model" and "tf-idf".
"Binary model" constructs lists indicating presence or absence of words. We can construct following binary lists for the example discussed above:
(1) [1, 1, 1, 1, 1, 0, 0, 0, 1, 1]
(2) [1, 1, 1, 1, 0, 1, 1, 1, 0, 0]
"TF-IDF" stands for "term frequency-inverse document frequency". TF-IDF

[5]Example from: https://en.wikipedia.org/wiki/Bag-of-words_model

is a method for emphasizing words that occur frequently in a document while at the same time de-emphasizing words that occur in many documents. For example words like “a”, “the”, “to” are words that appear in all documents, but this does not necessarily mean these words are important. Hence, a popular way to normalize the term frequencies is to weight a term by the inverse of document frequency.

2.2.2 N-gram model

The bag-of-words model is a document representation that ignores the word order. Only the occurrence of the words mattered. For instance, in the above example “John likes to watch movies. Mary likes movies too”, the bag-of-words representation will not reveal the fact that a person’s name is always followed by the verb “likes” in this text. As an alternative, the n-gram model can be used to store this spatial information within the text. Applied to the same example above, a bigram model will parse the text into the following units and store the term frequency of each unit, as before.

[
“John likes”,
“likes to”,
“to watch”,
“watch movies”,
“Mary likes”,
“likes movies”,
“movies too”,
]

2.2.3 Part-of-Speech (POS) tagging

POS tagging is the problem of assigning each word in the sentence with the part of speech it assumes in the sentence. Previously, this was done using rule-based systems, but over time, this problem has been solved using classification algorithms. The Penn Treebank project (Taylor et al., 2003) uses up to 45 tags.

For example, the sentence The Grand Jury commented on a number of

other topics is labeled as **“ The/DT grand/JJ jury/NN commented/VBD on/IN a/DT number/NN of/IN other/JJ topics/NNS ./. ”**
POS Tagging often proves to be a useful feature for many classification tasks.

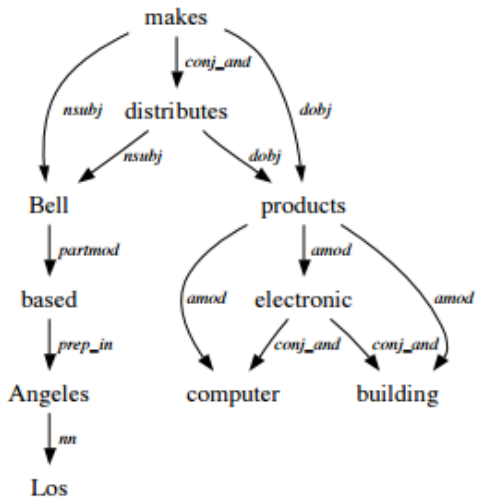
2.2.4 Syntactic dependency relations

Syntactic dependency parsers represent the syntactic structure of a sentence in terms of binary relations between tokens. For example, a verb is linked to its dependents (arguments/modifiers). Collectively, these relations form a tree or tree-like graph.

Stanford dependencies are a representation that categorizes/labels the head-dependent relation types. The current representation contains approximately 50 grammatical relations. A complete list can be found in Stanford dependency manual (De Marneffe and Manning, 2008).

Here’s an example sentence: *Bell, based in Los Angeles, makes and distributes electronic, computer and building products.* The corresponding dependencies are shown in Figure 2.3.

Figure 2.3: Graphical representation of the Stanford parser dependencies example



2.2.5 Topic modeling

Topic modeling is a statistical modeling technique for discovering the topics that occur in a collection of documents. The intuition behind topic models is, given that a document is about a particular topic, one would expect particular words to appear more or less frequently. For example[6], in a document about dogs, the words "dog" and "bone" will occur with high frequency. Similarly, in a document about cats, words "cat" and "meow" will occur with high frequency. While in both documents the words "the" and "is" will occur with similar frequencies. A document typically concerns multiple topics in different proportions, i.e., in a document that is 10% about cats and 90% about dogs, there would probably be about 9 times more dog words than cat words.

The "topics" produced by topic modeling techniques are clusters of similar words. A topic model captures this intuition in a mathematical framework, which allows examining a set of documents and discovering, based on statistics of the words in each, what the topics might be and what each document's balance of topics is.

In unsupervised learning, topic modeling algorithms, such as Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF), help us offer insights into large collections of unstructured text data.

Latent Dirichlet Allocation

Blei et al. (2003) describe LDA as a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document.

[6]The example is from Wikipedia: https://en.wikipedia.org/wiki/Topic_model

Non-negative Matrix Factorization

NMF[7], as the name suggests, is an algorithm based on Matrix Factorization (or decomposition). Matrix Factorization can be described as:
"Given a n x m matrix, R, find two smaller matrices, P and Q with k-dimensional features, e.g., where P has size n x k and Q has size m x k such that their product approximates R."
$R \approx P$ x Q^T
P is the features matrix, Q is the weights matrix. When they are multiplied as dot product, we obtain the sum of latent features by the weights for each element in the original matrix.

Non-negative matrix factorization returns features with no negative values. Therefore, all features must be positive or zero values. NMF is often used for unsupervised clustering algorithms.

2.3 NLP and Social Media

Recently, social media has gained immense popularity, ranging from discussion forums, blogs, Twitter (microblogging), Facebook, Instagram (photo-sharing), etc. These platforms allow generation of huge amounts of text in an informal environment. With popularity, social media has developed its own messaging trends inclusive of abbreviations, acronyms, neologisms, punctuation, and emoticons, to name a few. Existing words have taken new meanings, like "wall", "troll", "hashtag". Social media developed a vocabulary of its own.

NLP applications consisting of rule-based systems would need to add new rules to take into account the new words and abbreviations being developed everyday in order to extract useful information. Machine Learning applications are more adaptable to growing language.

NLP can analyze language patterns to understand text. One of the most compelling ways NLP offers valuable intelligence is by tracking sentiment the tone of a written message (tweet, Facebook update, etc.) and tag that text as positive, negative or neutral.

Much can be gleaned from sentiment analysis. Companies can target un-

[7]http://www.slideshare.net/BenjaminBengfort/non-negative-matrix-factorization

happy customers or find their competitors' unhappy customers, and generate leads. These are referred to as actionable insights[8] - findings that can be directly implemented into PR, marketing, adverting, and sales efforts.

A popular task is to predict user opinions, e.g., predicting the outcome of political campaigns from social media, or building user profiles to predict their likes and dislikes. The task we have undertaken also makes use of Twitter (social media) data.

There are limitations to what NLP algorithms can handle today. For instance, the tweeted phrase "You're killing it!" may either mean "You're doing great!" or "You're a terrible gardener!" No automated sentiment analysis that currently exists can handle this level of nuance.

Furthermore, certain expressions ("ima") or abbreviations ("#ff") fool the program, especially when people have maximum 140 characters to express their opinions, or when they use slang, profanity, misspellings and neologisms. Finally, much of social media interaction is personal, expressed between two people or among a group. Much of the language reads in first or second person ("I", "you", or "we"). This type of communication directly contrasts with news or brand posts, which are likely written with a more detached, omniscient tone.

2.3.1 Challenges in Social Media Data

The authors of the book Natural Language Processing for Social Media (Farzindar and Inkpen, 2015) note that the standard NLP methods applied to social media texts are confronted with difficulties due to non-standard spelling, noise and limited sets of features for automatic clustering and classification. Social media are important because the use of social networks has made everybody a potential author, so the language is now closer to the user than to any prescribed norms. Blogs, tweets, and status updates are written in an informal, conversational tone- often more of a stream of consciousness than the carefully thought out and meticulously edited work that might be expected in traditional print media. This informal nature of social media presents new challenges to all levels of automatic language processing.

[8] http://mashable.com/2011/11/08/natural-language-processing-social-media/#kmBBKU3Y7qqz

At the surface level, several issues pose challenges to basic NLP tools developed for traditional data. Inconsistent (or absent) punctuation and capitalization can make detection of sentence boundaries quite difficult - sometimes even for human readers, as in the following tweet: "#qcpoli enjoyed a hearty laugh today with #plq debate audience for @jflisee #notrehome tune was that the intended reaction?" Emoticons, incorrect or non-standard spelling, and rampant abbreviations complicate tokenization and part-of-speech tagging, among other tasks. Traditional tools must be adapted to consider new variations such as letter repetition ("heyyyyyy"), which are different from common spelling errors. Grammaticality, or frequent lack thereof, is another concern for any syntactic analysis of social media texts, where fragments can be as commonplace as actual full sentences, and the choice between "there", "they are", "they're", and "their" can seem to be made at random.

Social media are also much noisier than traditional print media. Like much else on the Internet, social networks are plagued with spam, ads, and all manner of other unsolicited, irrelevant, or distracting content. Even by ignoring these forms of noise, much of the genuine, legitimate content on social media can be seen as irrelevant with respect to most information needs. André et al. (2012) demonstrate this in a study that assesses user-perceived value of tweets. They collected over forty thousand ratings of tweets from followers, in which only 36% of tweets were rated as "worth reading", while 25% were rated as "not worth reading". The least valued tweets were so-called presence maintenance posts (e.g., "Hullo twitter!"). Pre-processing to filter out spam and other irrelevant content, or models that are better capable of coping with noise are essential in any language-processing effort targeting social media.

Several characteristics of social media text are of particular concern to NLP approaches. The particularities of a given medium and the way in which that medium is used can have a profound effect on what constitutes a successful summarization approach. For example, the 140-character limit imposed on Twitter posts makes for individual tweets that are rather contextually impoverished compared to more traditional documents. However, redundancy can become a problem over multiple tweets, due in part to the practice of retweeting posts. Topic drift is also more prominent in social media text than other published content.

2.4 Summary

This chapter introduced the nomenclature for Natural Language Processing and Machine Learning. We discussed commonly used techniques, features, classifiers, and performance measures for NLP task.

In our research, we focus on statistical NLP approaches which are based on Machine Learning algorithms. We apply supervised learning methods on social media text. To do so, we will perform experiments using a variety of features, and try several classifiers to develop a system that is able to predict users at-risk of depression from Twitter, without human intervention.

In the next chapter, we will present the literature review performed to discover the need for an automated mental illness detection system, the reliability of social media as source of data, and the existing systems developed to identify mental illness from social media.

CHAPTER 3

LITERATURE REVIEW

Mental illnesses, such as depression, are difficult to diagnose resulting in high under-diagnosis. This is, in part, due to the absence of a laboratory test for most forms of mental illnesses. Diagnosing mental illness is often based on self-reported experiences, behaviors reported by relatives, and a mental status examination.

This chapter highlights our investigations in three keys areas. First we evaluate the need for a system that can improve the process of detecting mental illnesses. Secondly, we analyze the reliability of social media as a source of data. Finally, we look at the existing systems created to solve similar problems, i.e., detecting mental disorders from social media.

3.1 About Mental Illness

Mental Illness, also known as the mental disorder, is defined as a "syndrome characterized by clinically significant disturbances in an individual's cognition, emotion regulation, or behavior that reflects a dysfunction in the psychological, biological, or developmental process underlying mental functioning" (American Psychiatric Association (APA, 2013)). According to the Canadian Mental Health Association (CMHA, 2016), 20% of Canadians belonging to varied demographics experience mental illness during their lifetime, and around 8% of adults go through major depression at some time in their lives. It has been recorded that approximately 1% of the population have experienced manic depression which is also known as bipolar disorder, and 1% have experienced schizophrenia. According to the World Health Organization (WHO, 2016) statistics, nearly 20% of children and adolescents have experienced mental illnesses and half of these mental illnesses start before the age 14. Also, around 23% deaths in the world were due to mental

and substance use disorders.

The annual Report (2014-15) published by the Mental Health Commission of Canada (MHCC, 2016) reports on the broad implications of mental illness, where from nearly 4000 Canadians that die each year by suicide, 90% of them were identified as having some form of a mental disorder. According to WHO (2016), each year around 800,000 individuals have committed suicide, making it the leading cause of death within the age group of 15-29. In addition, there are 20% more attempted suicides. According to WHO suicide is a preventable health problem and to be successful in preventing suicide, it is of great importance to identify depression and harmful use of alcohol so that necessary interventions can be provided.

Hence, to identify depression, we look at the cause for mental illness. WHO (2016) lists social, psychological, and biological factors as the three main factors that determine the level of mental health. For example, social factors will include socio-economical aspects such as low level of education, poverty, gender, discrimination, and stressful work conditions while psychological and biological factors include personality and genetic aspects. WHO also states that individuals with mental disorders have a higher probability of being ill with certain other diseases such as diabetes and cardiovascular diseases.

Apart from the severity of mental disorders and their influence on one's mental and physical health, the social stigma or discrimination in the forms of rejection, isolation, abuse, and fear of embarrassment associated with mental disorders cause the affected individuals to be neglected by the community. Furthermore, the Stigma and social misbeliefs such as "mental disorders cannot be cured", have resulted in individuals with mental illnesses to stay away from treatment (WHO, 2016). The unwillingness to obtain the necessary treatments can also be identified from the CMHA (2016) statistics, which state that nearly 49% of individuals who have identified themselves as being diagnosed with either depression or anxiety have not visited a doctor to obtain the proper treatment. Once untreated and depending on the severity of the diagnosis, an individual might have a higher probability of experiencing suicide ideation.

Although there has been an increase in the rate of treatment between 1990 and 2003 in regards to mental illness, the quality of treatment still falls below the minimal standards of quality (Kessler et al., 2003). It was also noted that treatment was provided to patients without disorders. Hence

it is suggested by Kessler et al. (2003) to put more emphasis on symptom screening procedures.

One of the initial screening procedures currently in use is to carry out surveys or questionnaires to assess the mental health status of a population. In such surveys or questionnaires, the interviewee could be vulnerable to memory bias and also adapt to the guidelines prescribed by the assessor. Survey responses could also be based on a part of the population being targeted. Conducting such studies is a financial burden for the institutions and also the error propagation within the dataset being gathered could be high.

Existing statistics and research provide evidence that a significant portion of the population suffer from mental disorders which often go undetected. Affected individuals hesitate to seek help due to the stigma associated with mental illness. Untreated mental illness can result in diseases such as diabetes and cardiovascular disease. It can also lead to an individual attempting suicide. Although there are treatments and preventive measures in place for individuals suffering from mental illness, reaching these individuals is a challenge even with surveys and questionnaires.

The discussion in this section establishes that there is a need for a cost-effective screening system that is able to identify individuals at-risk of mental illnesses at a large scale. To avoid memory-bias (relating to questionnaires and surveys), the system would require real-time data. As discussed in Section 2.3.1, social media updates are informal, contain conversational tone, and are more of a stream of consciousness. These attributes make social media an attractive source for obtaining real-time data. Being able to apply Text Mining methods to social media profiles would allow us to obtain real-time user data which can target a wider population in comparison to questionnaires at a much lower cost.

Next, we explore the reliability of social media data by looking at its existing applications.

3.2 Social Media Text Mining

The use of social media to detect mental illness could bring valuable results due to the rapid growth in using social media within different communi-

ties. Social media has become an integral part of everyday life where many people have started sharing their day-to-day activities. Such real-time data portraying one's daily life could reveal invaluable insights into ones cognition, emotion, and behavioral aspects, which could be hard to obtain from structured surveys. With its rapid growth among different demographics, social media can be a significant contributor to the process of mental disorder detection and prevent serious consequences like suicide.

3.2.1 Problems solved with the help of social media data

Researchers all around the world are making use of social media data to make predictions. For example Twitter is being used for event detection (Atefeh and Khreich, 2013), predicting seasonal influenza (Achrekar et al., 2012), crime prediction (Gerber, 2014), studying cyber-bullying (Xu et al., 2012) to name a few applications.

In regards to Facebook, researchers focus on more user-centric studies. Facebook status updates are used for recognizing user personality (Farnadi et al., 2013), mood and well-being (Wang et al., 2014). User's profile data is used for generating product recommendations (Gottschlich et al., 2013). Traud et al. (2012) studies the social structure of Facebook friendship networks at colleges and universities.

Instagram data was used for studies as well. Hochman and Schwartz (2012) analyze images from New York and Tokyo and show cultural visual differences in terms of local color usage, cultural production rate, and varied hue's intensities. Instagram was also used for detecting and tracking disease outbreaks as it comes with accurate geo-spatial information (Xie et al., 2013). A recent study carried out by Reece and Danforth (2016) discovered that there's a correlation between mood and color by observing the hue, saturation, and brightness of Instagram photographs.

Milne et al. (2016) organized a CLPsych 2016 shared task making use of ReachOut[1] forum data. The task involved identifying forum posts as "crisis, red, amber, or green" indicating the urgency by which a post requires a moderator's attention.

In addition to researchers, corporations are also leveraging social media

[1]ReachOut is set up to help young people get through everything from everyday issues to tough times

data to solve real-world problems. CBC reporter, Ireton[2] reports that Advanced Symbolics uses social media data to predict outcomes of Political Campaigns. SAS corporation successfully leveraged big data to predict risk of suicide among Canadian youth[3]. Qntify is using Social media data to help people with PTSD.

Successful results presented by researchers and corporations validate that Social Media data in fact can be leveraged as data source to make predictions at a large scale.

3.2.2 Use of social media data to identify mental disorders

The most important aspect to look into when considering detection of mental disorders through social web mining, is to understand the possibilities of using social media content to identify mental illness in a given population. A major contribution for this task is attributed to Choudhury (2013)(2014)(2015). De Choudhury investigates to what extent social media can be used as a source of information to understand mental illness among individuals as well as within a population. It is reported that among American adults, where 69% are online users, 66% are identified as using Facebook, followed by 20% using LinkedIn and 16% using Twitter. This gives an extensive heterogeneous representation of a society within different social media platforms, thus creating a rich source of sensors and allowing the identification of psychological and social behavioral patterns in longitudinal data.

De Choudhury et al. note that the use of language and emotion embedded within the postings could indicate depression. e.g., postings that represent worthlessness, guilt, helplessness, and self-hatred can be considered as indicators of depression. Additionally, increased use of first person singular pronouns in comparison to second and third person pronouns is also an indicator of depression. In contrast to linguistic features, withdrawal from social activities and changes within the social network relationships are also highlighted as possible indicators of depression. They also note that multimedia content can be used as an indicator. E.g., sharing more photos could be

[2]Now or then: will real-time data or past history better predict U.S. election? available at: http://www.cbc.ca/news/canada/ottawa/predicting-us-election-data-versus-history-1.3823042

[3]http://www.sas.com/en_ca/insights/articles/analytics/local/leveraging-big-data-to-predict-the-risk-of-suicide-among-canada.html

an indicator for an individual to be inversely correlated with depression. In addition to the number of photos, the content of such photos such as facial expressions are also identified as strong indicators. Similar to De Choudhury's work, Reece and Danforth (2016) discovered a relation between Instagram filters and depressed users. They show that "inkwell" filter, which turns color photos to black and white, was the most commonly used filter among depressed participants.

Choudhury (2013) emphasized the usefulness of identifying mental illness such as depressive disorder through social media posts and building tools that can be used by health care authorities and also by the user to take precautionary measures and to obtain the necessary treatments to avoid further impairment. Moreover, they raise awareness of the ethical implications of probing social media content to identify indicators of mental health disorders with the intent of informing health care authorities or relevant parties such as family members.

Other notable research in the same area includes, identification of suicide related risk factors from Twitter conversation (Jashinsky et al., 2014), CLPsych 2015 shared task organized by NAACL, and developing an annotation scheme for depressive disorder (Mowery et al., 2015).

- Jashinsky et al. (2014) filtered at-risk tweets from Twitter stream using keywords and phrases created from suicide risk factors, then compared values for suicide tweeters against national data of actual suicide rates from the Centers for Disease Control and Prevention. They found a strong correlation between state Twitter-derived data and actual state age-adjusted data.

- The North American Chapter of the Association for Computational Linguistics (NAACL) organized a Clinical Psychology Shared Task 2015 (CLPsych2015) organized by Coppersmith et al. (2015a) to detect and diagnose mental health problems from Twitter data. The challenge involved 3 tasks. Predicting depression users from control group, Predicting PTSD users from control group, and differentiating between depression and PTSD users. The proposed solutions to the challenge are discussed in more detail in section 3.3.2.

- Looking at the considerable amount of work being done to detect men-

tal illness from social media, creating a standard annotation scheme seems almost necessary. Mowery et al. (2015) have developed such annotation scheme for depressive disorder symptoms focusing specifically on Twitter Data.

3.2.3 Social media and self-disclosure

As social media interactions reside in a more naturalistic setting, it is important to identify to what extent an individual has disclosed his/her personal information and whether accurate and sufficient information is being published to determine whether or not a person has a mental disorder. Self-disclosure, as defined by Joinson and Paine (2009), is to reveal the unknown facts about one's self so that they become shared knowledge, or, in other words, the "process of making the self known to others". Park et al. (2012) have identified that, in addition to sharing depressed feelings, Twitter users are more likely to self-disclose to the extent where they reveal detailed information about their treatment history. To detect the stability of a social media platform to be used as a source to identify depression, the pilot study done by Park et al. (2012) revealed that, from a Twitter dataset collected using the word "depression", 42.04% of the tweets contained one's depressed feelings. Compared to CES-D (Centre for Epidemiologic Studies Depression Scale) score, a strong correlation was identified between the particular score obtained by a user and the use of words under different sentiment predictors such as anger, anxiety, sadness, and causation. The study also identified that certain demographic factors have a contribution towards depression, where participants without college degree (lower education) are more prone to depression compared to those with a college degree, and those with regular jobs are less likely to have depressive symptoms.

To identify the level of disclosure, honesty, and truthfulness expressed within mental health related social media forums, Balani and De Choudhury (2015) have used a perceptron classifier with adaptive boosting. The model performed with an accuracy of 74.8% compared with a precision of 0.74 and a recall of 0.86 in identifying high or low self-disclosure. With the use of the classifier, authors showed that reddit forums on mental health present high self-disclosure. It was identified that such posts tend to share fear, beliefs,

private, and sensitive information. Also, users with temporary accounts in the Reddit social media platform and specifically in the mental health forums have conveyed increased negativity, low self esteem, self-attention and cognitive bias (Pavalanathan and De Choudhury, 2015). In addition, the number of temporary accounts, which are being used, is six times higher in mental health forums compared to other Reddit forums. In deriving the conclusions, Pavalanathan and De Choudhury (2015) used features based on four main categories: affective, cognitive, linguistic and social, which were identified by using the Linguistic Inquiry Word Count (LIWC). The four categories were selected from nearly 32 categories of psychological constructs. In addition LIWC contains nearly 22 linguistic (e.g., auxiliary verbs, adverbs, prepositions), 12 punctuation (e.g., commas, periods), 7 personal concern (e.g. religion, home, work) and 3 spoken (e.g., nonfluencies, assents, fillers) categories (Pennebaker et al., 2007). The words in the posts could belong to either one or more word categories or subcategories where many of the categories are hierarchically organized.

To discover the extent to which users of temporary Reddit accounts have disclosed their sensitive information, Pavalanathan and De Choudhury (2015) looked into frequently used n-grams, that are identified through qualitative measures associated with the posts that tend to discourse extensive information with limited restraint. Presumably, the reasons for high self-disclosure could be because users do not need to share their personal information such as name, gender, age and location unlike in other social media platforms such as Facebook. Even though such highly self-disclosed posts have received fewer up-votes compared to down-votes, the level of social support was considerably high, indicating the willingness of such online communities to provide help and support. Furthermore, it was also identified that the higher the level of self-disclosure, the longer the person has been active within the particular forum.

To determine if there is any correlation between the personality of self-reporting individuals (on their mental health) and the use of language, Preotiuc-Pietro et al. (2015a) used age, gender and personality (identified using the five-factor model of personality), affect and intensity feature categories. The linguistic features were identified using the Linguistic Inquiry Word Count (LIWC) and also by using the most frequent n-grams (1 to 3-grams). In addition, a set of topics was identified as a method of reducing dimensionality to

improve the model's stability and reduce over-fitting. From the self-reported posts on post-traumatic stress disorder (PTSD) and depression, it was recognized that personality detected through text has enhanced the performances of the predictive model. Even though the use of language between the control group and the self-reporting users could be clearly distinguished using the model, certain limitations were identified when trying to differentiate the users into PTSD and depression groups. Also, an overlap on personality has been identified between the two groups.

3.3 Automatic Methods for Detecting Mental Disorders

With sufficient backing of literature to prove the popularity of social media platforms and its extensive use in disclosing personal information, it is important to look into the methods introduced by researchers to detect different mental illnesses in a given social media platform. In this section, we first look at the methods used for detecting depression, including levels of depression. Next we look at the systems submitted to CLPsych 2015 shared task for depression, PTSD, and control users.

3.3.1 Detecting depression

Both De Choudhry, Counts, & Horvitz (2013) and De Choudhry, Gamon, Counts, & Horovitz (2013) used crowd-sourcing to identify Twitter users reported to be in depression according to the standard psychometric measures. De Choudhry et al. have used behavioral characteristics identified under engagement, egocentric social graph, depression language, emotion and linguistic style to determine the cues of major depressive disorder symptoms. With an accuracy of 70% and a higher precision of 0.74 for the depression class, authors managed to identify vulnerable individuals to have depression before the start of their major depressive disorder. As contributors to the mentioned accuracy, reduced social activity, increased negative affect, clustered social network of the individual, raised interpersonal and medical fears and increased expression in religious involvement, were identified as strong indications leading towards major depressive disorder. Even though the im-

pacts of the egonetwork among individuals who are vulnerable to depression are being identified as considerably small, further analysis has shown that such networks are tightly clustered and close-knitted.

Similar to De Choudhury et al. (2013), Tsugawa et al. (2015) have demonstrated the effectiveness of using social media platforms among Japanese Twitter users in order to recognize depression. In addition to using Twitter user's activity history, a web based questionnaire was used to collect ground truth data in predicting the existence of depression for the Twitter users. In addition to the features used by De Choudhury et al. (2013), Tsugawa et al. used bag of words and word frequencies to identify the ratio of tweet topics. Even though subtle changes in behavioral features were identified between their research and the one done by De Choudry et al. which could be due to cultural aspects, the authors have identified similar analytical patterns for the use of negative words, posting frequency, retweet rate, and the tweets containing URL's. Feature engineering using Twitter user activity positively contributed towards a classification accuracy of 69%, with 0.64 precision, and 0.43 recall using support vector machine (SVM) classifiers. Topics identified using topic modeling also added positive contributions to the predictive model compared to the use of bag-of-words model, which could result in overfitting. Regarding the amount of Twitter data required in identifying depression, the authors have highlighted that at least two months of data is sufficient and data over a longer period could lead to lower accuracy.

To identify the level of depression among selected users and to increase the stability of the dataset, De Choudhury, Counts, et al. (2013) have used CES-D (Center for epidemiologic studies Depression Scale) and BDI (Beck Depression Inventory) standardized clinical surveys. Using a similar set of features under Twitter posts, user, and ego-network used by De Choudhry, Gamon et al. (2013), an additional feature named time under Twitter posts' feature category was used to distinguish users with depression. This is mainly because online social activities of users with signs of depression increase during evenings and nights. The model obtained an accuracy of 73% with a precision of 0.82 in predicting depression indicative posts. One of the key outcomes from the research is the calculation of social media depression index (SMDI) to scale the level of depression demonstrated by an individual through his/her postings on a given day. Reconfirming the higher level of accuracy obtained when calculating the SMDI, a significant correlation was

identified between SMDI values and the reported depression rates in different states of United States.

Determining the level of depression can be considered as an important factor when introducing interventions so that the relevant authorities can prioritize users according to level of depression. Schwartz et al. (2014) created a regression model to predict the level of depression based on Facebook status updates with moderate performances outperforming the baseline accuracy. The model was capable of identifying the change in the rate of depression over seasons, and specifically a decrease from summer to winter. The model was trained using features identified from n-grams, linguistic behavior and LDA topics. The model has made a unique contribution to the use of continuous values rather than discrete to identify the level of depression other than depression itself. Even though Choudhury, Counts & Horvitz (2013) developed similar measurements to identify the social media depression index (SMDI), they used their classification model for predicting the individuals who were susceptible to depression. Schwartz et al. on the other hand used nonclinical Facebook users who opted to complete 100 questions about their personality and shared access to their status updates. For feature engineering, relative frequency of n-grams (1 to 3-grams), probabilities of the topic (out from 2000 topics identified using LDA) given a user, 64 LIWC categories, sentiment analysis using sentiment lexicons, and the number of words were used. The model achieved better results compared to the baseline results which were obtained by using only sentiment features. Through the results and aligning with the literature in depression and its correlation with change of seasons, the authors concluded that depression can be successfully identified through the use of language in social media, and clear patterns of change in the degree of depression can be identified over the seasons where depressive linguistic style is more prominent in winter than summer.

The successful use of NLP techniques in identifying the progress and level of depression of individuals in online therapy could bring greater insights to clinicians, in order to apply interventions effectively and efficiently. McCabe et al. (2014), uses LDA for topic modeling on data gathered from an online psychological therapy provider to identify topics about therapy. LIWC and two machine-learning approaches were compared to identify the most suitable approach for calculating the sentiment values. The final set of independent variables was the therapist's identity number, higher level

features (e.g., client age, gender, session number), topic sentiment, words (unigrams) and n-grams. For the dependent variable, the binary conversion of Patient Health Questionnaire (PHQ-9) score was considered. The accuracies obtained for predicting the severity of the symptoms were comparable to face-to-face data. Also, it was identified that the use of linguistic features can be considered as more valuable in predicting the progress of a patient compared to sentiment and topic-based analysis.

The importance of identifying labels according to sentiment values was highlighted in the research conducted by Yu and Ho (2014). The authors use latent semantic analysis (LSA) and independent component analysis (ICA) to identify theoretical features from the psychiatric texts to derive the labels. Honkela and Hyvarinen (2004) explain LSA and ICA as follows. "LSA is a simple method for automatic generation of concepts that are useful, e.g., in encoding documents for information retrieval purposes. However, these concepts cannot easily be interpreted by humans. Independent component analysis applied on word context data gives distinct features which reflect syntactic and semantic categories. Thus, ICA gives features or categories that are both explicit and can easily be interpreted by humans". The reason for using ICA is to reduce the risk of performance decrease in the classification task due to term overlapping that could occur within the discovered latent concepts using LSA. Yu and Ho (2014) used SVM classifier with bag-of-words features, LSA, ICA, and a combination of LSA and ICA features to identify multiple emotion labels. The best performance for the classification task was obtained by using both LSA and ICA features, to identify multiple emotion labels, with an accuracy of 65% for 800 concepts. With greater precision and recall for ICA and LSA compared to word level features, authors concluded that features identified at the concept level can be used as powerful indicators in identifying emotion labels to be associated with the psychiatric texts.

3.3.2 CLPsych 2015 shared task

With considerable amount of individuals in different communities being diagnosed with at least some type of mental illnesses and also due to its adverse impact on suicidality, researchers have highlighted the importance of early detection of mental disorders. The naturalistic environment presented

through social media is continuously attracting more and more individuals, where ultimately it could become an integral part of one's life. Such an environment mirroring one's life makes it an ideal platform to identify mental health conditions. Due to importance of early detection and social media being a mirror of one's life, has increased the enthusiasm among researchers to delve into social media to detect mental health conditions. Text being one of the main mediums of communication within social media platforms, different shared tasks were initiated to encourage state of the art approaches to be demonstrated in a global platform.

The Computational Linguistics and Clinical Psychology (CLPsych) 2015 shared task, provided a platform for researchers around the world to demonstrate their state of the art approaches in identifying mental health conditions through social media platforms. Three tasks were given to participants, which include differentiating PTSD users from a control group , users with depression from the control group , and users with depression from users with PTSD. The task used self-reported data on Twitter on PTSD and depression according to procedure introduced by Coppersmith et al. (2014). Looking into the best performing models, varied approaches were identified in creating the predictive models such as: supervised topic modeling, rule-based, and, character language models.

Coppersmith et al. (2015b) experimented with a combination of systems submitted for the task and produced a system which outperformed the best solution implemented by Resnik et al. (2015) . In implementing the combined solution, Coppersmith et al. (2015b) used all the measures submitted by the 47 systems. Considering all the solutions provided by different researchers, topic modeling was identified to be a leading contributor for higher prediction accuracies. Even though linguistic features without using machine learning techniques have provided predictive power, it is not as significant as the contribution of topic modeling demonstrated by several participants.

Resnik et al. (2015), who was ranked first in the CLPsych 2015 shared task created 16 systems based on features derived using supervised LDA, supervised Anchor (for topic modeling), lexical TF-IDF and a combination of all. They highlighted the importance of TF-IDF features, which has made a significant contribution in obtaining higher accuracies used by itself as well as in combination with supervised topic modeling. The authors used SVM classifiers with linear or RBF kernel. They grouped the dataset into weeks

to avoid the issue of the dataset being too large or small. Finally, the dataset was pre-processed using basic pre-processing methods (e.g., remove emoticon character codes, removing stop words, lemmatizing). With 80% training and 20% testing data split, they managed to obtain the best results for all three tasks using an SVM classifier with linear kernel and using the big vocabulary for computing the models using all the features. The classifier obtained an average precision above 0.80 for all the three tasks and a maximum precision of 0.893 for differentiating PTSD users from the control group. Aligning with the literature, they identified that, out of the three tasks, differentiating PTSD users from users with depression to be the most difficult, because depression is identified as the most common mental health condition to be found in PTSD patients.

Preotiuc-Pietro et al. (2015b) used user metadata and textual features from the corpus provided by the CLPsych2015 shared task to develop a linear classifier to predict users having either one of the mental illnesses. They used bag-of-words, approach to aggregate word counts, topics derived from clustering methods, and Meta data (e.g., followers, followees, age, gender) from the users Twitter profile as the main feature categories. Several clustering approaches: brown clustering, NPMI word clusters, Word2Vec clusters, GloVe Word Clusters, LDA Word Clusters and LDA ERWord Clusters (ER) were used to reduce data sparsity and dimensionality of the feature space. With the use of logistic regression, linear support vector machines (LinSVM) and an ensemble of classifiers, authors managed to obtain an average precision above 0.80 for all three tasks and with maximum score of 0.867 for differentiating users in the control group from the users with depression. The model produced its best performances for smaller false positive rates.

The use of the supervised LDA and the supervised Anchor model was proven to be highly successful compared to the unsupervised clustering approaches and even more effective than using linguistic methods such as the use of n-grams and other lexicon based approaches. Resnik et al. (2015) have proven that such approaches can be successfully used in identifying users with depression, who has self-disclosed their mental illnesses on Twitter. Compared to the basic LDA clustering approach, supervised LDA approach includes document labels such as opinion analysis. Extending the basic LDA model, the anchor algorithm assigns an anchor word to each topic being identified. The key differentiator between the supervised LDA model

and the anchor algorithm is that rather than considering all the individual documents in the dataset, anchor model only requires a matrix with similar dimensionalities and word occurrences. Similar to supervised LDA, supervised anchor algorithm implements joint modeling, which is associated with document-level metadata. In agreement with Resnik et al. (2015), Coppersmith et al. (2015b) confirm that enhanced performances can be obtained by using LDA-like models such as supervised LDA and supervised anchor algorithm. In addition, they highlight the importance of using information priors, prevalence, and aggregation of data.

3.4 Summary

This chapter established that there is a need for a system that can help authorities to identify at-risk individuals on a large scale. We then established that social media is a widely-used and reliable source of data for research purposes and that it can be used for mental illness detection. We then corroborated that text mining can be used for mental illness detection from social media with reasonable performance.

Using these key aspects, in our research, we will use Twitter data to test if people at-risk of depression can be identified. We will further investigate if distressed tweets can be identified using text-mining techniques. We note that identifying the treatment requirement for a mental disorder is a complex clinical decision. It is also important to note that measuring the severity of the disorder is also a difficult task that could only be done by a highly-trained professional with the use of different techniques such as text descriptions, clinical interviews, and their judgment (APA, 2013). Considering the greater complexity of the procedures and the level of skills involved in identifying mental disorder and the necessary treatments, detecting at-risk users to mental illness within social media by using web mining and emotion analysis techniques could be considered as a primary step to create a focused group of population. Detecting mental disorder within social media will not be a platform for labeling an individual as a patient with a certain disorder, but only contributing as a platform for raising an alarm so that the relevant authorities could take necessary interventions to further analyze the predicted user to confirm his/her state of mental health.

Table 3.1: Summary: systems to identify depression from social media

Reference	Model	Features	Results
Choudhury, Gamon, Counts, Horvitz (2013)	SVM	engagement, ego-network, emotion, linguist. Style, dep. Language, demographics	accuracy: 70%; precision: 0.74
Tsugawa et al (2015)	SVM	features obtained from user activities can be used to predict depression; topics of tweets estimated with topic model are useful features; approximately 2 months of observation data are necessary for recognizing depression, longer observation periods may worsen accuracy.	accuracy: 69%
Schwartz et al (2014)	regression	Identifies the rate of change of depression over seasons. n-grams, linguistic behavior and LDA topics.	
Yu and Ho (2014)	SVM	Bag of Words, LSA, ICA to obtain multiple emotion labels	accuracy: 65%
CLPsych 2015			
Resnik et al. (2015)	SVM	supervised LDA, supervised anchor, lexical TF-IDF and a combination of all	precision above 0.80
Preotiuc-Pietro et al (2015b)	LinSVM	Bag of words, topics derived from clustering methods, meta data from user's profile	precision: 0.867

It is of great concern to respect ethical facets about the use of social media data and its privacy. Based on Twitter's policy, we take precaution to not release the gathered data. For this task, we also disregard any identifying user information. We take care of user's privacy and their ethical rights to avoid further distress.

In the next chapter, we present our data collection method and annotation guidelines.

CHAPTER 4

DATA COLLECTION AND ANNOTATION

4.1 Data Collection

The data for this research was collected by Advanced Symbolics, which is a consulting firm based in Ottawa that collaborated with our research team. The data was collected from Twitter posts (tweets) using the #Bellletstalk hashtag. Bell Let's Talk is a campaign created by Bell Canada to help reduce stigma and promote awareness and understanding of mental health issues. Canadians opened up the dialogue on mental health, contributing more than 122 million tweets, texts, calls and social media shares on Bell Let's Talk Day, helping raise more than $6.1 million for mental health initiatives. The data was collected for the year 2015 and was limited to Canadian users. 156,612 tweets were obtained from 25,362 users [1]. Only data made public by users was collected for this task. The dataset consists of 18 variables, most of them provided by Twitter API and some computed by the Advanced Symbolics team. These are listed in table 4.1

The reason we chose the #BellLetsTalk campaign is that for training our classifier we needed a source where we would find tweets relevant to mental health. Observing the data at a high level, we noticed that there were a lot of retweets and campaign publicity tweets. Retweets were discarded as they do not convey a user's own thoughts. This was done by removing all tweets that contained "RT" in them. Publicity campaigns were removed as they did not convey text related to mental health. There were also several statements that did not convey users' feelings, and so they were removed. Examples of tweets that we filtered out are listed below.

- " # BellLetsTalk is donating 5 cents to mental health research & care

[1] http://www.ctvnews.ca/health/bell-let-s-talk-breaks-records-raises-more-than-6m-for-mental-health-1.2211607

Table 4.1: List of tweet variables

variable name	**description**
Location	Location disclosed by User
Metro	Metro regions, e.g., Toronto, Ottawa
ID	Tweet's ID. This is unique for every tweet
Node	Node is an ID unique to every Twitter user
Name	Name provided by user on Twitter
Screen_name	User's chosen Twitter handle
Description	Description provided by user on Twitter
Image	Hyperlink to user's profile picture
Text	Text of tweet
Geo	Coordinates of a user's location, if disclosed by user
URL	URL used in a tweet
Media	Media (image, clip) attached with a tweet
Message_time	Time at which a tweet was posted
Retweet_count	Number of times the tweet was retweeted
Favourite_count	Number of times the tweet was favourited
Created_at	Retrieval time of tweet
In_reply_to_user_id	Node id of user being replied to
Favourites_count	Number of tweets favourited by the user

for every tweets! How do you support #mentalhealth? @healthy_minds"

- "RT @KristinBower: FACT: total # of 12-19 yr olds in CDA at risk for developing #depression is a staggering 3.2 million. #BellLetsTalk @Linda "

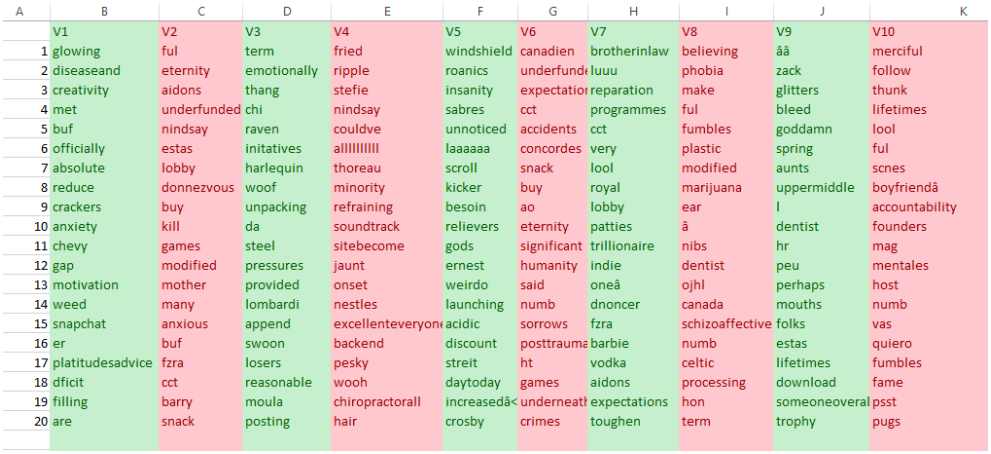
- "Tomorrow is .@Bell_LetsTalk day! Help end the stigma. Show your support. #mentalhealthawareness #mentalhealth #BellLetsTalk"

Publicity tweets and statements irrelevant to our task were primarily identified by looking at repetitive tweets. The words commonly occurring in these tweets were used to filter them. Next we tried LDA and NMF topic modeling techniques to identify frequently occurring topics and associated terms. Results are shown in Figures 4.1 and 4.2. Looking at the results for NMF (Fig. 4.2), we don't notice any obvious topics while results for LDA (Fig. 4.1) are much more insightful. e.g., Topic 4 in LDA results refer to the large number of tweets raising awareness for the BellLetsTalk campaign and raising donations. Topic 6 refers to the tweets where users urge other Twitter users to keep talking about mental illness, while Topic 8 refers to tweets regarding end

Figure 4.1: LDA results

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
1	lets	mental	support	will	dont	just	help	mental	today	people
2	day	health	amp	every	know	today	can	illness	tweet	one
3	talk	important	thanks	tweet	someone	make	get	stigma	tweets	depression
4	need	time	thank	bell	like	keep	awareness	end	million	many
5	love	get	much	donate	youre	year	please	around	now	life
6	conversation	illness	great	initiatives	never	canada	raise	help	bell	suffer
7	take	think	good	using	everyone	time	mental	time	donates	way
8	listen	canada	work	well	feel	talking	health	conversatic	well	anxiety
9	always	everyone	friends	donates	even	tweeting	money	health	twitter	still
10	help	people	cause	time	cant	going	everyone	thanks	tomorrow	can
11	everyone	help	see	today	just	clarahughes	always	awareness	time	dont
12	still	suffer	twitter	anxiety	alone	amazing	cant	make	every	think
13	now	day	year	tomorrow	see	come	every	need	will	see
14	talking	know	just	canada	always	lets	illness	still	day	talking
15	never	stigma	always	lets	get	day	time	alone	everyone	everyone
16	feel	thanks	talking	talk	going	tomorrow	donates	keep	end	feel
17	one	way	depression	twitter	take	now	feel	always	know	cause
18	year	alone	life	make	now	conversation	great	amazing	suffer	listen
19	every	always	dont	please	tweet	amp	know	amp	help	need
20	mental	amazing	time	someone	depression	thanks	amp	anxiety	life	alone

Figure 4.2: NMF results



	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
1	glowing	ful	term	fried	windshield	canadien	brotherinlaw	believing	âå	merciful
2	diseaseand	eternity	emotionally	ripple	roanics	underfunde	luuu	phobia	zack	follow
3	creativity	aidons	thang	stefie	insanity	expectatio	reparation	make	glitters	thunk
4	met	underfunded	chi	nindsay	sabres	cct	programmes	ful	bleed	lifetimes
5	buf	nindsay	raven	couldve	unnoticed	accidents	cct	fumbles	goddamn	lool
6	officially	estas	initatives	allllllllll	laaaaaa	concordes	very	plastic	spring	ful
7	absolute	lobby	harlequin	thoreau	scroll	snack	lool	modified	aunts	scnes
8	reduce	donnezvous	woof	minority	kicker	buy	royal	marijuana	uppermiddle	boyfriendâ
9	crackers	buy	unpacking	refraining	besoin	ao	lobby	ear	I	accountability
10	anxiety	kill	da	soundtrack	relievers	eternity	patties	â	dentist	founders
11	chevy	games	steel	sitebecome	gods	significant	trillionaire	nibs	hr	mag
12	gap	modified	pressures	jaunt	ernest	humanity	indie	dentist	peu	mentales
13	motivation	mother	provided	onset	weirdo	said	oneâ	ojhl	perhaps	host
14	weed	many	lombardi	nestles	launching	numb	dnoncer	canada	mouths	numb
15	snapchat	anxious	append	excellenteveryon	acidic	sorrows	fzra	schizoaffective	folks	vas
16	er	buf	swoon	backend	discount	posttrauma	barbie	numb	estas	quiero
17	platitudesadvice	fzra	losers	pesky	streit	ht	vodka	celtic	lifetimes	fumbles
18	dficit	cct	reasonable	wooh	daytoday	games	aidons	processing	download	fame
19	filling	barry	moula	chiropractorall	increasedâ<	underneath	expectations	hon	someoneoveral	psst
20	are	snack	posting	hair	crosby	crimes	toughen	term	trophy	pugs

the stigma. We Combined the keywords and statements identified manually and from LDA. These were used to remove tweets using pattern matching. A list of these words / phrases is given below:

("2015 campaign", "#BellLetsTalk commercial", "http", "howiemandel", "ads", "Jan 28" , "promote #BellLetsTalk", "#BellLetsTalk day", "#BellLetsTalk campaign", "Thanks for the RT", "Bell commercials", "conversation going", "Keep it up", "5 cents", "Jan. 28", "January 28" , "join the conversation", "Retweet", "thanks for the follow", "thanks for RT", "#BellLetsTalk 2015", "%", "days until #BellLetsTalk", "countdown to #BellLetsTalk is on", "please use the hashtag #BellLetsTalk", "find out what you can do to help", "#BellLetsTalk progress report", "commercial", "mental*

health awareness", "hashtag", "support #BellLetsTalk", "Stephen Harper", "how are you really feeling today", "reduce the stigma", "end the stigma", "break down the barriers")

Tweets containing two or more of the following terms (identified from LDA) were also removed.
("help", "awareness", "raise", "mental", "please", "today", "health", "tweeting", "can", "dont", "still", "thanks", "suffer", "important", "alone", "going", "even", "everyone", "donate", "stigma", "support", "bell", "initiatives", "conversation")

The above mentioned keywords were used to remove entire tweets. We refer to this process as cleaning. Cleaning process helped us to reduce the dataset size significantly for a more focused analysis.

After cleaning the dataset, we identified users who self-reported depression, i.e., they claimed to be suffering from depression or claimed to have suffered from depression at some point in life. This data was collected using keywords in combination with first-person pronouns ("I", "me", "you"). We further hand-labeled this data to confirm we had the right users. We noticed that some of these users were actually talking about a friend or family member suffering from depression. We were able to identify about 95 users who in fact did talk about their own diagnosis of depression.

We also noticed that many of the users claimed to suffer from anxiety disorders, and some talked about past suicide attempts. It might be interesting to look at users suffering from anxiety to see if they are more prone to depression; of course, only if there is any medical reason to support this claim.

Next, we obtained one-year tweet histories for the users that we retained with self-reported depression, we call this the self-disclosed set. We also obtained one-year tweet histories from those users who did not report depression, as a control set.

From the self-disclosed set and the control set, we created two datasets.

- 60Users: comprising of 30 users from self-disclosed set and 30 users from control set. This dataset was used for training and testing purposes.

- 100Users: comprising of 50 users from self-disclosed set and 50 users

from control set. This dataset was used for testing purposes only.

4.2 Data Annotation

For this research our focus is on supervised learning techniques, i.e., we will train classifiers on labeled data. Hence we must first ask human annotators to label a portion of the dataset.

For the 60Users dataset, we selected users who had between 100-300 tweets so annotators do not have to read too little or too many tweets for any one user. For the 100Users dataset we selected random users. We made sure that these users were not part of 60Users dataset. The users in both datasets were those who participated in BellLetsTalk campaign. The 60Users dataset was labeled at tweet-level and user-level while 100Users dataset was only labeled at user-level.

4.2.1 Tweet-level annotation

Before we began labeling, We prepared an annotation scheme based on paper "Towards Developing an Annotation Scheme for Depressive Disorder Symptoms: A Preliminary Study using Twitter Data" by Mowery et al. (2015). This annotation scheme labeled each tweet with "Psycho-social stressor", "Depression symptom" or "no evidence of clinical depression" at primary level. They break down each class into further categories and sub categories. e.g., a user talking about being awake at 4am in a tweet may be labeled as "depression symptom" at primary level, "disturbed sleep" at secondary level or "Difficulty in falling asleep" at tertiary level. See Appendix C for details. Using this annotation scheme, we decided to label at secondary level.

For the tweet-level annotations, we appointed two annotators. Annotator1 was a graduate student without knowledge of psychology, and Annotator2 was an undergraduate student with a minor in psychology.

Annotators were provided the 60Users dataset along with annotation guidelines in Appendix C. Using this annotation scheme, the two annotators labeled tweets from a few users as a trial. We found that the disagreement was high and we were missing a lot of tweets that we thought qualify as distressed, but could not be labeled as such, as they did not fit into any of the

categories in the annotation scheme. Therefore, we decided to change our annotation scheme to comprise of distress levels 0-3. We again performed a trial on 10 users. After labeling the 10 users, we reviewed the data labeled as distressed. The agreement for these was better than the earlier experiment. We did still have disagreements on levels of distress from 2-3, but we agreed on whether a tweet conveys some level of distress. Based on this trial, the annotators discussed and decided on what should fall in level 2 or level 3 distress. We also discovered that there were a lot of tweets that belonged to 0-level distress class and wanted to remove some of them, as annotation was a time-consuming task. For this, we obtained a list of positive negative words from AFINN net (Nielsen, 2011) and computed counts of positive and negative words for each tweet. Our intuition was to remove tweets containing positive words and keep those containing negative words. However, looking at the labeled data from 10 users, this intuition proved to be false. Tweets with positive words were often labeled as distressed. Next, we looked at neutral tweets, i.e., tweets with 0 positive and 0 negative words. In most cases, these were labeled with a 0-level of distress, and therefore, we decided that this would a good parameter to use for removing irrelevant tweets from our data. This helped reduce the number of tweets from 12,000 to almost 9,000. After removing the neutral tweets, annotators continued to label the remaining 50 users. While labeling, annotators were not given the information of whether a user belonged to the self-disclosed group or the control group, in order to avoid bias in labeling.

A total of 8,753 tweets belonging to 60 users were labeled. This resulted in 452 disagreements among annotators. The results of the annotation are shown in Table 4.2.

Table 4.2: Distribution of tweet labels with distress level 0-3

	annotator2			
annotator1	0	1	2	3
0	7,993	241	23	2
1	84	233	72	0
2	5	13	70	7
3	1	1	3	5

The Kappa value for the agreement between the two annotators for distress level 0-3 is 0.59. We report kappa instead of percentage agreement, because

kappa compensates for the agreement by chance. As it can be seen in table 4.2, both annotators agreed that 7,993 tweets belong to 0 or 'no-distress' class, while only 760 tweets indicate some level of distress. The data is highly imbalanced. Because of so few distress tweets, we decided to perform primary experiments for a binary class problem, that is, distress and no-distress as classes with the 7,993 tweets belonging to 'no-distress' class and the 760 tweets belonging to 'distress' class. All tweets that were not 0 were classified as 1 or 'distress' class. The results for the binary annotations are shown in Table 4.3. The kappa value for 2-annotator agreement for the binary-class problem was 0.67.

Table 4.3: Distribution of tweet labels with distress level 0-1

	annotator2		
annotator1		0	1
	0	7,993	266
	1	90	404

4.2.2 User-level annotation

60Users dataset

After completing the tweet-level annotations of 60Users dataset, annotator1 was asked to look at the distressed tweets of each user and assign user-level labels. The guidelines for the user-level annotations are as following:

- Depressed: User shows clear signs of depression, or shows signs that could result in depression in near future. There is enough reason for a public health member or doctor to investigate further.
- Not depressed: User does not show any signs of depression.
- Self-Reported: User shows no signs of depression, but claims he/she has been diagnosed with depression.

Annotation results of 60Users dataset by annotator1 are shown in Table 4.4

Table 4.4: Distribution of user labels

	depressed	not depressed	self reported
annotator1	13	35	12

100Users dataset

For user-level annotation of 100Users dataset, we appointed 3 annotators. Annotator1 was an undergraduate student with a minor in psychology, Annotator2 was a graduate student with a major in psychology. Annotator3 was a graduate student without knowledge of psychology.

This dataset was annotated after experiments were completed and models were trained and hence guidelines were adjusted accordingly. Annotator1 and Annotator2 were provided with an undersampled version of 100Users dataset. The dataset was undersampled using the best model trained for tweet-level classification task (discussed in Chapter 6 section 6.1). In order to undersample, we predicted tweet-class for all tweets in this dataset and removed tweets that were predicted as "no-distress". The reason for undersampling was to speed up annotation by removing some of the irrelevant tweets. Since this dataset is used for testing purposes only, we wanted the labels to consist of "depressed" and "not-depressed" only, keeping in mind the desired outcome and the selected user-level classification model. The new guidelines for user-level annotation are as follows:

- Depressed: User shows clear signs of depression, or shows signs that could result in depression in near future. There is enough reason for a public health member or doctor to investigate further. Additionally, if user self-reports depression but there are no other tweets indicative of depression.

- Not-depressed: User does not show any signs of depression.

From a total of 100 users, annotators disagreed on 14 users. The results for the annotations are shown in Table 4.5. The Kappa value for the agreement between the two annotators is 0.706. The disagreements among the annotators were resolved by annotator3.

We noticed a discrepancy in results shown in Table 4.5. The table shows 95 users and not 100. Looking into this issue revealed that during under-

Table 4.5: 100 users - annotations

		annotator1	
annotator2		not-depressed	depressed
	not-depressed	41	10
	depressed	4	40

sampling, for 5 users all tweets were predicted as "no-distress" and hence removed from the dataset. Therefore these 5 users were missing from the data provided to annotators. We later looked at the tweets for the missing users. All 5 users were found to belong to "control" set. Their tweets consisted mostly of retweets, URLs and commercial tweets. Examples of tweets from all 5 users are shown in Table 4.6. After going through all tweets for these 5 users manually, we decided to label these users as "not depressed".

Table 4.6: Examples of tweets from 5 missing users

#gadvcares Absolutely LOVE this! http://t.co/1KHWyV9WLp
#yyj #trafalgar DO not miss out on making wonderful travel memories this summer with Trafalgar! Let me help plan... http://t.co/u13k20doHK
RT @CruiseRadio: Five Airport Travel Tips #travel #airport http://t.co/7f4VNd2aZl http://t.co/EAjF6jW44h
RT @lonelyplanet: What to do on the #Panama Canal: adventure, wildlife & village life http://t.co/BJYHYfH0ke #lp #travel http://t.co/OVcnhv
Selling your home this year? Check out these #protips for choosing the right renovation projects: http://t.co/Xm5W2OSFf8
Easter gift to me. New Granite! http://t.co/dTQnRpjsHC
RT @LouRinaldiMPP: Jan. 28(Tomorrow) every tweet at #BellLetsTalk @Bell_LetsTalk = 5 cent donation to programs dedicated to mental health -
RT @AMStandard: Our Serin Faucets have a #bold, striking design that will add a modern look to your #bathroom http://t.co/1XcLERti3a
We're Open Until 8pm Today If You're Hungry! http://t.co/EPMXW1L1mo
RT @Solomomtravel: Enter to win $200 in gift cards for the best restaurants in the #Toronto beaches! http://t.co/hlayzVRNZH @hogtownsmoke @

4.3 Summary

This chapter explains the data preparation process. We collected public tweets for the campaign BellLetsTalk. From these tweets, we identified users who self-disclose depression. For these self-disclosed users, all tweets for the year 2015 were collected in self-disclosed dataset, while all tweets from year

2015 for non-self-disclosed users were collected in control dataset. Using self-disclosed and control datasets, we further prepare a dataset consisting of 60 users and another dataset consisting of 100 users. The 60Users dataset is labeled by 2 annotators at tweet-level and by 1 annotator at user-level. The 100Users dataset is labeled by 2 annotators at user-level and a 3rd annotator then resolves conflicts among the 2 annotators.

In the next chapter, we discuss the experiments performed to train both tweet-level and user-level classifiers.

CHAPTER 5

METHODOLOGY AND EXPERIMENTS

5.1 Experimental Setup

For this research all development is done in R version 3.3 (R Development Core Team, 2008) using Rstudio IDE (RStudio Team, 2015). Data Preparation, feature extraction, and classification tasks are performed using a variety of R packages. All classifiers were used from R's Caret package (Kuhn et al., 2012). Classifiers were trained using 10-fold cross validation to avoid over-fitting and then tested on a held-out test set. The results presented in Chapter 6 are for the the test set only, presenting recall, precision, f-measure of positive class.

5.2 TweetClass, UserClass

Our task consists of two components. First, to predict distress at tweet-level and second, to predict depression at user-level. This means our end goal is to train two classifiers. The first classifier predicts the TweetClass, while the second classifier predicts the UserClass.

For our experiments (performed on 60Users dataset), our training dataset must consist of features extracted from the tweets and the class variable. Class variable TweetClass consists of categories "distress" and "no-distress" while class variable UserClass consists of categories "depressed" and "not-depressed".

For our tweet-level classification task, we have class labels from the two annotators, with disagreements for some tweets. In order to obtain one TweetClass variable, we had two options: 1) to ask a third annotator to

label the disagreements; and 2) to generate labels based on the existing annotations from the two annotators. We chose the second approach, that is to generate labels from the existing annotations to save time and effort.

For this, we created two attributes either1 and both1. In either1, if either of the annotators labeled a tweet as distressed, the final annotation is 1 (distressed), and 0 (no-distress) otherwise. In case of both1, the tweet is annotated as 1 only when both annotators agreed it is 1; otherwise it is annotated as 0. See Table 5.1 for the details. We created both possible annotations, with the intention to see which performs better in experiments. The reason for choosing "either1" as TweetClass is that looking at the large imbalance, this gave us relatively more distressed tweets. On the other hand, the reason for using "both1" as TweetClass is that we can be sure that annotations are only for distressed tweets and may result in distinct identifiable features for the classifier to differentiate between the depressed and non-depressed classes.

We carried out a few experiments to see which attribute gives better results. There was very little to no difference between the results, so we decided to use either1 as TweetClass. Hence, TweetClass is an OR function of Annotator1 and Annotator2's labels. This resulted in 7,993 tweets labeled as 0 (no-distress) and 760 tweets labeled as 1 (distressed).

Table 5.1: Tweet annotations to TweetClass

annotator1	**annotator2**	**either1**	**both1**
0	0	0	0
0	1	1	0
1	0	1	0
1	1	1	1

At user-level, we only have annotations from 1 annotator so can directly be assigned to UserClass variable. We did, however, have another dilemma. Our task is a binary classification task i.e., to predict if a user has depression or not. There is no third option of "self-reported". Now we needed to clarify our task. If our intention is to find only those users whose tweets indicate depression, then the users labeled as "self-reported" would become "not depressed". However, if our intention is to find users who suffer from depression or are at-risk of depression regardless of whether their tweets, in our dataset, indicate this or not, then users labeled with "self-reported"

would become "depressed". There is also a possibility that "self-reported" users had distressed tweets in the past that are not part of our dataset, which consists of tweets from 2015. Therefore, we decided to investigate both scenarios in our experiments.

5.3 Training and Test Datasets

From the 60Users dataset, We divided the dataset to keep 40 users for training set and 20 for test set. The split was done randomly. In case of tweet-level classification, the tweets belonging to 40 users were part of training set and the tweets belonging to the other 20 users were part of the test set.

5.3.1 Why we need a training and a test set

When using Machine Learning, the goal is to develop a model that makes accurate predictions for a given task. One useful property of machine learning is that it not only makes predictions on data that it has seen, but also for data it has not seen. Jason Brownlee explains this well in his post[1] with the help of an example. Let's take an example of a hypothetical binary classification problem the goal of which is to classify data instances as red or green. For this problem let's assume there exists a perfect model, or a perfect function that can correctly discriminate any data instance from the domain as red or green. In the context of a specific problem, the perfect discrimination function very likely has profound meaning in the problem domain to the domain experts. We want to think about that and try to tap into that perspective. We want to deliver that result. Our goal when making a predictive model for this problem is to best approximate this perfect discriminant function.

We build our approximation of the perfect discriminant function using sample data collected from the domain. It is not all the possible data, it is a sample or subset of all possible data. If we had all the data labeled, there would be no need to make predictions, because the answers could just be looked up.

[1] "A Simple Intuition for Overfitting, or Why Testing on Training Data is a Bad Idea" http://machinelearningmastery.com/a-simple-intuition-for-overfitting/

The data we use to build our approximate model contains structure within it pertaining to the ideal discriminant function. The goal of data preparation is to best expose this structure to the modeling algorithm. The data also contains things that are irrelevant to the discriminant function such as biases from the selection of the data and random noise that perturbs and hides the structure. The selected model to approximate the function must overcome these obstacles.

The predictive model attempts to approximate a true discriminant function from a sample of data. For this, we want to use an algorithm that does not pick up and model all the noise in our sample, yet it generalizes beyond the observed sample data. It makes sense that we could only evaluate the ability of a model to generalize from a data sample on data that it had not seen before or during training.

The flaw with evaluating a predictive model on training data is that it does not inform us how well the model generalizes to new unseen data.

A model that is selected for its accuracy on the training dataset rather than its accuracy on an unseen test dataset is very likely to have lower accuracy on unseen data. The reason is that the model is not general. It has specialized to the structure in the training dataset. This is called overfitting.

5.3.2 Tackling overfitting

To tackle overfitting, it is suggested[2] to divide a dataset into training and test datasets where the predictive model is created using the training dataset and the performance is tested on unseen samples from the test dataset.

Another way to tackle overfitting is the use of cross-validation, common example of which is 10-fold cross validation. In 10-fold cross validation, the entire dataset is split into 10 parts and the algorithm is run 10 times. Each time the algorithm is run, it is trained on 90% of data and tested on 10%, and each run of the algorithm will change which 10% of the data the algorithm is tested on.

[2]http://machinelearningmastery.com/how-to-choose-the-right-test-options-when-evaluating-machine-learning-algorithms/

5.4 Tweet-level baseline experiment: SVM with BOW

As a preliminary experiment, we ran SVM on our data with Bag of Words (BOW) as feature set. This gave an accuracy of almost 95% with precision 1 and recall 0.01 (Table 5.2). At the first glance, the results appear very good as the accuracy is high. However looking at the confusion matrix (Table 5.3), the results are not convincing. It is clear from the confusion matrix that almost all the data is labeled as non-distressed, which makes us wonder why the accuracy is so high.

The reason for the high accuracy is that the true class for 95% of the data is actually "non-distressed" and with everything labeled as class 0, all of these instances are labeled correctly. Hence, the accuracy is simply a reflection of the percentage of class 0 instances. What our model learned here is that, by predicting everything as class 0, it can achieve a high accuracy. This is known as class imbalance problem.

Table 5.2: Tweet-level: baseline performance

accuracy	**precision**	**recall**	**F-measure**
0.9469	1.0000	0.0111	0.0219

Table 5.3: Tweet-level: baseline confusion matrix

	predicted 0	**predicted 1**
actual 0	1,589	0
actual 1	89	1

5.5 The Class Imbalance Problem

Any dataset that exhibits unequal distribution between its classes can be considered imbalanced. Commonly, imbalanced data corresponds to datasets that exhibit significant or even extreme imbalance. It is not uncommon to see imbalance in the order of 100:1, 1000:1, where one class severely out-represents another. In such cases, we require a classifier that will provide high accuracy for the minority class, without severely jeopardizing the accuracy on the majority class. Furthermore, this also suggests that the conventional evaluation practice of using singular assessment criteria, such as the overall

accuracy or error rate, does not provide adequate information in the case of imbalanced learning. Therefore, more informative assessment metrics such as the receiver operating characteristics (ROC) curves, precision-recall curves, and cost curves, are necessary for conclusive evaluations of performance in the presence of imbalanced data. These are discussed in section 5.5.3.

A lot of research has been done to tackle this problem. The proposed solutions can be divided into 5 main categories, namely, sampling methods, cost-sensitive methods, kernel-based methods, active learning methods, and one-class learning method. He and Garcia (2009) discuss all of these techniques in detail in their survey paper. For our research, we look at sampling methods and the one-class learning method.

5.5.1 Sampling methods for imbalanced learning

Studies show that for several base classifiers, a balanced data set provides improved overall classification performance compared to an imbalanced dataset. These results justify the use of sampling methods for imbalanced learning. The dataset can be balanced by adding or removing instances. In case of random oversampling, a set of random instances from minority class are selected. These are replicated and added to the dataset to increase the distribution of the minority class instances. In case of random undersampling, a set of random instances are selected from majority class and they are removed to lower the distribution of the majority class in the dataset. Both random oversampling and undersampling allow varying the degree of class distribution, to any desired level.

These do have some negative consequences. Random undersampling can cause the classifier to miss important concepts pertaining to the majority class. While random oversampling can lead to overfitting. These can be avoided by using informed undersampling methods. Informed undersampling methods make use of supervised learning with EasyEnsemble, BalanceCascade, or K-nearest neighbor (kNN) algorithm to systematically select which majority class samples to remove. EasyEnsemble samples several subsets from the majority class, trains a learner using each of them, and combines the outputs of those learners (Liu et al., 2009). BalanceCascade trains the learners sequentially, where in each step, the majority class examples that

are correctly classified by the current trained learners are removed from further consideration (Liu et al., 2009). kNN selects majority examples that are close to some of the minority examples (Mani and Zhang, 2003).

The dataset can also be balanced using synthetic samples. The Synthetic Minority Oversampling Technique (SMOTE) is a powerful method that has shown great success in various applications. SMOTE algorithm creates artificial data based on the feature space similarities between existing minority examples. SMOTE algorithm makes use of neighboring examples. In the SMOTE algorithm, the problem of over-generalization is largely attributed to the way in which it creates synthetic samples. Specifically, SMOTE generates the same number of synthetic data samples for each original minority example and does so without consideration to neighboring examples, which increases the occurrence of overlapping between classes. Various adaptive sampling methods have been proposed to overcome this limitation (He and Garcia, 2009).

5.5.2 One-class learning method

One-class learning and novelty detection has also attracted much attention. This category of approaches aim to recognize instances of a concept by using mainly, or only, a single class of examples rather than differentiating between instances of both positive and negative classes. This includes one-class SVM and the auto-associator method. Specifically, Raskutti and Kowalczyk (2004) suggested that one-class learning is particularly useful in dealing with extremely imbalanced data sets with high feature space dimensionality.

5.5.3 Assessment Metrics for the Class Imbalance Problem

In addition to the large variety of algorithms to solve class imbalance problem, it is important to assess those algorithms in a standardized way. There are several performance assessment metrics used to assess imbalanced datasets.

Singular assessment metrics

In case of an imbalanced dataset, accuracy and error rate can be deceiving. For example, Considering minority class a positive and majority class a negative, if a dataset includes 5% minority examples and 95% majority examples, a naive approach of classifying every instance to be a majority class would provide 95% accuracy. 95% accuracy appears to be good; however, it fails to reflect the fact that 0 percent of minority examples are identified. That is to say, the accuracy metric in this case does not provide adequate information on a classifier's functionality with respect to the type of classification required. It becomes difficult to analyze the results when the evaluation metrics are sensitive to data distributions. Instead of accuracy, other evaluation metrics are often adopted to provide comprehensive assessments of imbalanced learning problems, namely precision, recall, and F-measure. He and Garcia (2009) define these as:
Precision $= TP/(TP + FP)$
Recall $= TP/(TP + FN)$
F-measure $= [(1 + Beta)^2.Recall.Precision]/[(beta)^2.Recall + Precision]$, where beta is a coefficient to adjust the relative importance of precision vs. recall. (usually beta =1)

Intuitively, precision is a measure of exactness (i.e., from the examples labeled as positive, how many are actually labeled correctly), whereas recall is a measure of completeness (i.e., how many examples of the positive class were labeled correctly). These two metrics, much like accuracy and error, share an inverse relationship between each other. However, unlike accuracy and error, precision and recall are not both sensitive to changes in data distribution. Precision is sensitive to data distribution, while recall is not. On the other hand, that recall is not distribution dependent is superfluous because an assertion based solely on recall is equivocal, since recall provides no insight to how many examples are incorrectly labeled as positive. Similarly, precision cannot assert how many positive examples are labeled incorrectly. Nevertheless, when used properly, precision and recall can effectively evaluate classification performance in imbalanced learning scenarios. Specifically, the F-measure metric combines precision and recall as a measure of effectiveness of classification in terms of ratio of the weighted importance on either

recall or precision as determined by the Beta coefficient set by the user. As a result, F-measure provides more insight into the functionality of a classifier than the accuracy metric, while remaining sensitive to data distributions. Though F-measure is an improvements over accuracy, it is still ineffective in answering more generic questions about classification evaluations. For instance, how can we compare the performance of different classifiers over a range of sample distributions?

Receiver Operator Characteristics (ROC) curves

The ROC assessment technique makes use of the proportion of two single-column-based evaluation metrics, namely, true positive rate (TP_rate) and false positive rate (FP_rate), which are defined as:
$TP\_rate = TP/P_c$
$FP\_rate = FP/N_c$
where P_c is the count of actual positive instances and N_c is the count of actual negative instances.

The ROC graph is formed by plotting TP_rate over FP_rate, and any point in ROC space corresponds to the performance of a single classifier on a given distribution. The ROC curve is useful because it provides a visual representation of the relative trade-offs between the benefits (reflected by true positives) and costs (reflected by false positives) of classification in regards to data distributions.

In order t o assess different classifiers' performance, area under the ROC curve (AUC) is often used as an evaluation criteria.

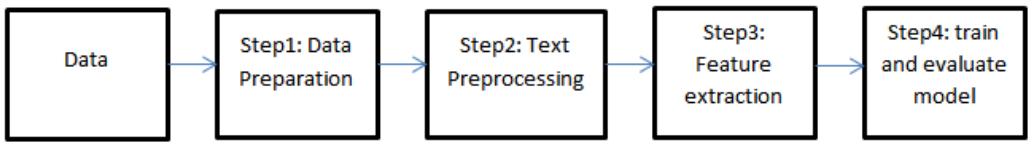
We use under-sampling m ethods and SMOTE to deal with class imbalance. We also look at precision, recall and F-measure for comparing performance of classifiers[3].

5.6 Text Pre-processing

Now that we have training and test sets from 60Users dataset, we can proceed to pre-processing and feature extraction.

[3]We do not compare performance using AUC as it is obtained using 10-fold cross-validation on training set and may not be an accurate measure of how model performs on unseen test data.

Figure 5.1: Process for developing a prediction model



Normalization (also called pre-processing) is a process that converts list of words to a more uniform sequence. This is useful when we require text to be in a certain format before performing operations on it. Normalization is often domain specific. There is no one solution for all applications. Examples of normalization include removing symbols or punctuation, replacing characters like $200 with words "two hundred dollars", changing text to lowercase etc. In context of Twitter, emoticons such as :), :(are often replaced with their textual forms i.e., happy, sad correspondingly.

In our application we normalized the tweets by removing hyperlinks, punctuation (except , and $), special characters, and @BellLetsTalk. We also convert text to lowercase. Finally we removed tweets with less than 5 words. Cleaning phase discarded entire tweets, while normalizing phase removed characters from a tweet's text.

In many applications, normalizing includes removing stopwords, but we skipped this step primarily because we wanted to keep words like "I", "me", "you" (first person pronouns). Literature review shows that first person pronouns are useful predictors when predicting depression. Secondly, we wanted to maintain the readability of the text.

5.7 Feature Engineering

After pre-processing, we now have text in a desired format. The next step is to extract features from text. Extracted features include counts of Positive and Negative words, counts of depression-related words, counts of first person and second person pronouns, and the Bag of Words.

These features are computed for both the training and test sets. If a model is trained with 7 features, the test set must also contain the same 7 features. In our case, the polarity words module gives 4 features, the depression words give one feature, and the pronouns module returns 2 features. However, the

number of features returned by the Bag of Words (BOW) module is not fixed.

5.7.1 Polarity words

To count the number of polarity words, i.e., Positive and Negative words, we first need to define which words are positive and which words are negative. For this we use The AFINN net's AFINN-111(Nielsen, 2011) word list. AFINN is a list of English words rated for valence with an integer between -5 (negative) and +5 (positive). The words are manually labeled by Finn Arup Nielsen. The AFINN-111 contains a list of 2,477 words and phrases. We process AFINN-111 to obtain 4 features. vNegativeTerms (score -5, -4), NegativeTerms (score -3, -2, -1), PositiveTerms (score 1, 2, 3), and vPositiveTerms (score 4, 5).

To count polarity words, documents/tweets are first tokenized. Tokenization means to break a sentence into words. In our implementation, tokens are split by space, punctuation and numbers. For example, the sentence "Been awake since 6am for no reason" contains 8 tokens (see Table 5.4). Once we have tokens, we can simply count the polarity words among these tokens and store them as feature counts.

Table 5.4: Tokenization example

sentence	tokens
"been awake since 6am for no reason"	"been", "awake", "since", "6", "am", "for", "no", "reason"

5.7.2 Depression words

To count depression terms, we needed to define which terms to count. For this we use a list of 204 depression-related terms from (Maigrot et al., 2016) (see Table 5.5). A feature was calculated that counts the number of depression terms in a tweet. Instead of counting the frequencies of each depression term, we count all depression terms as a single feature.

Table 5.5: List of depression words

“agonized”,“aid”,“alienated”,“alienation”,“alone”,“anger”,“angry”,“anguish”, “anguished”,“antidepressant”, “anxiety”,“anxious”, “attempt”,“awful”, “barren”,“beaten”,“better place without me”, “blas”,“bleak”,“bleeding”, “blue”, “child”,“communication”,“concern”,“confusion”,“courage”,“crestfallen”,“cruel”, “crushed”,“crying”, “cycle”,“dead”,“death”,“Death-seeking”,“debilitating”,“defeated”, “dejected”,“demoralized”,“depressed”, “depression”,“descent”,“desolate”, “despair”,“despondent”,“detriment”,“devalued”,“devastated”,“die”, “disappointed”, “discouraged”,“discrimination”,“disease”,“disinterest”,“disinterested”,“dismal”, “disorder”, “dispirited”,“distracted”,“distressed”,“doctor”,“dog days”,“done with life”, “doomed”,“down”,“downcast”, “downhearted”,“drained”,“drugs”,“effect”,“empty”, “endure”,“esteem”,“family”,“fatalistic”,“fatigued”,“fear”, “fed up”,“feelings”,“fight”, “finality”,“friends”,“gain”,“gloomy”,“glum”,“grief”,“grieved”,“grieving”,“grim”, “hard work”,“health”,“help me”,“helpless”,“hopeless”,“hopelessness”,“Hot-line”,“hurt”, “I cry”,“I’m crying”, “I’m done”,“immune”,“improvement”,“in despair”,“inability”, “inactivity”,“indifferent”,“insecure”, “interested”,“involvement”,“irritable”,“isolated”, “isolation”,“joyless”,“kill”,“kill myself”,“lack”, “life quality”,“lost”, “media”,“medication”,“medicine”,“melancholia”,“melancholy”,“mental”,“miserable”, “misunderstanding”,“moody”,“morose”,“necessary”,“need”,“negative”, “not pretty”,“nothing”,“option”, “overcome”,“pain”,“panic”,“parents”,“passionless”, “patience”,“patient”,“pattern”,“pay attention”,“peers”, “pessimistic”,“pills”, “pleasureless”,“prescription”,“prevent”,“prevention”,“progress”,“protect”,“quantity”, “reality”,“reckless”,“regretful”,“requirement”,“rotten”,“sadness”,“scared”,“security”, “Self-destructive”, “separation”,“seriousness”,“signs”,“skills”,“solitary”,“somber”, “sorrowful”,“struggle”,“studies”,“succor”, “suffer”,“suicide”, “sullen”,“sympathetic”,“symptoms”, “tearful”,“teenagers”,“terrified”,“therapy”, “thoughts”,“tired”,“tormented”,“torture”, “tragedy”,“tragic”,“treat”,“treatment”,“trouble”,“troubled”, “uncertain”,“uncomfortable”,“unfulfilled”,“unhappy”,“unique”,“upset”, “victim”,“warning”,“weepy”, “woeful”,“worried”,“worry”,“worthless”,“zero”

5.7.3 Pronouns

We count the first person and second person pronouns as features called “firstPronounCount” and “secondPronounCount”, respectively. First person pronouns are defined as “i”, “me”, “my”, “myself”, “mine”, “we”, “our”, “ours”, “ourselves”, “us”. Second person pronouns are defined as “you”, “your”, “yours”, “yourself”, “yourselves”.

5.7.4 Bag of words

Bag Of Words is simply a term frequency matrix (see section 2.2). In contrast with polarity words, depression words, and pronouns, where number of features is fixed, Bag of words does not return a fixed number of features.

If we compute BOW features for training set and test set independently,

features may vary in which case trained model will not be able to make predictions on the test set. To avoid this, we compute BOW features from the training set only. Let's say we get 100 features on training set. In case of test set, we only compute values for these given 100 features. This ensures that the trained model can make prediction on any new test data.

5.8 Tweet-level Classification

The goal of training a tweet-level classifier is to make predictions whether a given tweet indicates distress or not. We performed several experiments with a variety of features and classification algorithms.

1. with 7 features (vNegTerms, negTerms, posTerms, vPosTerms, depressionWordCount, firstPronounCount, secondPronounCount)
    - train svmLinear classifier (**exp1-svm-original**)
    - train svmLinear with data balanced by SMOTE (**exp1-svm-SMOTE**)
    - train svmLinear with data balanced by undersampling (**exp1-svm-Down**)
2. with 2917 features (vNegTerms, negTerms, posTerms, vPosTerms, depressionWordCount, firstPronounCount, secondPronounCount, BOW)
    - train svmLinear classifier (**exp2-svm-original**)
    - train svmLinear with data balanced by SMOTE (**exp2-svm-SMOTE**)
    - train svmLinear with data balanced by undersampling (**exp2-svm-Down**)
3. with 2910 features (BOW)
    - train svmLinear classifier (**exp3-svm-original**)
    - train svmLinear with data balanced by SMOTE (**exp3-svm-SMOTE**)
    - train svmLinear with data balanced by undersampling (**exp3-svm-Down**)
4. training data undersampled to remove two-thirds of non-distress tweets per user. with 1319 features (vNegTerms, negTerms, posTerms, vPosTerms, depressionWordCount, firstPronounCount, secondPronounCount, BOW)

- train svmLinear classifier (**exp4-svm-original**)
- train svmLinear with data balanced by SMOTE (**exp4-svm-SMOTE**)
- train svmLinear with data balanced by undersampling (**exp4-svm-Down**)

Experiment1 uses only features derived from text, Experiment2 uses BOW in addition to text-derived features. Experiment3 uses only BOW features. Experiment4 is similar to Experiment2 except that the data is undersampled to remove two-thirds of non-distress tweets from each user. In all four experiments we try variation of feature combinations and sampling techniques to find out which classifier performs the best.

Experiments were also performed using Random forest, SVM with Radial kernel, and Naive Bayes algorithms but are not discussed here as results were unsatisfactory.

5.9 CLPsych2015 Dataset

Before we discuss the user-level classification, it is important to mention another dataset that we obtained. Some models were trained on CLPsych2015 dataset and then tested on our BellLetsTalk dataset.

The dataset for the CLPsych2015 shared task is not publicly available. It was obtained from Glen Coppersmith, one of the organizers for the task, by providing an ethical exemption document. Although the tweets are public data, the labels prepared by the shared task organizers are not.

The dataset consists of 1,746 users. The training set contains 327 depression users, 246 PTSD users, and, for each, an age and gender matched control user. The test set contains 150 depression users, 150 PTSD users, and an age and gender matched control user for each.

Coppersmith et al. (2015a) explain that the dataset was collected by looking for phrase "I was just diagnosed with X" where X can be depression or PTSD. A human annotator then evaluated each such statement of diagnosis to remove jokes, quotes, or any other disingenuous statements. For each user, the most recent 3,000 public tweets were included in the dataset. The tweet in which the genuine statement of diagnosis was found was removed, to avoid any bias created from the data sampling technique.

Our interest in this dataset was for the reason that it was large, labeled, and balanced unlike our own BellLetsTalk dataset.

From the CLPsych2015 dataset, we separated 327 depressed users and their corresponding control users from the training set, for our experiments. We trained 5 models on this dataset. They can be described as:

- **Model1**: with 9,481 features (negWordCount, vNegWordCount, posWordCount, vPosWordCount, depressionWordCount, firstPronounCount, secondPronounCount, DepressedTweetCount, TotalTweetCount, depressedTweetsPercentage, BOW)
  - train svmLinear classifier
- **Model2**: with 9,483 features (negWordCount, vNegWordCount, posWordCount, vPosWordCount, depressionWordCount, firstPronounCount, secondPronounCount, DepressedTweetCount, TotalTweetCount, depressedTweetsPercentage, BOW)
  - train randomForest classifier
- **Model3**: with 10 features (negWordCount, vNegWordCount, posWordCount, vPosWordCount, depressionWordCount, firstPronounCount, secondPronounCount, DepressedTweetCount, TotalTweetCount, depressedTweetsPercentage)
  - train randomForest classifier
- **Model4**: with 9,483 features (negWordCount, vNegWordCount, posWordCount, vPosWordCount, depressionWordCount, firstPronounCount, secondPronounCount, DepressedTweetCount, TotalTweetCount, depressedTweetsPercentage, BOW limited to those from depressed users)
  - train randomForest classifier
- **Model5**: with 9,381 features (negWordCount, vNegWordCount, posWordCount, vPosWordCount, depressionWordCount, firstPronounCount, secondPronounCount, DepressedTweetCount, TotalTweetCount, depressedTweetsPercentage, BOW without stopwords)
  - train randomForest classifier

Model1 uses SVM while Model2-5 use Random Forest. Model1 uses text-derived features and BOW, Model3 uses only text-derived features. Model2, 4, 5 use text-derived features and BOW. They differ in the words included in BOW representation.

5.10 User-level Classification

The goal of training a user-level classifier is to predict whether a given user may suffer from or is at-risk of suffering from depression.

For the user-level classification, we start by merging data for each user. The merging module takes care of three things. First, it combines the texts of all the tweets of the user into a single document. Second, it sums up the features calculated during the tweet-level classification, i.e., sums up the polarity word counts, depression word counts, and pronoun counts. Third, it uses the predictions from the tweet-level classifier and counts the number of distressed tweets as a new feature "depressedTweetCount". The merging process is equivalent to recomputing features. In addition to "depressedTweetCount", 2 additional features are added namely "TotalTweetCount" and "depressedTweetsPercentage". TotalTweetCount is a numeric variable that indicates the number of tweets by a user in the dataset. DepressedTweetsPercentage is a numeric variable indicating percentage of depressed tweets posted by user.

5.10.1 Preliminary experiments

After dataset is prepared, we run several experiments to find a model that provides highest recall. These experiments 5-11 below were done by setting self-reported users as "depressed" in scenario "a" and self-reported users as "not depressed" in scenario "b".

5. with 10 features (negWordCount, vNegWordCount, posWordCount, vPosWordCount, depressionWordCount, firstPronounCount, secondPronounCount, DepressedTweetCount, TotalTweetCount, depressedTweetsPercentage) trained on 60Users dataset

    - train svmLinear classifier (**exp5-svm-Orignal**)

- train svmLinear with data balanced by SMOTE (**exp5-svm-SMOTE**)

6. with 2406 features (negWordCount, vNegWordCount, posWordCount, vPosWordCount, depressionWordCount, firstPronounCount, secondPronounCount, DepressedTweetCount, TotalTweetCount, depressedTweetsPercentage, BOW) trained on 60Users dataset
    - train svmLinear classifier (**exp6-svm-Orignal**)
    - train svmLinear with data balanced by SMOTE (**exp6-svm-SMOTE**)
7. model1 trained on CLPsych2015 dataset (**exp7-svm-Original**)
8. model2 trained on CLPsych2015 dataset (**exp8-rf**)
9. model3 trained on CLPsych2015 dataset (**exp9-rf**)
10. model4 trained on CLPsych2015 dataset (**exp10-rf**)
11. model5 trained on CLPsych2015 dataset (**exp11-rf**)

Experiment5 and 6 differ by use of BOW in Experiment6. They are both trained on training set derived from 60Users dataset. Experiment7-11 are trained on CLPsych2015 dataset.

5.10.2 Improvement

Next, we attempt to improve precision and recall by focusing on users whose tweets actually reflect depression and are not just self-reporting depression. The rationale behind this is that users who self report depression but their tweets are not indicative of depression might make it difficult for the classifier to find useful features. By removing these users from the training set, we hope that some patterns might become more prominent. We do this by following these three steps:

- Training Data
    - step1: identify and remove users labeled as "self-reported" from the training set. Set the prediction of these users as "depressed".
    - step2: train on the remaining users, and make predictions.

- step3: merge results from step1 and step2 to obtain the final confusion matrix.

- Test Data
  - step1: identify users who self-report depression but have less than 10% distressed tweets (using text matching); mark these users as depressed.
  - step2: make predictions on the remaining users.
  - step3: merge results from step1 and step2 to obtain final confusion matrix.

At step2, we repeat the experiments "exp5-svm-SMOTE" and "exp11-rf". These experiments, "**exp14-5**" and "**exp14-11**", correspondingly return much better results than the preliminary experiments. These will be discussed further in chapter 6 section 6.2.

5.10.3 Introducing more features

In a final experiment (**exp15**), additional features are generated. A total of 116 features were used for this experiment. These are listed and described in Table 5.6. This new data was used to train a variety of classifiers including Linear SVM, Random forest, Naive Bayes, weighted Linear SVM, svmLinear2 with linear kernel, k-nearest neighbor, SVM with radial kernel, and SVM with polynomial kernel. Through this experiment, we wanted to see if large number of features has a significant impact on performance compared to previous experiments (using 7 or 10 features).

5.11 Other Experiments

Additional experiments performed at tweet-level include using one-class SVM, Naive Bayes Multinomial, meta classifier, weighted Naive Bayes, and SVM with linear, polynomial, sigmoid, and radial kernels. Unlike in the previous experiments (performed using caret package in R), these algorithms were tested in WEKA.

Table 5.6: Additional features

feature_groups	**feature names**	**description**
polarity_words (4)	negWordCount, vNegWordCount, posWordCount, vPosWordCount	counts of polarity words
pronoun (2)	firstPronounCount, secondPronounCount	counts of Pronouns
Tweet Count (1)	Tweet count	Count of depressive tweets
depression percentage (1)	depression_percentage	percentage of depressive tweets
community features (5)	favourite_count, favourites_count, replies, retweets_by_user, mentions	features related to user interaction. not derived from text
Liwc features (87)	wc, wps, dic, allpunct, period, comma, colon, semic, qmark, exclama, dash, quote, apostro, parenth, functionw, {LIWC dictionary word}	obtained from LIWC tool
sentiment features (12)	sentiment, anger.y, anticipation, disgust, fear, joy, sadness, surprise, trust, negative, positive, emoValence	relating to text sentiment
emoticon features (2)	emj_count, emj_senti	emoticon count and sentiment of those emoticons
readability features(2)	SMOG, Flesh.Kincaid	SMOG Readability Formula estimates the years of education a person needs to understand a piece of writing. The FleschKincaid readability tests are designed to indicate how difficult a reading passage in English is to understand.

Furthermore, for tweet-level classification, we also changed the distribution of the dataset to split into 80%-20% training and test set on the TweetClass variable instead of UserClass. This too did not improve the results. These results are not presented in Chapter 6, since they were unsatisfactory.

5.12 Summary

In this chapter, we discussed the experiments we performed for training tweet-level and user-level classifiers. For this, we pre-processed the data,

derived features from text, and then designed several experiments using various features and Machine Learning algorithms to train suitable classifiers. We also mentioned experiments that we tried, but did not perform well and hence are excluded from this manuscript.

In the next chapter, we will present results for the experiments described in this chapter and discuss the performance of the models.

CHAPTER 6

RESULTS AND DISCUSSION

For both tasks, tweet-level classification and user-level classification, we choose precision, recall & F-measure as performance measures. Precision and recall are selected instead of accuracy due to the data being imbalanced. Baseline experiments for tweet-level classification return an accuracy of 95% by classifying all samples as majority class, which is not a true reflection of classifier performance.

Precision is sensitive to data distribution, while recall is not. F-measure combines precision and recall as a measure of effectiveness of classification in terms of ratio of the weighted importance on either recall or precision.

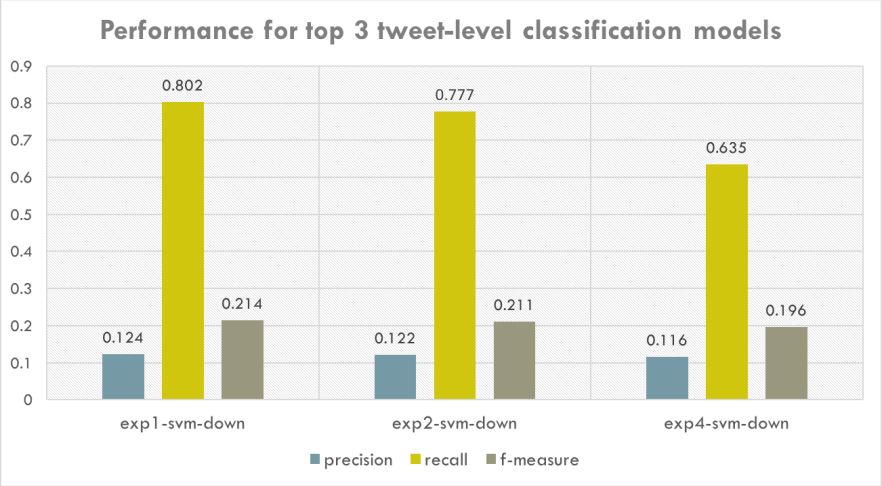
For measuring performance, we think that recall is somewhat more important for the task, therefore we aim at achieving high recall. This can be justified by keeping in mind the problem we are attempting to solve. In the context of detecting depression, a false positive (FP) is defined as a user who is predicted to have depression but does not actually suffer from depression. A false negative (FN) is defined as a user who is actually depressed but is predicted to not have depression. A classifier detecting more false positives would result in lower precision, the cost of which is that the state would need to invest more money to help users who are not actually depressed. On the other hand, a classifier detecting more false negatives would result in lower recall, the cost of which is that users suffering from depression will not get the help they need on time, which could lead to serious consequences, like suicide. So low recall could lead to loss of human life.

At the same time, we are trying to find a balance of precision and recall. A perfect recall of 1, with a very low precision (e.g., 0.2) is also not an acceptable outcome. In such cases, we look at F-measure which combines both precision and recall. In particular, we look at the precision, recall, and F-measure of the positive class, obtained on the test set.

6.1 Tweet-level Classification

Table 6.1 shows results obtained for tweet-level classification experiments. Results shown are for 20 test users. In case of tweet-level classification, the tweets belonging to the 20 test users were tested for distress. None of the classifiers performed exceptionally well on the task. The top 3 performing classifiers are indicated in bold (exp1-svm-Down, exp2-svm-Down, exp4-svm-Down), and the best classifier (exp1-svm-down) is indicated in italics. Figure 6.1 shows performance of top3 tweet-level classifiers.

Figure 6.1: Performance for top 3 tweet-level classification models



Looking at recall, it can be seen from these results that classifiers trained using balancing techniques such as SMOTE or undersampling perform much better than classifiers trained on imbalanced data. We also notice that using Bag of Words (BOW) as features does not affect performance too much. Exp1 provides the highest performance without the use of Bag of Words features, while exp2-4 which use BOW do not perform as well.

Appendix A presents results of tweet-level classifiers on training data using 10-fold cross validation. We notice that in Table A.1 recall and precision for exp2-4 is generally higher than recall for exp1. The major difference between exp1 and exp2-4 is absence of BOW in exp1. This shows that contrary to our conclusion about BOW from performance on testdata, BOW does improve performance in general, however it does not perform well on unseen data. This can be overcome by increasing the amount of training data.

Table 6.1: Tweet-level classification results

ModelName	Accuracy	Precision	Recall	F1
baseline	0.9469	1.0000	0.0111	0.0219
exp1-svm-Original	0.9337	NA	0.0000	NA
exp1-svm-SMOTE	0.7816	0.1706	0.5939	0.2650 4
exp1-svm-Down	***0.6102***	***0.1237***	***0.8020***	***0.2144***
exp2-svm-Original	0.9303	0.2222	0.0203	0.0372
exp2-svm-SMOTE	0.7711	0.1124	0.3553	0.1707
exp2-svm-Down	**0.6143**	**0.1219**	**0.7766**	**0.2107**
exp3-svm-Original	0.9303	0.2500	0.0254	0.0461
exp3-svm-SMOTE	0.7782	0.1188	0.3655	0.1793
exp3-svm-Down	0.5695	0.1010	0.6954	0.1764
exp4-svm-Original	0.8987	0.1867	0.1574	0.1708
exp4-svm-SMOTE	0.7641	0.1283	0.4416	0.1989
exp4-svm-Down	**0.6553**	**0.1161**	**0.6345**	**0.1962**

6.2 User-level Classification

Table 6.2, Table 6.3 and Table 6.4 show results obtained for the user-level classification experiments. Exp5-Exp7 (Table 6.2, Table 6.3) are trained on BellLetsTalk data , exp8-Exp11 (Table 6.2 , Table 6.3) are trained on CLPsych2015 data. Exp14-5 and Exp14-11 (Table 6.4) are an improvement of experiment 5 and experiment 11, respectively. Exp15 (Table 6.4) attempts to improve results by using a larger number of features. Figure 6.2 shows performance of top3 user-level classifiers.

Figure 6.2: Performance for top 3 user-level classification models

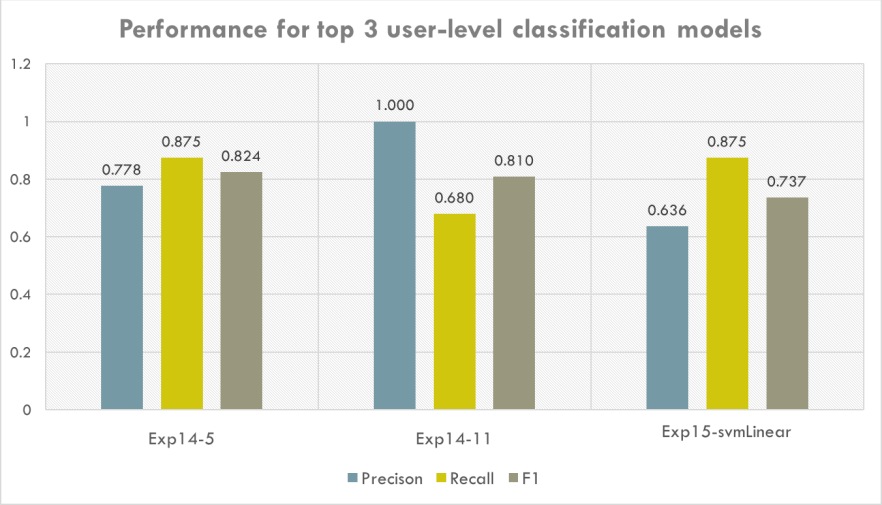


Table 6.2: User-level classification results with "self-reported" users considered as "depressed" (scenario a)

modelName	Accuracy	Precision	Recall	F1
Trained on BellLetsTalk data, tested on 20 users from BellLetsTalk data				
exp5a-svm-Orignal	0.6000	NA	0.0000	NA
exp5a-svm-SMOTE	0.6500	0.6666	0.2500	0.3636
exp6a-svm-Orignal	0.6000	NA	0.0000	NA
exp6a-svm-SMOTE	0.6000	0.5000	0.1250	0.2000
Trained on CLPsych2015 data, tested on 60 users from BellLetsTalk data				
exp7a-svm-Orignal	0.4333	0.4237	1.0000	0.5952
exp8a-rf	0.6500	0.8333	0.2000	0.3226
exp9a-rf	0.5833	0.5000	0.1200	0.1935
exp10a-rf	0.6500	0.8333	0.2000	0.3226
exp11a-rf	0.6333	0.6364	0.2800	0.3889

Table 6.3: User-level classification results with "self-reported" users considered as "not-depressed" (scenario b)

modelName	Accuracy	Precision	Recall	F1
Trained on BellLetsTalk data, tested on 20 users from BellLetsTalk data				
exp5b-svm-Orignal	0.8	NA	0	NA
exp5b-svm-SMOTE	0.8	0.5	1	0.6667
exp6b-svm-Orignal	0.8	0.5	0.5	0.5
exp6b-svm-SMOTE	0.75	0	0	NaN
Trained on CLPsych2015 data, tested on 60 users from BellLetsTalk data				
exp7b-svm-Orignal	0.2333	0.2203	1.0000	0.3611
exp8b-rf	0.8667	1.0000	0.3846	0.5556
exp9b-rf	0.7833	0.5000	0.2308	0.3158
exp10b-rf	0.8500	0.8333	0.3846	0.5263
exp11b-rf	0.8000	0.5455	0.4615	0.5000

Experiments 14 and 15 (Table 6.4) exhibit much better results in comparison to experiments 5 through 11. This shows that by removing users who self-report depression but this is not reflected in their tweets, the training process is improved. The classifier is able to find a pattern among the features of the training sample. The performance is also improved because many of the "self-reported" users which were previously being mis-classified

Table 6.4: User-level classification improved results

ModelName	Accuracy	Precision	Recall	F1
exp14-5	***0.8500***	***0.7778***	***0.8750***	***0.8235***
exp14-11	**0.8667**	**1.0000**	**0.6800**	**0.8095**
exp15-svmLinear	**0.7500**	**0.6364**	**0.8750**	**0.7368**
exp-15svmLinear-SMOTE	0.6000	0.5000	0.8750	0.6364
exp15-rf	0.6500	0.5714	0.5000	0.5333
exp15-rf-SMOTE	0.6000	0.5000	0.6250	0.5556
exp15-nb-SMOTE	0.7000	0.6000	0.7500	0.6667
exp15-nb-SMOTE	0.6500	0.5455	0.7500	0.6316
exp15-svmLinearWeights	0.7000	0.5833	0.8750	0.7000
exp15-svmLinearWeights -SMOTE	0.6000	0.5000	0.8750	0.6364
exp15-svmLinear2	0.6000	0.5000	0.8750	0.6364
exp15-svmLinear2-SMOTE	0.7000	0.5833	0.8750	0.7000
exp15-knn	0.6000	0.5000	0.7500	0.6000
exp15-knn-SMOTE	0.6500	0.5556	0.6250	0.5882

are now being labeled as "depressed" based on pattern matching rule instead the classifier making a prediction on these users.

The classification process can be explained with the following pseudo code, where $isSelfReported$ = tests if the user claimed to have been diagnosed with depression:

$if(isSelfReported$ && $percentageOfDistressTweets < 10\%)$

{ $PredictedClass$ ="depressed" }

$else$

{ $PredictedClass = PredictWithClassifier();$ }

For model "exp14-5" (Table 6.4), the high recall of 0.8750 indicates that of the users who are actually depressed, 87.50% are correctly predicted as depressed, and a corresponding high precision indicates that of the users predicted as depressed, 77.78% are predicted correctly[1].

Appendix B presents results for user-level classifiers on training data using 10-fold cross validation. We can see in Table B.1 that many classifiers perform extremely well on training data. This could be due to overfitting. This is one of the reasons why we present performance on held-out test set

[1]Experiment 14-5 was repeated using svmRadial instead of linear SVM, in order to investigate if other kernel setting work better. We found that there is a significant drop in recall when a radial kernel is used. Accuracy: 0.85; Precision: 1; Recall: 0.6250; F1: 0.7692

instead of 10-fold cross validation results in this chapter. Many classifiers that over-fit on training data perform poorly on testdata e.g., exp6a-svm-SMOTE and exp6b-svm-SMOTE provide perfect precision and recall (Table B.1) but on test data, performance drops significantly (Table 6.2 and Table 6.3 correspondingly). In comparison models that did not overfit on training data perform relatively well on test data.

In regards to exp14-11, we expected the model to out-perform other classifiers as it was trained on the a larger dataset obtained from CLPsych2015 shared task. Although, in terms of F1, it is comparable with exp14-5, the recall is lower than desired. Perhaps this is as a result of their data collection method. The data was collected using self-disclosure and verified that the self-disclosing tweets were in fact about the individual. It is unclear whether a human annotator looked at the remaining tweets for the self-disclosed users. In case the human annotator did not verify the presence of depressive tweets in the dataset, the model would be trained on a large number of non-distressed tweets and hence perform poorly on a new dataset.

Exp15-svmLinear is also comparable with exp14-5 and is considered the 3rd best classifier based on F1[2]. High recall obtained in exp15 shows that a large number of features help to improve classifier performance; however, in comparison to exp14-5 which only uses 10 features, there is no increase in performance when number of features is increased from 10 to 116.

6.3 Results on 100 Users

Predictions were made on 100users dataset using the best user-level classifier "exp14-5". The dataset provided to annotators was undersampled using "exp1-svm-Down" but the dataset used for testing was not undersampled. The results are shown in Table 6.5. The distribution of predictions can be seen as a confusion matrix in Table 6.6.

The recall on predicted users is consistent with results from 20 test users

[2]Exp14-5 and Exp15 were compared using t-test with the null hypothesis that the 2 models are statistically different. The p-value = 0.9189 obtained through paired t-test on the test data indicates that our hypothesis is false, meaning that the two models are NOT statistically different. We did not apply the t-test on the training data because the size of the data was different in the two experiments due to the resampling for balancing the data in one of the experiments.

Table 6.5: Prediction results on 100 users using classifier exp14-5

modelName	Accuracy	Precision	Recall	F1
100Users-exp14-5	0.65	0.5714	0.8182	0.6729

Table 6.6: 100Users - confusion matrix

	actual		
predictions		not-depressed	depressed
	not-depressed	29	8
	depressed	27	36

predicted earlier (Table 6.4), but precision and F1 is lower. We also note that from the predicted users, 44 users were identified as being self-reported, but only 1 of those users had less than 10% distress tweets, which means only 1 user was marked "depressed" using pattern matching rule and all other "depressed" predictions are made by the classifier.

To support our use of classifier "exp1-svm-Down" to undersample dataset for annotation we present the confusion matrix for the classifier in Table 6.7. Here low False Negatives shows that we did not remove too many "distress" tweets hence justifying its use for under-sampling. Examples in Table 4.6 also show the type of tweets that were removed. Using this method we reduced 117712 tweets to 4887 tweets, hence saving the annotator's time and effort.

Table 6.7: Exp1-svm-Down - confusion matrix

	actual		
predictions		no-distress	distress
	no-distress	1655	39
	distress	1119	158

6.4 Error Analysis

To investigate the mis-classifications, we first look at the users who were not easy to label (i.e., annotator1 and annotator2 had a disagreement) and also those users where one of the annotators made a comment. These are shown in Table 6.8. The predictions indicated in bold can be justified, even though they do not match the UserClass label (indicated by annotator3).

- Users “28246817”,“148176891”, “752005778”, “757433912” are labeled as “not depressed” but are considered at high risk of future depression. We do want the system to identify such users, and hence should not consider them as misclassified.

- User “47661600” is labeled as depressed due to the self-disclosure about suffering from depression in the past. It is possible that the tweets in our dataset, for 2015, do not indicate depression, but perhaps tweets from 2014 would.

- For the remaining users, it can be seen that the prediction corresponds to the expected label provided by at least one of the annotators.

Table 6.8: Annotator disagreements and comments

node	**Annotator1**	**Annotator2**	**Annotator3**	**comments**	**prediction**
20625590	depressed	depressed	depressed	self-reported;past	depressed
21132135	not depressed	depressed	depressed	self-reported	depressed
28246817	not depressed	not depressed	not depressed	higher risk	**depressed**
33651750	depressed	not depressed	not depressed		not depressed
47661600	depressed	depressed	depressed	self-reported;past	**not depressed**
90767819	not depressed	depressed	not depressed	self-reported;past	**depressed**
148176891	not depressed	not depressed	not depressed	higher risk	**depressed**
218566230	depressed	not depressed	not depressed		**depressed**
227916342	depressed	not depressed	not depressed		**depressed**
240443174	depressed	depressed	depressed	self-reported;past	depressed
365295508	depressed	not depressed	not depressed		**depressed**
406747719	depressed	not depressed	depressed		**not depressed**
477143152	not depressed	depressed	depressed		**not depressed**
566740897	depressed	not depressed	not depressed		**depressed**
752005778	depressed	not depressed	not depressed	higher risk	**depressed**
757433912	depressed	not depressed	not depressed	higher risk	**depressed**
865536330	depressed	not depressed	not depressed		**depressed**
983758176	not depressed	depressed	depressed	self-reported	depressed
1009070256	depressed	not depressed	not depressed		not depressed

Table 6.9 shows the remaining misclassified users. From these users, 4 can be attributed to lack of distress tweets in our dataset for self-reported users. Looking at the tweets, we noticed that many people self-reported during the BellLetsTalk campaign which took place in January. It is possible that the distress tweets for some of the self-reported users may be present in 2014 tweets.

Table 6.9: 100Users: misclassified users

node	UserClass	Prediction	comment
15881912	no	yes	
17782835	yes	no	self-reported
20691217	no	yes	
22733847	yes	no	self-reported
24389127	no	yes	
26327509	no	yes	
31976463	no	yes	
39599022	no	yes	
105141601	no	yes	
181848628	no	yes	
204926658	yes	no	self-reported
277513851	no	yes	
279366868	no	yes	
319929034	no	yes	
389830312	no	yes	
409471294	no	yes	
701054720	yes	no	
811595204	yes	no	self-reported
1534226707	no	yes	
1707744458	no	yes	
2190802568	no	yes	
2451754638	no	yes	

6.5 Comparison to Related Work

Table 3.1 from Chapter 3 provides a summary of previously existing systems developed by researchers to identify depression from social media. Many of the researchers provide the accuracy of their systems tested on different datasets. Resnik et al. (2015) and Preotiuc-Pietro et al. (2015b) report performance on the same dataset made available through CLPsych2015 shared task. We ran our best-performing User-Level classifier on the training set of CLPsych2015 shared task data and obtained a recall of 0.9167 and a precision of 0.4931, as shown in Table 6.10. These results are not comparable with those reported by Resnik et al. (2015) and Preotiuc-Pietro et al. (2015b), because the shared task uses precision at a certain recall level as the main performance measure, while we report standard precision and recall, and we selected our model to have a high recall. Secondly, the results we report in

Table 6.10 and Table 6.11 are for training users, as we were not provided with labels for test users by the organizers.

Table 6.10: User-level prediction - CLPsych 2015 users

accuracy	**precision**	**recall**	**F-measure**
0.4931	0.4950	0.9167	0.6429

Table 6.11: Confusion matrix for the users from the CLPshych 2015 dataset

	actual		
predictions		**not-depressed**	**depressed**
	not-depressed	24	27
	depressed	303	297

6.6 Summary

This chapter presented results for the experiments described in Chapter 5. We presented results for testset consisting of 20 users, instead of 10-fold cross validation results on the training set, because the models could overfit on training data, hence exhibiting unrealistic performance. For the user-level classification, we further provided performance figures on the 100Users dataset, annotated specifically for additional testing purposes.

CHAPTER 7

CONCLUSION AND FUTURE WORK

7.1 Conclusion

In this research, we collected public Twitter data for BellLetsTalk campaign. From this dataset, we identified 95 users who self-disclosed depression. For these 95 users, we collected their tweets for the year 2015. Similarly we collected 2015 tweets for users who did not self-disclose depression, these are called control users. From both the self-disclosed users and control users, we selected 30 users each who had between 100-300 tweets. Tweets from these 60 users were then hand-labeled by 2 annotators for distress level 0-3. In addition to tweets, the users themselves were also annotated as "depressed" or "not-depressed".

We then performed experiments to create 2 classifiers: a tweet-level classifier that predicts whether a given tweet indicates distress or not, and a user-level classifier that predicts whether a user suffers from depression. The experiments used a variety of features and algorithms in order to find a combination that performs well for our tasks.

The best tweet-level classifier obtained provides a recall of 0.80 with a precision of 0.12. This classifier was trained using SVM algorithm on data balanced using undersampling. It uses 7 features including polarity word counts, depression word counts, and pronoun counts.

The best user-level classifier obtained provides a recall of 0.8750 with a precision of 0.7778. This classifier was trained using SVM algorithm on user data balanced using the SMOTE algorithm. It uses 10 features including polarity word counts, depression word counts, pronoun counts, depressed tweet count, total tweet count, and percentage of depressed tweets. It also relies on a rule to remove users who self-disclose depression but have less than 10% distress tweets in order to improve performance.

The user-level classifier was further tested on 100 users to provide a recall of 0.8182 with precision of 0.5714. Many of the misclassifications for this testset can be justified by looking at Annotator disagreements.

In terms of contribution, we conclude that our user-level classifier can be used to find at-risk users from Twitter with a fairly high recall. Our tweet-level classifier on the other hand is not very accurate. It is not suggested for labeling tweets, but can be used to undersample user tweets to discard a large chunk of irrelevant tweets.

In terms of methodology, we conclude that a large number of features does not necessarily improve results. This conclusion is derived from experiments that use Bag of Word features, but do not affect the performance much compared to their counterparts that do not use the Bag of Words. Exp15 also shows that using 116 features instead of 10 features did not have a significant effect on performance. We further add that Bag of Words feature helped improve performance on the training set, but did not scale well on new unseen data.

7.2 Future Work

We have several suggestions to improve or extend our work.

7.2.1 Possible improvements

Following are some suggestions which may or may not improve classification results, but it would be a good idea to find out by performing an experiment:

- From an implementation point of view, we recommend using the training data to find the best parameters to optimize the SVM models for detecting users at-risk of depression. We also suggest trying ensemble classification techniques. Looking into word embeddings and deep learning methods is another direction of future work.
- We recommend taking negations into account while deriving features.
- Now that we have labeled data for 160 users at user-level, train on 70% (112) users instead of 40 users. This is likely to improve the results.

- In section 6.4, we discussed that many mis-classifications were not actually mis-classifications. Many results of the classifier correspond to the label of at least one of the annotators. So, for the training data, we could ask a second annotator (possibly a psychologist) to verify if the annotations are correct and re-train the model, and ask another annotator to review the 60Users dataset at user-level, to improve the training of the model.

- In addition to text derived features, perhaps community features (such as user's location, their network) can provide further insights.

- We could verify if users who "self-report" depression in our current dataset have depressive tweets in 2014.

7.2.2 Extensions

In order to extend our work, a major contribution could be to develop an application that shows a visual representation of the results e.g., an application that uploads tweets at the end of a day, identifies users that need further verification, and highlights the users' distressed tweets to allow for a human to judge if the user requires further attention.

We also recommend investigating patterns of depressive behavior over weekly or monthly basis. Perhaps, focusing on tweets around the time when user behaves differently from their norm, would provide insights which are overlooked when analyzing year-round tweets.

We came across a large number of tweets talking about anxiety compared to those talking about depression. Hence, we suggest investigating if users talking about anxiety are at a greater risk of depression or other mental illnesses. If this hypothesis is true, then users suffering from anxiety disorders can be given a higher priority when looking for at-risk users in general population. Similarly, training an unsupervised model to identify symptoms and stressors of depression may also prove to be useful.

Further, we suggest using community features provided by Twitter API to map data at population level to study trends such as investigating if a certain region contains more users at-risk of mental illness, or if users of certain age are more prone to depression.

Appendices

A Tweet-level Classification Results on Training Data

The following results report the performance of tweet-level classifiers on the training data using 10-fold cross validation.

Table A.1: Tweet-level classification results on training data

modelName	**Accuracy**	**Precision**	**Recall**	**F1**
exp1-svm-Orignal	0.902628848	NA	0	NA
exp1-svm-SMOTE	0.782774127	0.232432432	0.534635879	0.324004306
exp1-svm-Down	0.653407126	0.190905191	0.790408526	0.307532827
exp2-svm-Orignal	0.904877205	1	0.023090586	0.045138889
exp2-svm-SMOTE	0.901591145	0.497335702	0.994671403	0.663114269
exp2-svm-Down	0.75147008	0.281281281	0.998223801	0.438891058
exp3-svm-Orignal	0.905223106	1	0.026642984	0.051903114
exp3-svm-SMOTE	0.90003459	0.49339207	0.994671403	0.659599529
exp3-svm-Down	0.718955379	0.256868132	0.996447602	0.408445577
exp4-svm-Orignal	0.786219081	1	0.140319716	0.246105919
exp4-svm-SMOTE	0.9545053	0.854938272	0.98401421	0.914946325
exp4-svm-Down	0.810954064	0.568250758	0.998223801	0.724226804

B User-level Classification Results on Training Data

The following results report the performance of user-level classifiers on the training data using 10-fold cross validation.

Table B.1: User-level classification results on training data

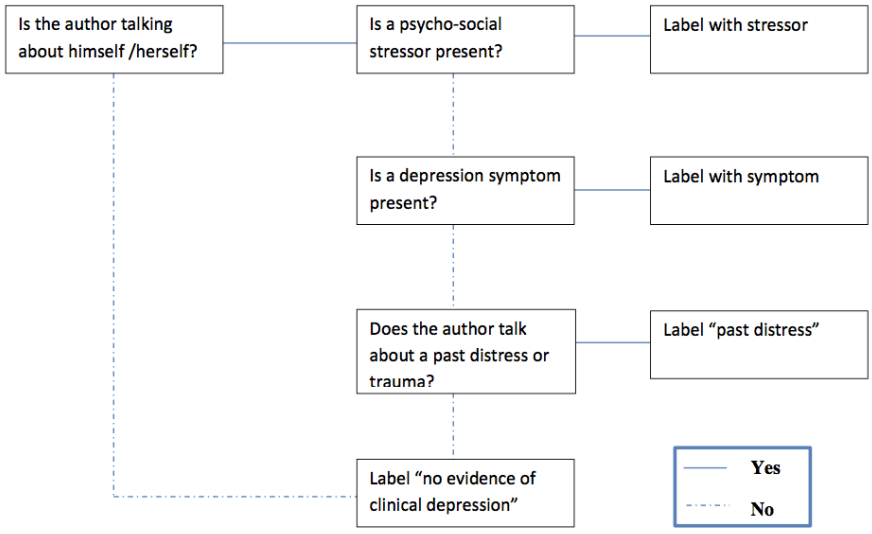
modelName	**Accuracy**	**Precision**	**Recall**	**F1**
scenario a. self-reported users considered as depressed				
exp5a-svm-Original	0.65	0.6154	0.4706	0.5333
exp5a-svm-SMOTE	0.7479	0.8621	0.4902	0.6250
exp6a-svm-Original	1	1	1	1
exp6a-svm-SMOTE	1	1	1	1
scenario b. self-reported users considered as not-depressed				
exp5b-svm-Orignal	0.775	NA	0	NA
exp5b-svm-SMOTE	0.6984	0.8333	0.3704	0.5128
exp6b-svm-Original	1	1	1	1
exp6b-svm-SMOTE	1	1	1	1
trained on CLPych2015 data				
exp7-svm-Original	1	1	1	1
exp8-rf	1	1	1	1
exp9-rf	1	1	1	1
exp10-rf	1	1	1	1
exp11-rf	1	1	1	1
improved results				
exp14-5	0.875	0.8571	0.6667	0.7500
exp14-11	1	1	1	1
additional features				
exp15-svmLinear	0.875	0.7619	1	0.8648

C Annotation Schema for Labeling Tweets

Clinical evidence of depression	Psycho-social stressors	**Problems with expected life course w.r.t. self**	• Serious medical condition • Failure to achieve important goal
		Problems with primary support group	• Death of a family member • Health problems in a family member • Serious disability of a family member • Separation/divorce/end of serious relationship
		Problems related to the social environment	• Death of friend • Death of celebrity or person of interest • Social isolation • Inadequate social support personal or romantic • Living alone • Experience of discrimination • Adjustment to lifestyle transition
		Educational problems	• Academic problems • Discord with teachers or classmates • Academic problems • Inadequate or dangerous school environment
		Occupational problems	• Firing event • Unemployment • Threat of job loss • Stressful work situation • Job dissatisfaction • Job change • Difficult relationship with boss of coworker
		Housing problems	• Homelessness • Inadequate housing • Unsafe neighborhood • Discord with neighbors or landlord
		Economic problems	• Major financial crisis • Regular difficulty in meeting financial commitments • Poverty • Welfare recipient
		Problems with access to health care	• Inadequate healthcare services • Lack of health insurance
		Problems related to legal system / crime	• Problems with police, arrest • Incarceration • Litigation • Victim of crime
		Weather	
		Media	• Music • Movie or tv • Book
		Other psychosocial and environmental problems	• Natural disaster • War • Discord with caregivers

	Depression Symptoms	**Low mood**	• Feels sad • Feels hopeless • "the blues" • "feels down"
		Anhedonia	• Loss of interest
		Weight change or change in appetite	• Increase in weight • Decrease in weight • Increase in appetite • Decrease in appetite
		Disturbed sleep	• Difficuly in falling asleep • Difficulty in staying asleep • Waking up too early • Sleeping too much
		Psychomotor agitation or retardation	• Feeling slowed down • Feeling restless or fidgety
		Fatigue or loss of energy	• Feeling tired • Insufficient energy for tasks
		Feelings of worthlessness or excessive inappropriate guilt	• Perceived burdensome • Self-esteem • Feeling worthless • Inappropriate guilt
		Diminished ability to think or concentrate, indecisiveness	• Finding concentration difficult • Indecisiveness
		Recurrent thoughts of death , suicidal ideation	• Thoughts of death • Wish to be dead • Suicidal thoughts • Non-specific suicidal thoughts • Active-suicidal ideation with any method without intent to act • Active suicidal ideation with some intent to act, without specific plan • Active suicidal ideation with specific plan and intent • Completed suicide
No evidence of clinical depression			• Political stance or personal opinion • Unsubstantiated claim or fact • Inspirational statement or advice

Annotation schema for labeling tweets



NOTE: label with all possible symptoms/stressors, separated by semicolon, on two lines. Example:

Stressor: exam; loneliness

Symptom: lack of sleep

Example:

"Hate it when you say/see/do something & it makes you feel so annoying, & like you want to drop of the face of the earth #anxietyproblems"

Stressor: Feelings of worthlessness or inappropriate guilt

#BellLetsTalk about how scared I was to talk about my depression & anxiety until I've seen how many people support days like this.

Label: Past Distress

D Other Experiments

The following results present performance of exp5 and exp6 performed using algorithms other than svmLinear.

Table D.1: Some other classifiers tested for exp5

model	**classifier**	**Accuracy**	**Precision**	**Recall**	**F1**
original - test	svmlinear	0.8	na	0	na
smote - test	svmlinear	0.75	0.4444	1	0.6154
original - test	nb	0.75	0.4	0.5	0.4444
smote - test	nb	0.8	0.5	0.75	0.6
original - test	cforest	0.8	na	0	na
smote - test	cforest	0.85	0.6	0.75	0.6667
original - test	knn	0.75	0	0	NaN
smote - test	knn	0.65	0.2	0.25	0.2222
original - test	svmLinearWeights	0.9	0.75	0.75	0.75
smote - test	svmLinearWeights	0.55	0.3077	1	0.4706
original - test	rf	0.8	na	0	na
smote - test	rf	0.8	0.5	0.75	0.6
original - test	svmLinear2	0.9	1	0.5	0.6667
smote - test	svmLinear2	0.85	0.5714	1	0.7273
original - test	svmPoly	0.8	na	0	na
smote - test	svmPoly	0.75	0.4	0.5	0.4444
original - test	svmRadial	0.8	na	0	na
smote - test	svmRadial	0.8	0.5	0.5	0.5

Table D.2: Some other classifiers tested for exp6

model	**classifier**	**Accuracy**	**Precision**	**Recall**	**F1**
original - test	svmlinear	0.6	0.5	0.375	0.4286
smote - test	svmlinear	0.7	0.75	0.375	0.5
original - test	nb				
smote - test	nb				
original - test	cforest	0.6	0.5	0.125	0.2
smote - test	cforest	0.6	na	0	na
original - test	knn	0.55	0.4	0.25	0.3077
smote - test	knn	0.5	0.375	0.375	0.375
original - test	svmLinearWeights	0.7	0.75	0.375	0.5
smote - test	svmLinearWeights	0.65	0.6	0.375	0.4615
original - test	rf	0.75	0.8	0.5	0.6154
smote - test	rf	0.7	0.6667	0.5	0.5714
original - test	svmLinear2	0.7	0.75	0.375	0.5
smote - test	svmLinear2	0.65	0.6	0.375	0.4615
original - test	svmPoly	0.6	na	0	na
smote - test	svmPoly	0.65	0.5714	0.5	0.5333
original - test	svmRadial	0.6	na	0	na
smote - test	svmRadial	0.4	0.4	1	0.5714

REFERENCES

Achrekar, H., Gandhe, A., Lazarus, R., yu, S.-H., and Liu, B. (2012). Twitter improves seasonal influenza prediction. *Proceedings of the International Conference on Health Informatics*.

André, P., Bernstein, M., and Luther, K. (2012). Who gives a tweet?: evaluating microblog content value. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 471–474. ACM.

APA (2013). *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub.

Atefeh, F. and Khreich, W. (2013). A survey of techniques for event detection in twitter. *Computational Intelligence*, 31(1):132164.

Balani, S. and De Choudhury, M. (2015). Detecting and characterizing mental health related self-disclosure in social media. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, pages 1373–1378. ACM.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Choudhury, M. D. (2013). Role of social media in tackling challenges in mental health. *Proceedings of the 2nd international workshop on Socially-aware multimedia - SAM '13*.

Choudhury, M. D. (2014). Can social media help us reason about mental health? *Proceedings of the 23rd International Conference on World Wide Web - WWW '14 Companion*.

Choudhury, M. D. (2015). Social media for mental illness risk assessment, prevention and support. *Proceedings of the 1st ACM Workshop on Social Media World Sensors - SIdEWayS '15*.

Choudhury, M. D., Counts, S., and Horvitz, E. (2013). Social media as a measurement tool of depression in populations. *Proceedings of the 5th Annual ACM Web Science Conference on - WebSci '13*.

CMHA (2016). Facts about mental illness.

Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., and Mitchell, M. (2015a). Clpsych 2015 shared task: Depression and ptsd on twitter. *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*.

Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., and Mitchell, M. (2015b). Clpsych 2015 shared task: Depression and ptsd on twitter. *NAACL HLT 2015*, page 31.

Coppersmith, G., Harman, C., and Dredze, M. (2014). Measuring post traumatic stress disorder in twitter. In *ICWSM*.

Cristianini, N. and Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.

De Choudhury, M., Gamon, M., Counts, S., and Horvitz, E. (2013). Predicting depression via social media. In *ICWSM*, page 2.

De Marneffe, M.-C. and Manning, C. D. (2008). Stanford typed dependencies manual. Technical report, Technical report, Stanford University.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer.

Farnadi, G., Zoghbi, S., Moens, M.-F., and De Cock, M. (2013). Recognising personality traits using facebook status updates. In *Proceedings of the workshop on computational personality recognition (WCPR13) at the 7th international AAAI conference on weblogs and social media (ICWSM13)*. AAAI.

Farzindar, A. and Inkpen, D. (2015). Natural language processing for social media. *Synthesis Lectures on Human Language Technologies*, 8(2):1–166.

Gerber, M. S. (2014). Predicting crime using twitter and kernel density estimation. *Decision Support Systems*, 61:115125.

Gottschlich, J., Heimbach, I., Hinz, O., et al. (2013). The value of users' facebook profile data-generating product recommendations for online social shopping sites. In *ECIS*, page 117.

Han, J., Pei, J., and Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

He, H. and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284.

Hochman, N. and Schwartz, R. (2012). Visualizing instagram: Tracing cultural visual rhythms. In *Proceedings of the Workshop on Social Media Visualization (SocMedVis) in conjunction with the Sixth International AAAI Conference on Weblogs and Social Media (ICWSM–12)*, pages 6–9.

Honkela, T. and Hyvarinen, A. (2004). Linguistic feature extraction using independent component analysis. In *Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on*, volume 1, pages 279–284. IEEE.

Jashinsky, J., Burton, S. H., Hanson, C. L., West, J., Giraud-Carrier, C., Barnes, M. D., and Argyle, T. (2014). Tracking suicide risk factors through twitter in the us. *Crisis*, 35(1):5159.

Joinson, A. N. and Paine, C. B. (2009). Self-disclosure, privacy and the internet. *Oxford Handbooks Online*.

Kessler, R., Berglund, P., Demler, O., and et al (2003). The epidemiology of major depressive disorder: Results from the national comorbidity survey replication (ncs-r). *JAMA*, 289(23):3095–3105.

Kuhn, M., from Jed Wing, C., Weston, S., Williams, A., Keefer, C., and Engelhardt, A. (2012). *caret: Classification and Regression Training*. R package version 5.15-044.

Liu, X.-Y., Wu, J., and Zhou, Z.-H. (2009). Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):539–550.

Maigrot, C., Bringay, S., and Azé, J. (2016). Concept drift vs suicide: comment l'un peut prévenir l'autre? In *16ème Journées Francophones Extraction et Gestion des Connaissances, EGC 2016, 18-22 Janvier 2016, Reims, France*, pages 219–230.

Mani, I. and Zhang, I. (2003). knn approach to unbalanced data distributions: a case study involving information extraction. In *Proceedings of workshop on learning from imbalanced datasets*.

Manning, C. D. and Schütze, H. (1999). *Foundations of statistical natural language processing*, volume 999. MIT Press.

McCabe, R., Howes, C., and Purver, M. (2014). Linguistic indicators of severity and profess in online text-based therapy for depression. Association for Computational Linguistics.

MHCC (2016).

Milne, D. N., Pink, G., Hachey, B., and Calvo, R. A. (2016). Clpsych 2016 shared task: Triaging content in online peer-support forums. In *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*, pages 118–127.

Mowery, D., Bryan, C., and Conway, M. (2015). Towards developing an annotation scheme for depressive disorder symptoms: A preliminary study using twitter data. *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*.

Nielsen, F. Å. (2011). Afinn.

Park, M., Cha, C., and Cha, M. (2012). Depressive moods of users portrayed in twitter. In *Proceedings of the ACM SIGKDD Workshop on healthcare informatics (HI-KDD)*, pages 1–8.

Pavalanathan, U. and De Choudhury, M. (2015). Identity management and mental health discourse in social media. In *Proceedings of the 24th International Conference on World Wide Web*, pages 315–321. ACM.

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., and Booth, R. J. (2007). The development and psychometric properties of liwc2007. *UT Faculty/Researcher Works*.

Preotiuc-Pietro, D., Eichstaedt, J., Park, G., Sap, M., Smith, L., Tobolsky, V., Schwartz, H. A., and Ungar, L. (2015a). The role of personality, age, and gender in tweeting about mental illness. *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*.

Preotiuc-Pietro, D., Schwartz, M. S. H. A., and Ungar, L. (2015b). Mental illness detection at the world well-being project for the clpsych 2015 shared task. *NAACL HLT 2015*, page 40.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Raskutti, B. and Kowalczyk, A. (2004). Extreme re-balancing for svms: a case study. *ACM Sigkdd Explorations Newsletter*, 6(1):60–69.

Reece, A. G. and Danforth, C. M. (2016). Instagram photos reveal predictive markers of depression. *arXiv preprint arXiv:1608.03282*.

Resnik, P., Armstrong, W., Claudino, L., and Nguyen, T. (2015). The university of maryland clpsych 2015 shared task system.

RStudio Team (2015). *RStudio: Integrated Development Environment for R.* RStudio, Inc., Boston, MA.

Schwartz, H. A., Eichstaedt, J., Kern, M. L., Park, G., Sap, M., Stillwell, D., Kosinski, M., and Ungar, L. (2014). 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*, pages 118–125.

Taylor, A., Marcus, M., and Santorini, B. (2003). The penn treebank: an overview. In *Treebanks*, pages 5–22. Springer.

Traud, A. L., Mucha, P. J., and Porter, M. A. (2012). Social structure of facebook networks. *Physica A: Statistical Mechanics and its Applications*, 391(16):4165–4180.

Tsugawa, S., Kikuchi, Y., Kishino, F., Nakajima, K., Itoh, Y., and Ohsaki, H. (2015). Recognizing depression from twitter activity. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 3187–3196, New York, NY, USA. ACM.

Wang, N., Kosinski, M., Stillwell, D. J., and Rust, J. (2014). Can well-being be measured using facebook status updates? validation of facebook's gross national happiness index. *Social Indicators Research*, 115(1):483–491.

Webb, A. R. (2003). *Statistical pattern recognition.* John Wiley & Sons.

WHO (2004). Promoting mental health: Concepts, emerging evidence, practice: Summary report.

WHO (2016). Mental health: a state of well-being.

Xie, Y., Chen, Z., Cheng, Y., Zhang, K., Agrawal, A., Liao, W.-K., and Choudhary, A. (2013). Detecting and tracking disease outbreaks by mining social media data. *Dimensions*, 17(16):16–70.

Xu, J.-M., Jun, K.-S., Zhu, X., and Bellmore, A. (2012). Learning from bullying traces in social media. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 656–666, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yu, L.-C. and Ho, C.-Y. (2014). Identifying emotion labels from psychiatric social texts using independent component analysis. In *COLING*, pages 837–847.