University of Brasília

Institute of Exact Sciences
Department of Computer Science

UnB Sense: a web application to probe for signs of depression from user profiles on social media

Otto Kristian von Sperling

Undergraduate thesis submitted as a partial requirement to receive
the degree of Bachelor of Science in Computer Science

Supervisor
Prof. Dr. Marcelo Ladeira

Brasília
2019

University of Brasília

Institute of Exact Sciences
Department of Computer Science

UnB Sense: a web application to probe for signs of depression from user profiles on social media

Otto Kristian von Sperling

Undergraduate thesis submitted as a partial requirement
to complete the Bachelor of Science in Computer Science

Prof. Dr. Marcelo Ladeira (Supervisor)
CIC/UnB

Prof. Dr. Jorge Henrique Cabral Fernandes
CIC/UnB

Prof[a]. Dr[a]. Germana Menezes da Nobrega
CIC/UnB

Prof[a]. Dr[a]. Elisa Dias Becker Reifschneider
CAEP/UnB

Prof. Dr. Edison Ishikawa
Coordinator of the Bachelor of Science in Computer Science

Brasília, 05 of December of 2019

Dedication

This work is dedicated to all of those who have lost faith in technology. Hopefully, we can draw a better picture that tools do not have will, humans do. And if humans want to be kind to one another, we can. And if humans want to make tools that help us all be a little more human, we can. So if you are not satisfied with the way things are going, speak up, hold your ground, and put your money where your mouth is. Maybe you are the one who has to take the first step.

Acknowledgements

First of all, I would like to acknowledge and thank my mother, Claudia de Melo Cardoso, for bearing, caring, nurturing and providing for me and my sibilings even when life was at its hardest. She is a warrior of a woman, independent and strong. I also wouldn't have been able to finish this work and this stage in life if it weren't for my partner and love, Paula. Her companionship gives me the sobriety that I need to be my best self, her love gives me drive to always strive for new horizons. Last but not least, for all the hard and honest questions, for the misunderstood sense of humor, for the comments that help me keep my feet on the ground and look at things with more pragmaticality. For all that and more, I am thankful to my supervisor, Dr. Marcelo Ladeira.

Abstract

Research on computerized models that help detect, study and understand signs of mental health disorders from social media has been thriving since the mid-2000s for English speakers. In Brazil, this area of research shows promising results, in addition to a variety of niches that still need exploring. Thus, we construct a large corpus from 2941 users (1486 depressive, 1455 non-depressive), and induce machine learning models to identify signs of depression from our Twitter corpus. In order to achieve our goal, we extract features by measuring linguistic style, behavioral patterns, and affect from users' public tweets and metadata. Resulting models successfully distinguish between depressive and non-depressive classes with performance scores comparable to results in the literature ($F_1 = 0.798$, precision $= 0.806$, recall $= 0.807$). Last but not least, we develop an online platform to allow Twitter users to probe their profiles for signs of depression. By doing so, we hope to empower users to better understand their signals and to steer them to seek professional assistance whenever needed.

Keywords: data mining, depression, machine learning, online application, Twitter

Contents

List of Figures

List of Tables

Acronyms

API Application Programming Interface.

DM Data Mining.

EDA Exploratory Data Analysis.

ETL Extract Transform Load.

GCP Google Cloud Platform.

ML Machine Learning.

MLP Multilayer Perceptron.

MVC Model View Controller.

PCA Principal Component Analysis.

PWA Progressive Web App.

REST Representational State Transfer.

RF Random Forest.

SVM Support Vector Machine.

WMM Weighted Moving Mean.

Chapter 1

Introduction

> And once the storm is over, you won't remember how you made it through, how you managed to survive. You won't even be sure, whether the storm is really over. But one thing is certain. When you come out of the storm, you won't be the same person who walked in. That's what this storm's all about.
>
> – Haruki Murakami, 2016

Mental illness has become a major cause of disability worldwide, with an estimate of 300 million people suffering from depression [1]. The impact of depression goes beyond the individual sphere, as it leads to limitations in psychosocial, parental and professional functioning, it is also to blame for an increase in the number of disability days (i.e. days not worked). Overall, depression has been estimated to be the cause of 400 million disability days per year, which in turn results in an aggregate cost of U$ 210.5 billion in the United States alone [2]. By and large, the pain that comes with depression does not solely affect those who suffer, it stains all of us; thus, all of us should care. Some of us should even try to tackle the problem. Fortunately, technology has been shown to be a valuable ally in solving many of the predicaments of modern society. It should be no different for the problem of effectively preventing and treating depression and other mental illness. After all, there is already growing interest when it comes to mental health from the scientific community and society at large.

The ubiquity of social media has provided researchers with a treasure trove of data. By combining the power of Data Mining (DM) with what is already known about mental health from rich bodies of research in psychology, psychiatry, neuroscience and sociolinguistics, the relationship that many have with their social profiles can lead to further insights into human behaviour. It becomes possible to forecast the onset of depression [3] and post-traumatic stress disorder [4], among other mental illnesses. Despite its being too early, by some measures, for strong claims about the power of Data Mining in mental

health, researchers continuously add to a plethora of methods that can help identify and forecast mental illness with satisfactory performance and better explainability and interpretability. For instance, Hidden Markov Models [5] and word shift graphs [6] have been successfully employed to render more explainability of models' decisions [7].

1.1 Problem Definition

Although there have been growing efforts to treat and prevent mental illness through the use of technology as discussed later in Section 2.1.1, it is still a fairly recent field of research in Brazil. Due to most research being carried out in English, often times there is a gap between the latest findings and how they translate to other cultures and languages. With that in mind, we pose 3 questions that guide all of this work

1. What, if any, are the signals closely related to depression in the literature that can be extracted from social media profiles?
2. How different are the signals for users who report having been diagnosed with depression and users who have not?
3. How well can we classify users as belonging to each class via Machine Learning (ML) classifiers?

In short, there is still the need to support the claim that social media can be used to study mental illness, depression in particular. Many efforts have achieved strong results that suggest it to be possible in the United States. Nevertheless, the same kind of exploratory work has to be done in Brazil independently. We believe our cultures, our people, differ enough for a direct jump onto the state of the art to become rather reckless. Instead, research must be conducted to establish solid enough background so that future results are reliable and reproducible. This belief leads us towards a set of clear objectives as follows.

1.2 Objectives

In light of the aforementioned shortcomings, we employ well-established methods of data collection and feature extraction to induce machine learning classifiers that are able to distinguish between the signals of allegedly depressive and non-depressive users, in addition to applying statistical analysis in order to select truly significant features. Our results suggest that findings in the literature can be replicated and translated to the Brazilian culture. In short, the present work describes our efforts to detect signs correlated to

depression on Twitter as well as the development of our application to help users better understand their signals. Our main objectives with this work are:

1. Construction a data set with messages in Brazilian Portuguese and their respective metadata from Twitter via the development of a data pipeline that enables us to both retrieve new users and update the current data set.

2. Extraction of features from user linguistic style and behavioral patterns to induce machine learning classifiers that can distinguish between both depressive (positive) and non-depressive (control) classes.

3. Analysis of the significance of features derived from our work and other research efforts in Brazil. By doing so, we replicate some findings in the literature and refute others.

4. Development and deployment of a web application to help users understand their signals and prone them to seek professional help if needed.

By completing these 4 goals, we believe that we get a closer to the conclusion that data mining can be very helpful to study depression, and perhaps other mental illnesses. This document is a report on the work that has been done towards achieving these 4 goals.

1.3 Document Structure

For the sake of comprehension, we go over the structure of this dissertation.

- Chapter 2 provides an overview of the theoretical concepts and research milestones in the intersection of mental health, sociolinguistics, data mining and machine learning. Furthermore, we discuss the concepts, patterns and frameworks that underlie our application

- Chapter 3 discusses the methods we employed for the research part of this work; from data collection, to feature extraction, to machine learning models and statistical analysis.

- Chapter 4 presents our research results with information on the statistical significance of our features and the performance scores of our models, in addition to further considerations on our results.

- Chapter 5 deals with the process of development of the application, the software engineering side of our work.

- Chapter 6 makes the case for employing data mining methodos to research depression, and mental illness at large in Brazil. Furthermore, we set the ground for future work to extend our data set and improve our methods.

In the next chapter, we delve into the concepts that surround our research. The reader who is not familiar with concepts such as data mining, hypothesis test, Machine Learning (ML), web development and Representational State Transfer (REST) Application Programming Interface (API) can skip Sections 2.1.3 to 2.2.4 without much loss of comprehension. We take the chance to note that any material that is not of our own authorship has the appropriate reference to the source, otherwise they are ours.

Chapter 2

Theoretical Foundations

> The problem which I wish to examine afresh in this lecture, and which I hope not only to examine but to solve, may perhaps be described as an aspect of the old quarrel between the British and the Continental schools of philosophy – the quarrel between the classical empiricism of Bacon, Locke, Berkeley, Hume, and Mill, and the classical rationalism or intellectualism of Descartes. Spinoza, and Leibniz. In this quarrel the British school insisted that the ultimate source of all knowledge was observation, while the Continental school insisted that it was the intellectual intuition of clear and distinct ideas.
>
> – Karl Popper, 1962, p.04

2.1 Research

Science is a systematic approach to solving the problem of knowledge expansion and continuity. Nevertheless, as hinted by Popper, it does not mean that science itself is simply defined. Observation and deduction, two sources of knowledge, unfortunately often juxtaposed. There is little point in arguing whether empiricism or rationalism is the real science. The really good question here lies not in "what", but in "what for". What is science for?

In this chapter, we discuss the science that has been carried out over the last fifteen years towards understanding how computers can help detect, prevent and treat mental illness at large. Then we move on to give a broad overview of the concepts, definitions, algorithms, design patterns, frameworks and areas of research that underlie our work. Finally, we start the first section of this chapter with a survey of the background literature.

2.1.1 Related Work

Since the early 2000s, there have been rising efforts to leverage the power of technology to aid in understanding and preventing mental illness. From computerized analysis of written

texts that revealed predictive cues about neurotic tendencies and psychiatric disorders [8], to support for the claim that negative (cognitive) processing biases in resolving ambiguous verbal information can predict subsequent depression [9], much has been accomplished in the past couple of decades. Nevertheless, the social media boom in the early 2010s brought with it an ever growing flux of data that enabled researchers to derive further insights from signals correlated to depression and other mental illness. Through the use of Data Mining (DM) techniques, research showed that patterns of behaviour can be matched to real world events [10], and that symptomatic signals of major depressive disorder could be observed from status updates on Facebook [11].

When it comes to Twitter, Park and colleagues [12] found evidence that people post about their depression and treatment on the platform, and De Choudhury and colleagues [3] induced classifiers (precision = 0.742, recall = 0.629, F_1 = 0.681) to estimate the risk of depression before its onset by measuring behavioral attributes related to social engagement, emotion, linguistic styles, social network, and mentions of antidepressant medication. Coppersmith and colleagues [4] proposed heuristics to automate parts of the data set construction, which yielded, for depression alone, a data set much larger than what had been previously achieved, in addition to expanding the scope to other mental illness.

Resnik and colleagues [13] applied a variety of Supervised Topic Models on the data set created Coppersmith [4] and achieved promising results (AUC = 0.860) when classifying depressive versus non-depressive groups. Following research questions the methods employed by De Choudhury and Coppersmith [3, 4] and argues that more meticulous methods are needed to support stronger claims that Twitter data indeed enables detecting and forecasting depression, as argued by Reece and colleagues [7]. Interpretability of models is yet another strong argument made by Reece, and word shift graphs together with Hidden Markov Models are employed (precision = 0.852, recall = 0.518, F1 = 0.644) as an alternative that seems to identify signs not fully captured by either Linguistic Inquiry Word Count (LIWC) [14] or Affective Norms for English Words (ANEW) [15]. Having said all of that, claiming that the findings of Reece deem invalid the methods of De Choudhury and Coppersmith is open for debate but it's rather unlikely. Conversely, it is to be seen as an iteration upon the methods so as to yield more robust claims of whether social media truly captures some of the nature of the human mind.

The one strong argument in this work is that in addition to steering research towards prediction of the chance of mental illness in research data sets, comparable efforts should go into bridging the gap towards real-world tools that can aid practicing physicians, psychologists and care-takers to better understand their patients and the patterns they share. Thus, with the present work, we hope to take our first steps towards the goal of

detecting those who are in need now rather than later, and attract more research to this challenge in our home country by showing that, although people and cultures differ in countless ways, leveraging technology in favor of well-being and mental health is truly only a matter of grit and minds.

Selection Criteria

With the growing amount of accumulated knowledge in research databases, we must impose limits to what constitutes a reasonable corpus of literature to base one's own research. Throughout the first six months of our work, there have been three iterations of gathering data about related research, which were necessary in order to define our research questions and narrow the scope of our own exploration. To survey the literature, we took advantage of the *DBLP Computer Science Bibliography*[1] repository for research articles and performed queries with tokens related to DM and mental illness in both English and Portuguese. Due to the lack of results in Brazil, we also took advantage of the Google Scholar[2] search engine to query for tokens in Portuguese.

By and large, a great percentage of Brazilian research in the field of DM and mental health has taken place in 2018, which entails that many still have not been through peer-review. Nevertheless, we attempt to aggregate two Brazilian research efforts into our own in order to expand our selection of features and also contribute with validation of their results. The first is the depression lexicon proposed by Nascimento and colleagues [16], the second is the lexicon of activation and valence of words constructed by Kristensen and colleagues [17]. The former failed to show statistical significance in its features, while the latter achieved statistical significance in all of them. We discuss in more detail their features in Chapter 3 and their contribution in Chapter 4.

Lastly, we discuss our selection criteria. The first watershed that must be emphasized is the advent and wide adoption of social media in the early 2010s as mentioned previously. Therefore, we select some of the literature from before the early 2010s as a means to draw a fuller picture of the development of the field. Nonetheless, we focus on the research that strives to establish a link between real-world suffering to the digital breadcrumbs left by users on social media. We justify such decision, juxtaposed to going straight into more recent work that employs unsupervised learning and extends feature extraction, with the argument that it is still necessary to show that such methodologies can indeed be employed in Brazil without loss of integrity.

With that in mind, we sought research that was conducted between 2011 and 2017, and that took advantage of DM to explore whether or not signals extracted from social

[1] https://dblp.dagstuhl.de/
[2] https://scholar.google.com/

media can correlate to mental illness. So as to narrow our scope, we chose depression as the mental illness to be studied due to its being among the most epidemic in Brazil, afflicting over 11% of the population[1]. Finally, we were able to reach a reasonable number of articles to manually review – 21 articles to be more precise. From those, we selected the work of De Choudhury and colleagues [3] and Kristensen and colleagues [17] to draw inspiration for our feature extraction and result validation methods, as well as Coppersmith and colleagues [4] for our data collection and processing.

2.1.2 Hypothesis Test

The main purpose of statistics is to test a hypothesis. A hypothesis is an educated guess about relationships, often correlation, among observations. It needs to be testable, either by experiment or observation, otherwise it is deemed unfalsifiable and worthless for the purpose of science. Testing is done by computing the likelihood that an observed property happened by chance. That is where we introduce the concept of the null hypothesis.

In order to test a hypothesis, really there must be two. The first is called null hypothesis, and is usually the assumption of the status quo. For instance, if testing the effect of new medicine, the null hypothesis should be that there is no significant difference between the group who received the medicine and the control group. Stating null hypothesis is a great way to think critically and enables us to work with observations as if we were conducting proof by contradiction[3]. Still, where is the contradiction? This is where the alternative hypothesis takes place.

An alternative hypothesis is the consequence we get if the null hypothesis is rejected by our test. In the previous example of medicine, if enough people in the group that was given the medicine reacted to treatment with enough expressivity, then we reject the (null) hypothesis that there is no difference between groups, and accept the (alternative) hypothesis that indeed there is. The decider for rejecting or accepting the null hypothesis that is often used is called p-value.

The p-value is a widely accepted measure to represent how surprising observations are. That is not to say that it is infallible. Much of the blunder in science can be attributed to misuse of the p-value based on overlooked assumptions [18], lack of training in statistics [19] or, seldom, ill intent. Nonetheless, when the appropriate corrections are put in place, the power of the p-value rises.

Broadly speaking, in order to proceed with statistical analysis, one must first arrive at a sound hypothesis to test. The process of extracting knowledge from data to construct robust hypothesis is known as DM, which we explore next.

[3]This comment should be taken lightly and not as serious claim

2.1.3 Data Mining (DM)

In short, Data Mining is the process of applying a collection of techniques from computer science, statistics and mathematics to large amounts of data in order to:

- find patterns, anomalies, limitations and relationship among variables.
- discover models for the data, so as to enhance understanding and forecasting.
- test and validate hypothesis in order to yield robust insights from data.

The broad aim of extracting knowledge in databases boils down to being able to study, understand and forecast a wide range of systems and phenomena, both natural and man-made.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis is a strategy of data analysis that emphasizes maintaining an open mind to alternative possibilities. It is a philosophy or an attitude about how data analysis should be carried out, rather than being a fixed set of techniques. One benefit is that it minimizes prior assumptions and thus allows the data to guide the choice of appropriate models. Traditionally, EDA comprises residual analysis, data re-expression, resistant procedures, and data visualization. With the advance of highpower computing, the alternate taxonomy is goal oriented, namely, clustering, variable screening, and pattern recognition, many of which are supported by machine learning algorithms.

Feature Engineering and Data Split

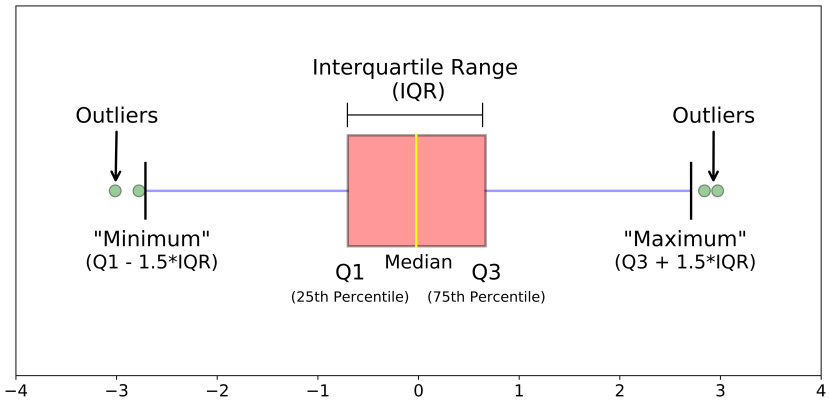
After having gathered knowledge about the data set, the next step is to transform the raw data into features that capture, to some extent, the underlying phenomena. Such features are used as input for machine learning algorithms to discover a good model for the data. Nevertheless, before selecting and learning models, it is critical to have a strategy to evaluate their performance that does not involve the same data used for training. Training and testing on the same data is likely to yield models that have memorized the data (overfitting) rather than learn the intended (or unintended) property that enables generalization to previously unseen data.

Common strategies include creating a "holdout" data set or performing cross-validation. The effect is that only a portion of the data is used to train models, so that there can be unbiased data with which to tune hyperparameters and evaluate the final model.

Machine Learning (ML)

There are different approaches to discovering properties of data sets. Machine Learning is one of them. Machine Learning is a field of DM that focuses on designing supervised and unsupervised learning algorithms that can discover models and make predictions on the data. However, DM can use other techniques besides, or in conjunction with, machine learning.

Figure 2.1: Components of a box plot



Source: https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

By and large, machine learning can be structured into five classes of tasks, namely **i)** anomaly detection, **ii)** association rule learning, **iii)** clustering, **iv)** classification and **v)** regression. A brief explanation of each follows.

i Anomaly detection is the identification of items, events or observations which do not conform to an expected pattern of variables in a data set, known as outliers. Outliers can be easily visualized via box plot, as shown in Figure 2.1 and Figure A.1

ii Association rule learning is the task of searching for relationships between seemingly independent variables in a data set via association rules [20]

iii Culstering is a method of unsupervised learning to group observation base on similarities that are not explicitly defined. K-Means is likely the most established clustering algorithm, with linear time complexity $\mathcal{O}(n)$. An example of applying clustering to a data set is presented in Figure A.2

iv Classification is the problem of identifying to which of a set of categories a new observation belongs, on the basis of a training set of data containing observations whose category membership is known.

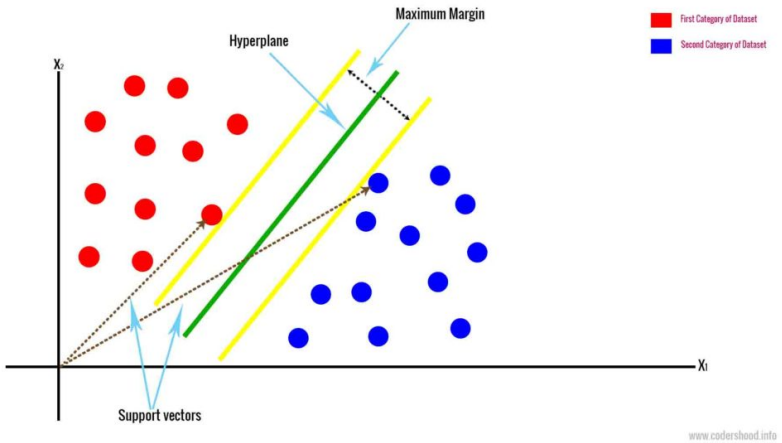
v Regression is an attempt to find a function which models the data with the least error in order to estimate the relationships among observations.

Supervised Learning Classifiers

In the present work, we focus on supervised learning algorithms to induce classifiers from our data. The decision to exclude unsupervised learning algorithms, except Principal Component Analysis (PCA) for feature selection (i.e., dimensionality reduction), is justified by the need for interpretable decisions due to the sensitive nature of working with mental illness.

- Random Forest (RF) is a collection of many decision tree classifiers where each tree receives the same input vector to classify. The forest chooses its classification based on the mean prediction of the individual trees. It is a witty way to solve the tendency decision trees have of overfitting train data.
- Support Vector Machine (SVM) is best explained as a technique to learn how to distinct observations of 2 or more classes by augmenting the dimensionality of the data and finding hyperplanes that split observation with maximum margin. The simple 2-dimensional example in Figure 2.2 helps us better understand the concept.
- Multilayer Perceptron (MLP) is an artificial network of perceptrons (neurons) where, conceptually, each perceptron draws a line to split the data, then feeds information to the next layer of perceptrons until the last (output) layer is compared to the expected result. The error is propagated backwards to adjust the contribution (weight) of each perceptron to the final answer.

Figure 2.2: SVM maximum margin and support vectors



Source: shorturl.at/wBEMT

Validation and Hyperparameter Tuning

Once the classifier is trained, it is a good idea to evaluate its performance on what is a partition of the data called the validation set. We approach the topic of spliting the data and the reasons for doing so in Section 3.3.1. The point of validation is to adjust the classifier's training parameters (e.g., loss function) in order boost performance. These training parameter are also known as hyperparameter.

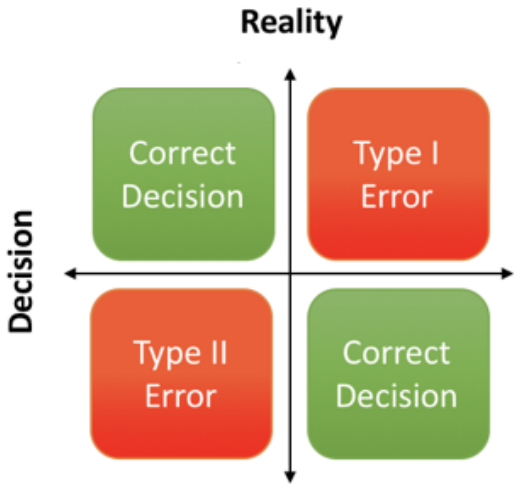
There is a variety of algorithms to help automate parts of the process of adjusting hyperparameters. We employ 2 simple algorithms, known as GridSearch and RandomSearch. We choose to use simple algorithms in this part of our work since it really is a study in DM and not particularly in machine learning. Hopefully, the distinction between the two has been made clear in previous sections.

Assessment

After finishing the process of hyperparameter tuning, we should be confident that our classifier has reached a state of good performance. But our own confidence is not enough to convince anyone that our model is indeed good. Instead, we must conduct tests on the test set, the part of the data that has not been used for learning or validation. These tests are summaries of our classifier's performance, and are usually conducted over a number of randomized runs so that confidence intervals can be reported along with the metrics. The most commonly used metrics are precision, recall, F_1 measure, specificity and accuracy, and can all be derived from a confusion matrix (Figure 2.3 and Figure A.3). We briefly define the performance metrics we utilize:

- *precision* as the rate of observations correctly classified as positive over all the positively classified observations. It gives a sense of how often the classifier tends to be right when it classifies observations as positive.
- *recall* as the proportion of actually positive observations that are correctly identified as such. It gives a sense of how nuanced the classifier is.
- *specificity* as the rate of true negatives over all observations classified as negative. It gives a sense of how often the classifier tends to be right when it classifies observations as negative.
- F_1 as a measure of a test's accuracy, by balancing precision and recall. It gives an overall idea of how certain and effective our classifier is.

Figure 2.3: Simple depiction of a confusion matrix



Source: https://towardsdatascience.com/precision-vs-recall-386cf9f89488

Summarization

In order to draw conclusions that lead to decision based on data, DM must providing compact representations of the data set, including visualization and report generation. This task is usually achieved with the implementation of dashboards.

2.2 Web Application

In this section, we provide a general overview of the concepts, patterns and frameworks used throughout the development of our web application. For the sake of order, we start at the very top of the abstraction stack with new standards for web applications, through an API framework, to finally end in the cloud. As a reminder, we do not aim to give a full picture of any of these concepts. Instead, this is a brief introduction, in addition to comments on the problems each tries to solve.

In short, our application is of two parts. To the user, it is a Progressive Web App (PWA) client developed with Vue.Js following the Model View Controller (MVC) design pattern, hosted by Google Cloud Platform (GCP) Firebase. Internally, it is a Representational State Transfer (REST) Application Programming Interface (API) developed with Python's Flask[4] framework, running serverless on GCP's Cloud Run. We unpack many of these concepts next.

[4]https://pypi.org/project/Flask/

2.2.1 Progressive Web Application (PWA) Standards

Progressive Web App (PWA)[5] standards is an attempt to improve the experience of developers and users alike. Nowadays, it is ordinary that a person may have over dozens of applications installed in their smartphones and machines, although most of usage goes to a handful of applications. The benefit for the user is that they no longer have to download applications in order to have a fuller experience. PWAs allow for most, if not all, of the functionaluty that native applications have.

When it comes to developers, due to PWAs being traditional web applications, they lift the need for third-party distributors such as Apple App Store and Google Play to interface with users. Furthermore, it is much easier to maintain because of its having a single code-base when compared to separate code-bases for Android, iOS and web. As mentioned before, this single code-base closely matches a traditional web application, hence developed with a combination of HTML, CSS and JavaScript. However, such tools are only as good as the patterns used to systematize development and function. We take a look at one of the most well-established design patterns for web applications in the next section.

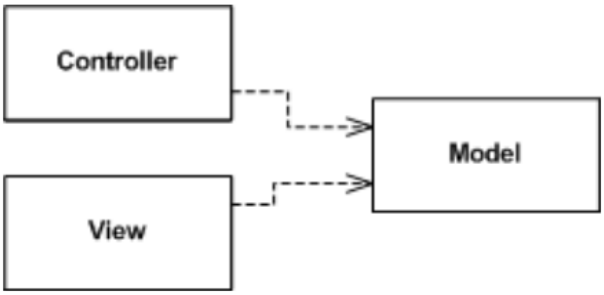
2.2.2 Model View Controller (MVC) Design Pattern

When dealing with web services, one must consider that data flow is not previously determined or bound in any way. This gives web applications somewhat of a asynchronous nature, which entails that having hard-coded sequences of actions that make assumptions about their inputs is a rather dangerous practice and unpleasant to the user. So, it seems that detaching logic from interface enables for a more interactive and secure experience in a web application. Model View Controller (MVC) is a widely adopted design pattern that aims to solve the problem of data flow and interactiveness.

Model View Controller design pattern assigns objects in an application one of three roles: model, view, or controller. The pattern defines not only the roles objects play in the application, it defines the way objects communicate with each other. Each of the three types of objects is separated from the others by abstract boundaries and communicates with objects of other types across those boundaries. The collection of objects of a certain MVC type in an application is sometimes referred to as a layer; for instance, model layer.

[5] https://developer.mozilla.org/en-US/docs/Web/Progressive_web_apps

Figure 2.4: Essential dependencies between model, view, and controller



Source: https://martinfowler.com/eaaDev/uiArchs.html#ModelViewController

Model Objects

Model objects encapsulate the data specific to an application and define the logic and computation that manipulate and process that data.

View Objects

A view object is an object in an application that users can see. A view object knows how to draw itself and can respond to user actions. View objects are typically decoupled from model objects in an MVC application.

Controller Objects

A controller object acts as an intermediary between one or more of an application's view objects and one or more of its model objects. Controller objects are thus a conduit through which view objects learn about changes in model objects and vice versa. Controller objects can also perform setup and coordinating tasks for an application and manage the life cycles of other objects.

2.2.3 Representational State Transfer (REST) Architecture

In short, Representational State Transfer architecture is one solution to the problem of sharing data. When developing a service or a study, the authors may feel compelled to give others access to their data and to the functions being implemented. For a variety of reasons, the authors may not want to give access to all of the data, or to the raw data. as well as to the code they have. Instead, REST defines ways to expose end-points to the web so that others can consume and input data. It defines basic operations [21] that an service must have in order to support most programming languages, and establishes a standard format for the data.

2.2.4 Serverless Architecture

Providing a service on the internet can be a tricky task. As it is virtually impossible to know beforehand the amount of traffic a web service will receive, it becomes extremely hard and costly to provide low latency and scalable services; that is, for an ordinary person or small to mid-sized companies.

The big tech companies, namely Google, Amazon, Microsoft, and smaller ones too have realized how profitable it could be to package all the infrastructure maintenance tasks that they have already mastered and provide it as a service for those who lack the skills or the initial investment required for infrastructure. The next big realization came with the wide adoption of containerized software, which means software that carries with it all of its dependencies. These cloud service providers realized that instead of having the user choose and control hardware, it would be best to abstract that layer too.

That is how "serverless" came to be. It is the concept that, provided with containerized software, cloud service providers take care of all the complexity of balancing load, scaling services horizontally, maintaining hardware and many other tasks that require specialized professionals. In the end, it enables new entrepreneurs and small businesses to have reliable services without the attached complexity – at the expense of control and privacy[6].

[6]https://www.gnu.org/philosophy/who-does-that-server-really-serve.en.html

Chapter 3

Research Methods

> In accordance to the Brazilian General Data Protection Regulation[1] and Twitter's Developer Policy[2], all of the data we collect is public and has been anonymized for the sake of privacy. No information that can lead to participant identification will ever be made public regardless of whether its author's making it public themselves. No other person but the authors of this article have access to the data. No contact has ever been or will ever be made to participants, as the focus of this work is neither to diagnose nor to provide treatment for those in need at this time.
>
> – Statement of Privacy

Having spent the first part of Chapter 2 in the theoretical foundations of our research, it is time to describe our methods for data collection, feature engineering and machine learning, as well as the intricacies of our data pipeline and the caveats of our methods. First, we start with the logic behind our data collection and some of its caveats.

3.1 Data Collection

Due to a lack of labeled corpora targeting depression in Brazil, we found an opportunity to direct our own data collection specifically for that purpose. All data collection took place between July, 2018 and May, 2019. Public messages posted between 2016 and 2019 have been collected for both depressive and non-depressive classes, for a total of 2941 Twitter users ($N_{depressive} = 1486$, $N_{non-depressive} = 1455$). Further explanation of the concepts and processes involved in this task are discussed in the next sections.

[1] Article 4(b) of Law 13.709/2018: https://bit.ly/32o1GaA

[2] Section F, item 2(b): https://developer.twitter.com/en/developer-terms/policy.html

3.1.1 Depressive Class (Positive)

We seek users who publicly state that they have been diagnosed with depression. We query Twitter for public messages in Portuguese that contain self-reports of depression (i.e., "I was diagnosed with depression"). The reason why people come out publicly with such self-reports is presumably to seek support from the community, to explain some of their behaviour to their peers, or to fight against the stigma of mental illness. Some may jokingly or disingenuously make such statements, though the motivation behind such behaviour escapes the scope of this work. Nevertheless, we presume that the majority of self-reports is indeed truthful and will statistically overbear disingenuous ones. We then retrieve up to 3200 of the most recent public tweets for each user and routinely updated their message pool. Users with fewer than 30 messages in total or more than 300 messages in a single given day were filtered out (Section 3.1.4) and the remaining 1486 users were considered as positive observations. Both upper and lower bounds of 300 daily messages and 30 messages in total were put in place to remove supposedly spamming and marketing accounts and to have at least enough samples to enable some statistical analysis — although the latter is a rule of thumb for statisticians rather than rigorous methods.

3.1.2 Non-Depressive Class (Control)

So as to be able to validate the data and to induce machine learning models that distinguish allegedly depressive from non-depressive signals, we query Twitter for public messages in Portuguese that do not contain self-reports of diagnosis of depression. Further efforts are made to remove from the control class users that also report diagnosis of anxiety, due to observed comorbidity between both disorders [22]. Next, we retrieve up to 3200 of the most recent public tweets for each user, routinely updated their message pool, and filter out users with fewer than 30 messages in total or more than 300 messages in a single given day (Section 3.1.4). The remaining 1455 users were considered as negative observations.

3.1.3 Caveats

No effort was applied in verifying either the veracity or the onset date of users'depressive disorder from self-report of diagnosis. It is often the case that no mention is made of when the diagnosis was first reached or whether it was the first diagnosis at all. Some users even go as far as reporting diagnosis from childhood and adolescence. For such cases, we conjecture that the high rate of recurrence of depression [23], can be used to interpret self-reports of diagnosis as an arguably reliable statement of relapse. Thus, behavioral

and linguistic style attributes of such users is assumed to more closely match those of users in the depressive class.It also follows that we establish no time-wise limit to when messages were posted other than the span between 2016 and 2019.

3.1.4 Data Pipeline Architecture

Extract Transform Load (ETL) is the process of developing a pipeline to collect data from a particular private or public database, performing data cleaning tasks such as outlier detection and imputation [24], in addition to making the data easily available to client applications. To build our own ETL pipeline, we employed Python 3.x and its plethora of modules (libraries) to achieve our goal.

Extract

We take advantage of Python's Tweepy[3], a wrapper for the Twitter API, to perform queries to Twitter. There is a sequence of 3 stages of queries and transformations that we make. We present here the tasks specific to extraction and in the transform section, its particular tasks:

1. *update_old_user*, which gets a batch of up to 3200 Tweet[4] objects for each of the users that are already in our database. The limit of 3200 objects is imposed by Twitter for the free-tier developer account.

2. *fetch_new_user*, which searches Twitter for new users to include in the positive and control class, following the logic of Section 3.1.1 and Section 3.1.2 respectively for up to 7 days before the day of query ($D_{query} - 7$).

3. *fetch_new_user_data*, which gets a batch of up to around 3200 Tweet objects for each new user found on the previous step.

This first stage is made much simpler thanks to the Twitter API. At this stage, our data is still kept in a raw JavaScript Object Notation (JSON) file, in addition to a Comma Separated Values (CSV) file to keep track of users. An important note on privacy is that in order to lessen the risk of leaking user information, despite only having access to exclusively public content, we cypher all data files with a 4096 bits RSA[5] public key of a particular machine used by the authors. In practice, this means that the data files on our server can only be decrypted by the holder of the private key that pairs with the public key used for cyphering, hence only the authors. One additional precaution we take is to

[3]http://www.tweepy.org/
[4]https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/tweet-object
[5]https://crypto.stanford.edu/ dabo/courses/cs255_winter03/rsa-lecture.pdf

hash each users' id before storing the JSON file since we need to access users' ids in order to retrieve new data. The dictionary with hashes and their respective id is also kept away from the server.

Next, we discuss the tasks performed in the "transform" side of operations. Note that tasks in "extract" and "transform" run intercalated in each of the steps.

Transform

Tasks in this stage of the pipeline are extremely important to assure data quality. The process of removing outliers and imputation (i.e., replacing missing values) can inject unwarranted biases into the distribution of variables if not done properly. In the present study, EDA led to the discovery of a common trend of highly active small clusters of users in both the positive and the control set. Further inspection revealed these to be advertising accounts and allegedly bot accounts. The result of this inquiry grounded the decision of establishing hard-coded thresholds for specific variables. For instance, suppose it takes 1 minute for a person to type and send a *tweet* and that the person repeats this process non-stop for 5 hours. Is it fair to say that this scenario is rather unlikely? Still, this hypothetical person would have posted half of what some of our outliers did. Thus, we argue that establishing hard-coded limits to some variables, instead of following the interquartile range rule, was an informed and fair decision.

There are a few other tasks that need performing at this stage of the pipeline. We highlight some of their aspects here, but reference the user to our repository[6] for further information when the code-base is opensourced. Main tasks involve:

1. *update_old_user*, which checks whether accounts of users in our database are still active. If not, flags the user as deactivated and stops collection attempts. If so, receives data from "extract" and:

   i) extracts hashtags ("#tag") and mentions ("@user") into new variables (columns);
   ii) removes hyperlinks, extra whitespace and non-Latin characters from text;
   iii) processes data into time series as described further in Section 3.2.1;
   iv) summarizes times series into features as describe in Section 3.2.2;
   v) splits the data via the holdout technique as in Section 3.3.1;

2. *fetch_new_user*, which hashes user identification key and removes all variables for the sake of privacy, but for "full_text" and"created_at".

[6]https://github.com/unbsense/etl_pipeline.git

3. *fetch_new_user_data*, which performs the same actions of task 1 (*update_old_user*) Really, they are the same module. We split it into two instances to make it clear that the module executes twice.

Load

Unfortunately, there is a caveat to this last stage of our pipeline, which stops it from being a "real" ETL pipeline. The load module can be run independently from the rest of the pipeline. This goes against one principal of ETL that is to have all compute and storage of data within the pipeline so as to avoid the overhead of copying large quantities of data to client applications. Having said that, the scale of our data collection is void if compared to a large big data analytics company. The pipeline is automatically triggered weekly so that it can stay up to date. All of that is the reason why we choose to have the load module be solely our database and keep machine learning methods in a separate module. For database, we run the NoSQL MongoDB[7] database. Regarding our machine learning function, it is more thoroughly discussed in Section 3.3.

3.2 Feature Extraction

Now, it is time to describe in more detail one of the most important tasks of the "transform" stage of our pipeline (Section 3.1.4). We start with a top-down exploration of the signals that correlate to depression in the literature, structure them as attributes (times series) and summarize them into numerical variables (features). These numerical variables are used to induce our ML classifiers, discussed in Section 3.3.

3.2.1 Attributes

Research suggests that as we give hints of our emotional state at every interaction with another human-being, so does it happen when we interact with our machines and devices. We explore that research in depth in Section 2.1.1. Now, we define the classes of signals and their respective attributes (Class: (0) attribute).

- Engagement: we define (1) *volume* as an attribute measured by the total number of messages per user, per day.
- Linguistic Style: simple natural language processing techniques were used to derive three attributes regarding pronoun use. (2) *fpp* is defined as the daily count of

[7]https://www.mongodb.com/

first-person pronouns in posts. The next two attributes follow the same logic for (3) *spp* second-person pronouns and (4) *tpp* third-person pronouns.

- Emotion: the unigram sentiment instrument ANEW-Br [17] was used to derive 2 attributes in this category. (5) *valence*, which ranges from unpleasant to pleasant, and (6) *activation*, which ranges from relaxed to tense (e.g., sadness and anger, with low and high activation respectively, and low valence for both).
- Depression Terms: We define (7) *depre_terms* as the ratio of the number of words in a message that belong to the depression lexicon for Brazilian Portuguese [16] to the total number of words in a message.
- Insomnia Index: Insomnia has been shown to have significant correlation with depression [25]. Thus, we define (8) *insomnia index* as the daily ratio of messages posted at night ("11PM-6AM") to messages posted during the day ("6AM-11PM").

All of these attributes (time series) are generated with Python's Pandas module and its "groupby" operations. Most signals represent a simple count os sum of the daily measures with the exception of "*insomnia index*", which is a ratio, hence a float between 0 and 1. In the next section, we discuss how these collection of time series per user are transformed into numerical features.

3.2.2 Feature Vectors

Machine Learning (ML) provides powerful supervised and unsupervised learning algorithms that can be employed to induce models for the distribution of variables in our data. In order to harness such power, we must transform our attributes (time series) into features (summaries) that capture some of the truth of the underlying phenomena.

Thus, each of our 8 attribute series that correlate with depression is summarized by 4 scores, namely *mean frequency* (μ), *variance*, *weighted moving average* and *entropy*. Given a time series $X_i(0), X_i(1), ..., X_i(t), ..., X_i(N)$ for the i^{th} attribute, the summaries are computed as follows:

1. *Mean frequency* (μ_i) as the average measure of the time series signal of an attribute over the entire period of analysis.

$$\frac{1}{N}\sum_{t=0}^{N} X_i(t) \tag{3.1}$$

2. *Variance* as the variation in the time series signal over the entire time period.

$$\frac{1}{N}\sum_{t=0}^{N}(X_i(t)-\mu_i)^2 \tag{3.2}$$

3. *Weighted Moving Mean (WMM)* as the weighted relative trend of a time series signal in a window of M (=7) days.

$$\frac{1}{N}\sum_{t=0}^{N}(X_i(t)-(1/(t-M))\Sigma_{(M\leq k\leq t-1)}X_i(k)) \tag{3.3}$$

4. *Entropy* as the measure of uncertainty in a time series signal.

$$-\sum_{t=0}^{N}X_i(t)log(X_i(t)) \tag{3.4}$$

The last step after computing our features is to standardize them. Standardizing observations (rows) means representing them with standard deviation points from the mean of the variable (columns). Let t be the observation for a given user, the standardization of the j^{th} summary number is defined by:

$$\bigcup_{j=0}^{31}\frac{X_j(t)-\mu_j}{\sigma_j} \tag{3.5}$$

Thus, each user is represented by a standardized 32-item feature vector with zero *mean* and unit *variance*. The importance of standardization comes from the inability of many learning algorithms to work with distinct scales of variables, running the risk of having the scale itself be a learned property when it really should not be.

As a final comment in this section, we must emphasize the importance of splitting the data into at least a train and test set before performing standardization so as to avoid injecting bias between sets. Such care must be taken in order to increase the likelihood that models can generalize to new observation rather than memorize our data, which is commonly called "overfitting". We explore next what processes we put in place in order to avoid overfitting and injection of bias while learning our models.

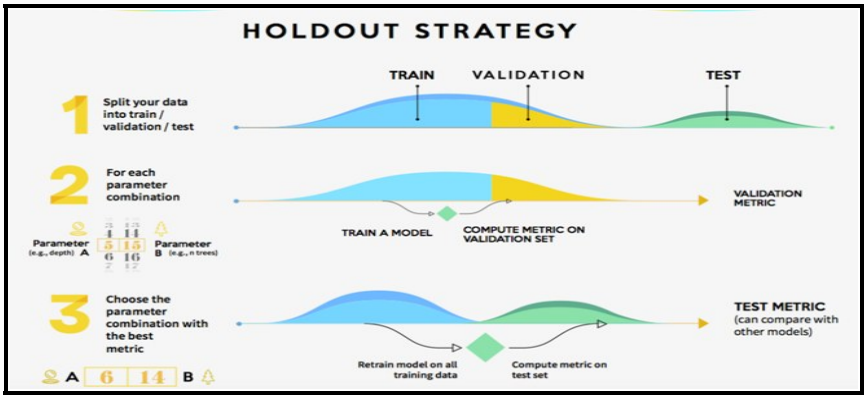
3.3 Machine Learning Classifiers

We train supervised machine learning classifiers to discriminate between depressive and non-depressive classes.

3.3.1 Splitting the Data

In simple terms, the holdout set is a partition of the data that is saved just to produce performance metrics for models. A validation set is the partition of the data used to adjust models' hypeparamenters in order to improve performance. It is used in conjunction with the train set. The train set itself is the partition of the data used to fit models. Having said that, there is much confusion in applied ML about what a validation set is and how it differs from a test set. To get a broader comprehension of the topic, we reference the reader to Section 2.1.3. In pratcice, our data set is randomly split into training (50%), validation (20%) and test (30%) sets.

Figure 3.1: Instructions for the Holdout technique



Source: https://shorturl.at/pINU5

3.3.2 Dimensionality Reduction

Based on the principle of Ockham's razor, which states that the simplest solution is most likely the right one, and to avoid overfitting, we employ Principal Component Analysis (PCA) to reduce dimensionality and capture 96% the variance in our data set, which in turn yields 13 principal components. The choice of 96% instead of the usual 95% of variance is justified by our results from hypothesis tests. In short, our hypothesis tests identified enough difference (p-value $< 1.5625e{-}3$) in the mean of 13 features (Table 3.1). When looking at the results of PCA for 95% of the variance, the resulting 12 principal components were largely affected by the 12 of the 13 relevant features. Thus, we ran PCA once again for 96% of the variance, and surely enough it resulted in 13 principal components.

It is important to note that PCA is an unsupervised learning algorithm to fit better axis for variables based on their eigenvectors. Nevertheless, common tools such as Pandas

Table 3.1: Selected features after dimensionality reduction.

fpp_variance	*volume_mean*
volume_entropy	*insomnia_variance*
insomnia_moving_mean	*valence_mean*
valence_variance	*valence_moving_mean*
valence_entropy	*activation_mean*
activation_variance	*activation_moving_mean*
activation_entropy	

enables us to trace which of the variables is most important when computing each principal component. To sum up, we were able to verify that PCA selected the 13 features from our hypothesis test (not with the same weight between them), which led us to use PCA as a sort of validation of our features rather using it to transform our variables.

3.3.3 Supervised Learning Models

Done with the stage of dimensionality reduction, it is time to discuss our classifiers induce with supervised learning algorithms. Here, we present specifics of our classifiers, in other words, both parameters and hyperparameters adopted. In order to have a better superficial understanding of how the algorithms work, we reference the reader to Section 2.1.3. As a technical note, all 3 classifiers are trained with Python's *scikit-learn*[8] module.

Random Forest (RF)

We utilize 100 iterations of randomized search with 5-fold cross-validation to optimize our RF hyperparameters. Our best set of hyperparameters is a 250-tree RF classifier with maximum depth of 15 and a minimum number of 20 samples required to split an internal node.

Support Vector Machine (SVM)

We employ grid search over 5,625 hyperparameter combinations with 10-fold cross-validation to optimize our SVM classifier. Our best performing SVM classifier has the hyperparameters *gamma* set to 'scale' and C to 45.

[8]https://scikit-learn.org/stable/index.html

Multilayer Perceptron (MLP)

We take advantage once again of grid search to find optimal settings over 2,500 hyperparameter combinations, which yields a thirteen-node two-hidden-layer MLP classifier with the hyperparameters *alpha* and *momentum* set to 0.01 and 0.9, respectively.

3.4 Closing Remarks

In this chapter we have discussed all the nuances of our methods and their caveats. We consider it best practice to be upfront about the limitations of our methods and results, so that future work can drawn from this work without injecting any unknown shortcomings or drawbacks. In the light of results, we go over ours for statistical significance of our features and performance of our models in the next chapter.

Chapter 4

Results

> If you find that you're spending almost all your time on theory, start turning some attention to practical things; it will improve your theories. If you find that you're spending almost all your time on practice, start turning some attention to theoretical things; it will improve your practice.
>
> – Donald Knuth, 1991, p.15

4.1 Statistical Analysis

We first present results of the statistical analysis for the 5 out of the 8 attribute classes. The exclusion of 3 attributes is given due to their lack of statistical significance for any of the extracted features. We use independent sample (unpaired) t-tests to compare the mean of the positive and control classes, for p-value $\leq \alpha = 0.05/32 = 1.5625e-3$ after Bonferroni correction for multiple comparisons. The values of the t-statistics and an indication of p-value $\leq \alpha$ on Table 4.1.

Table 4.1: Statistical significance of principal components

Attributes	**Entropy**	**Mean**	**WMM**	**Variance**
activation	4.6257 **	10.3910 **	9.3983 **	6.367147 **
fpp	0.1476	-0.9533	1.3704	4.014316 **
insomnia	-0.4773	0.2027	-12.1403 **	4.018704 **
valence	4.6273 **	8.7854 **	7.9036 **	4.545939 **
volume	-3.9258 **	-5.7723 **	-3.0485	1.050110

$df = 2939$ ** $p \leq \alpha$, after Bonferroni correction

Our statistical results align with De Choudhury et al [3] to some extend. Variance is the attribute that sees the highest number of significant features, although we argue that

mean and WMM are still more relevant. The higher overall t-values for features derived from both measures shows that they bear a heavy weight in rejecting the null hypothesis that the two classes — users with self-reported diagnosis of depression and users with no self-report of diagnosis — have similar signals of behaviour, language and cognition. It is a fact that we extract fewer signals in comparison. Nevertheless, we base ourselves on the argument that major underlying signals of depression can be captured from signs of insomnia [25], high attention to self [8], lack of energy and exhaustion [26], loss of social connectedness [9], and elevated negative emotion [12]. Thus, our choice of features seems adequate and yields positive results in classification.

4.2 Classification Performance

We choose to go beyond the usual set of performance scores for the sake of better understanding and comparison of our models. The average precision, recall, specificity and F_1 scores are reported in Table 4.2 with 95% confidence intervals (CI) over 100 randomized runs. Our best classifiers were the Multi-layer Perceptron and the Support Vector Machine due to their higher chance of detecting most users in the positive class (recall), in addition to doing the same for the control class (specificity). Overall, both the SVM and MLP classifiers have comparable performance, as shown by the intersecting confidence intervals of their classification performance scores in Figure 4.1. We opt to prioritize recall over other measures since such computerized tools are not meant to supplant professional diagnosis. Instead, they should enable physicians and care-takers to spend more time in direct contact with individuals who do indeed need their care.

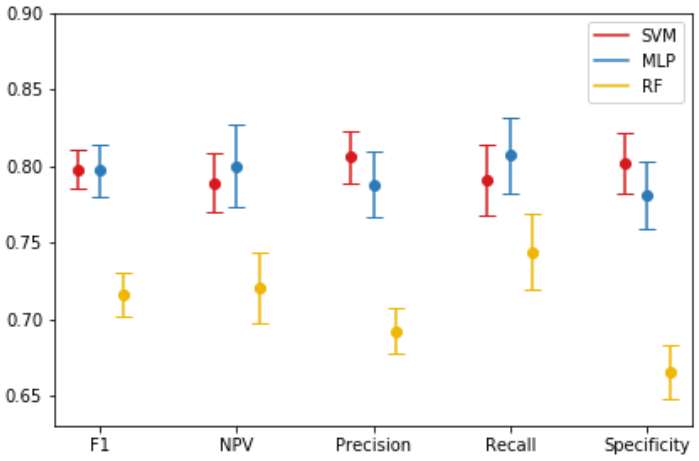
Table 4.2: Average performance scores over 100 randomized runs with 95% CI.

	F_1	precision	recall	specificity
SVM	0.798 ± 0.012	0.806 ± 0.017	0.791 ± 0.023	0.802 ± 0.020
MLP	0.797 ± 0.018	0.788 ± 0.021	0.807 ± 0.025	0.781 ± 0.022
RF	0.716 ± 0.014	0.692 ± 0.015	0.744 ± 0.025	0.665 ± 0.018
Baseline	0.502 ± 0.018	0.50 ± 0.018	0.502 ± 0.018	0.493 ± 0.013

Both the SVM and MLP classifiers improved over De Choudhury and colleagues [3], though the comparison is loosely formal and some considerations have to be made. First, De Choudhury extracts features from 1 year of user history seeking to predict the onset of depression, whereas we do so in a span of 3 years to detect users with overall observed behaviour correlated with depression in the literature; thus, direct comparisons

of complementary and yet distinct objectives is unsuited under rigorous analysis. Conversely, no effort has been made in this work to verify whether users in fact suffered from depression. De Choudhury employs both CES-D (Center for Epidemiologic Studies Depression Scale) [27] and Beck Depression Inventory (BDI) [28] to screen participants for depression. Nonetheless, Coppersmith and colleagues [4] demonstrates that trading off meticulous screening for larger amounts of data not only yields better prediction results, but also enables more robust machine learning techniques to be leveraged. Hence, we present our results with confidence despite the aforementioned caveats.

Figure 4.1: Comparative plot of CI for the classifiers' performance scores.



To some extend, our challenge differs in nature from Coppersmith and colleagues, as well as De Choudhury and colleagues. The cultural and language barrier takes its toll on natural language processing techniques as well as on the availability and accuracy of lexicons. For instance, the depressive lexicon for Brazilian Portuguese [16] failed to show statistical significance at this time, whereas Coppersmith and colleagues demonstrates the efficacy of using lexicons to construct novel methodologies. On the other hand, all four features derived from the ANEW-Br lexicon [17] reached statistical significance, which in turn suggests that, although nascent, this field of research is likely to thrive in Brazil.

4.3 Published Paper

When it comes to the acceptance of our work, it seems that we indeed approached the right topic at the right time. Our earlier work, an article entitled "Mining Twitter Data for Signs of Depression in Brazil" has received the award of best paper at the VII Symposium on Knowledge Discovery, Mining and Learning[1] held in Fortaleza, CE, in October, 2019.

[1] http://sbbd.org.br/kdmile2019/

The present work builds upon that one, aiming to bring to real life some of the benefits that our research results can yield. In this vein, we give a brief overview of the development process and functionality of our prototypical application's architecture.

4.4 Web Application

The intent of our developing an application is to provide an online space that can help people understand their depression signals, and encourage them to seek professional help when needed. In the future, such application can grow to host content creation, awareness campaigns, online treatment with certified professionals and even customization to fit companies needs.

4.4.1 Architecture

Our application is designed with two main components. The front end and the back end. In the front end PWA, we provide users with analytics of their public Twitter data, whilst servicing processed data from the server-side API. The two of them communicate direct and indirectly via HTTP requests and serverless functions, respectively. We include a list of dependencies and development tools in Section B.1 and Section B.2. Next, we give an overview of both the PWA client and the REST API.

4.4.2 Client Application

Thus far, the functionality of the front-end client is still very limited. That does not mean that little work has been put in developing it. In order to comply with most requirements of the concepts, patterns and frameworks discussed in Section 2.2, it took the authors some getting used to the overall structure. Nonetheless, the opensource community never disappoints. By forking the skeleton provided by the project Bento Starter[2], the task became more feasible. Next, we highlight our projects features with the appropriate remarks.

Home

As previously stated, the client web application still lacks in functionality. Nevertheless, we briefly describe the user experience and dive deeper into what takes place system-wise.

The user is greeted with the application name, UnB Sense, a somewhat vague greet and a button that opens in a new tab the "about us" page. With regards to what happens

[2]https://github.com/kefranabg/bento-starter

Figure 4.2: Web client home page.

Source: https://app-sense1.firebaseapp.com/home/

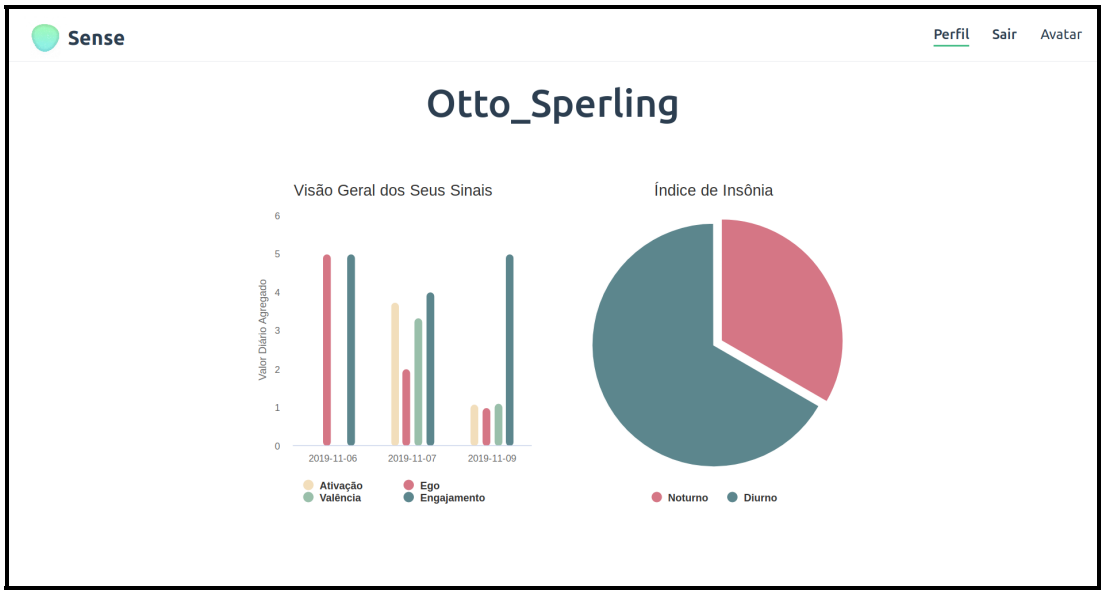
internally, the application has already mounted the object "store", which is responsible for data structures as well as flow, and plays the role of "model" (referenced as store) from MVC. In simpler terms, the "store" object is the only object that can directly modify the state of variables, and the data hub for all other objects and components.

The main "view" object (referred to as view) of our application is simply a placeholder for the navigation bar and the home page view. This ability to handle views as if the were instances of classes creates vast opportunities for modularity, code reusability and layering of functions. It also supports inheritance, which makes it easier to enforce the same style of as many objects or components as needed.

Log In

Our components behave as "controller" objects (referenced as controllers), which are responsible for interfacing the requests of views and the store. Suppose then that the user clicks the "log in" button. The view object triggers a specific method of the controller, that decides whether to return anything or to send a request to the store. The store receives the request for a "getter" (which can be thought of as a function) that checks if the user state in the store is set to logged in and an access token is registered. If not, the store activates GCP's federated sign-in serverless function; which performs the authentication between Twitter's Oauth access delegation and GCP's Firebase. Thus, the process of authentication is fully handled by our cloud service provider.

Figure 4.3: Web client dashboard.



Source: https://app-sense1.firebaseapp.com/profile

After the "log in" request is successful, Firebase returns user metadata collected from their public profile to the store, which mutates its required states. Finally, it triggers a method of the controller, that indicates to the view that the store has been updated. The view object updates its states and triggers another controller object responsible with selecting the correct route, the "router".

Before we continue, there is another important point regarding privacy to note here. Even though Firebase returns an access token for the user account, we choose not to use the token and favor explicitly public data, so as to say data we can see even without the token. We also set the required permissions to the lowest invasive configuration that Twitter allows.

Dashboard

Last but most important, our dashboard is the place we want to provide users with insights into their own signals. For now, the interface is simple, too simple. Having said that, we make the commitment to continuously develop it, and hopefully one day, we get the community to take part in it so that it no longer depends on any single person.

Regardless, we shall get back to point. As soon as the profile controller is mounted, it sends another request to the store. This time, it needs data for user reports, so the store triggers both the Firebase function and our API running on GCP's Cloud Run. It sends requests to both for distinct reasons. From Firebase, the stores expects to receive data fast so that it can display all reports even if the are out-dated. Meanwhile, the API

performs the task of checking and updating the attributes and features (Section 3.2.1 and Section 3.2.2, respectively). The rationale behind this scheme is that the user should not have to wait to start using the application. Thus, it is preferable to display out-dated data than no data at all. This is one important principle of PWAs.

Once the store receives either of the requests, it triggers the controller, which in turn triggers the view so that it can be updated and display the new data. Lastly, the user can log out of the account by pressing the "log out" button, which makes the view trigger the controller, which triggers the store, which mutates its states to have the user set to "logged out" and the access token removed. The last step is to trigger the router to redirect the user to the home page.

4.4.3 REST API

The simplest way to explain our API is to imagine it as if it were a reduced version of our ETL data pipeline (Section 3.1.4). Instead of running the modules extract, transform and load for a group of users, it does the same processing for the user in scope. Our ML classifiers are pre-loaded into the container that runs on GCP's Cloud Run. That is far from ideal for real production deployment. Ideally, these classifiers need to be periodically calibrated or fully retrained to guarantee performance and fairness. This is indeed one of the top priorities in our roadmap for future versions. Summing up, so as to avoid being a bore to the reader, instead of discussing here again all the intricacies of our ETL task, we reference the curious or the forgetful to Chapter 3, where we explore such task in detail.

4.5 Closing Remarks

In this chapter, we have presented our results of statistical analysis, classification task via machine learning models, peer-reviewed publication as well as the development of a prototypical application. In the next chapter, we discuss the implications of data mining to society, aspects of ethics that must be considered by researchers and businesses alike, concepts of privacy and security, as well as present arguments for how open source software can be used to create monetary profit while respecting citizens at large.

Chapter 5

Social Impact of Data Mining

> I could have made money this way, and perhaps amused myself writing code. But I knew that at the end of my career, I would look back on years of building walls to divide people, and feel I had spent my life making the world a worse place.
>
> – Richard Stallman, 2002, p.14

5.1 Healthcare Research

In Chapter 2 we present how social media data is used for a wide range of studies into the nature of the human mind. However, the use of social media data goes beyond mental health, into many other domains of healthcare. Not only researchers are interested in applying data mining to healthcare, companies are spending millions of dollars to advanced their capability to predict a person's future healthcare needs based on their social media content, data from smart devices and medical background as well as broader profiling techniques (e.g. gender, income, education level, etc) with the aim to personalize care and augment the efficacy of treatments.

Moreover, there is a number of other potential benefits of using social media in research, including the ability to reach larger numbers of participants at a much lower cost than ever before, greater opportunities for interaction across extended time periods, the possibility to mitigate bias that can surface when it comes to direct contact between researchers and participants, in addition to generating new ways to broadcast research results to a larger audience. However, multiple concerns have been raised about the implications of employing data mining on social media data by privacy advocates, researchers[1], policymakers[2] and patients at large. Next, we discuss some of the risks of such practices.

[1] https://tinyurl.com/tls9yka
[2] https://tinyurl.com/wouc2ms

Risks

There is controversy involving what it means to handle of social media data ethically, and no clear consensus has emerged among researchers. In Chapter 3, we argue that the public property of the data we collect, the care we take not to contact any user of the platform in addition to anonymizing and securing the data, can arguably exempt us from undergoing the scrutiny of an Internal Review Board. Nonetheless, in hindsight it seems that it was a poor rather than good choice. Here it lies one aspect of the risks that mining social media data poses to society as a whole. Current legislation on data protection and consent lags behind the potential of these new technologies, and the ethical principles remain relatively underdiscussed in computer science departments across Brazil.

During a post hoc analysis, we find that Conway [29] suggest a taxonomy of 10 ethical considerations specifically relevant to the use of social media data in research:

- privacy;
- informed consent;
- ethical theory;
- institutional review board (IRB)/regulation;
- traditional research versus social media (e.g. Twitter) research;
- geographical information;
- researcher lurking;
- economic value of personal information;
- medical exceptionalism;
- benefit of identifying socially harmful medical conditions.

Each of these considerations aim to mitigate one or more risks that research may incur. However, it is hard to prospect whether the industry will adhere to any considerations made solely by academia. Instead, regulation seems to be the way moving forward. In the past years, there have been far too many cases of user privacy being broken by both big and small technology companies. For instance, a 2014 survey [30] found that among the 600 most used mobile health apps available for Android and iOS, only 183 (30.5%) had privacy policies, of which two thirds (66.1%) had no mention of the app itself in their policies. Furthermore, Facebook received significant criticism regarding its covert "emotional contagion" study [31] conducted in 2012 which aimed to prospect how changing users' timeline content could induce distinct emotional states, and that involved

approximately 600 000 of its users, from whom no research consent was obtained, to whom no study information was provided, and who were unable to withdraw from the study.

This is one example of how lack of transparency and centralized data, coupled with research practices that do not take into account any ethical considerations, can lead to abuse of privacy of social media users and indeed cause harm. In the same vein, we explore next how an alternative technological paradigm has been finding itself useful in safeguarding individuals privacy and rights.

5.2 Free Software and Community

Today's internet has wildly diverged from the original distributed concept envisioned by Tim Berners-Lee and colleagues towards a centralized network controlled by the few. From Edward Snowden's leakage of NSA's global surveillance to Cambridge Analytica's deceitful use of Facebook data to interfere with elections in multiple countries, it has become increasingly clear that centralization of data does not go hand in hand with user privacy.

All in all, many computer scientists agree that free and open source software (FOSS)[3] is of great benefit to us all. However, the general population still does not understand the difference between free software and software that is free (no cost). Simply put, the former lets you use it and change it however you'd like, the latter uses you and your data however it likes. Fortunately, there is a growing wave of FOSS and distributed technologies (e.g., blockchain) that aim to protect user privacy over big-tech profit, in addition to data protection regulations that many governments, including Brazil's, have been pushing forward.

For our part, due to mental illness being a sensitive topic, we argue that opening the source code of our application is a good way to ensure transparency to users of what is done with their data. Nevertheless, maintaining an application and its infrastructure does indeed occur in costs. This is where the concept of crowd-funding, which is the practice of asking users to financially support a service they would like to see grow, becomes a cornerstone of our project. We do not intend for our application to be a source of profit, instead we hope it can become somewhat resemblant of a public service that is maintained by the community, akin to tax-payer funded programs, and that people can directly contribute to its growth.

[3] https://www.fsf.org/about

Chapter 6

Conclusion

> The first way in which science is of value is familiar to everyone. It is that scientific knowledge enables us to do all kinds of things and to make all kinds of things. Of courseif we make good things, it is not only to the credit of science; it is also to the credit of the moral choice which led us to good work. Scientific knowledge is an enabling power to do either good or bad - but it does not carry instructions on how to use it. Such power has evident value - even though the power may be negated by what one does with it.
>
> – Richard Feynman, 1988, p.01

The question we ask at the begining of Chapter 2 still rings in the ears. What is science for? What is the use of it? Does it exist for its own sake? We believe not.

It seems to us that a tool is only as good as the use it is given, although science could not be further from a simple tool. Instead, science gives us the most reliable methods to answer questions that are dear to our hearts and minds. Thus, good versus bad is of full responsibility of those who choose to wield the scientific method towards a particular goal, and it must not be treated as mere unattended consequence. In this vein, we choose to employ the methods of data mining and machine learning as an attempt to explore how useful they can be to help identify depression before it is too late. In this final chapter, we take time to reflect on all that has been done in this work, but also on some of the shortcomings; how it aims to contribute to society in Brazil, as well as the opportunities for future work.

On broad strokes, we **i)** construct a data set in Portuguese through an automated data pipeline built with Python, **ii)** extract features that capture some of the underlying signals of depression based on the literature of psychology, psychiatry and sociolinguistics, **iii)** induce ML classifiers to distinguish the depressive from non-depressive class, in addition to presenting the performance scores (with 95% CIs) that compare to, if not improve upon, findings in the literature [3, 4, 12]. Last but not least, we **iv)** develop a RESTful

API to package our data processing modules and **v)** implement a PWA client to consume data from our API.

6.1 Contribution

The main contribution of this work is that it shows the feasibility of applying methods of data collection, feature extraction and machine learning to identify signs correlated with depression on social media in Brazil. We hope our work can catch the interest of the next generation of undergraduates and steer them into applying science for the sake of mental health, for mental illness is a growing challenge that is not likely to slow down in the years to come.

Fortunately, we are able to support the findings of Kristensen and colleagues [17], since all features extracted with the use of the ANEW-Br lexicon were found to be significant and greatly help in classification. On the other hand, we are unable to confirm the findings of Nascimento and colleagues [16] with the Portuguese lexicon for depression. It is important to mention that since this is not a replication study, we do not claim invalid the methods of Nascimento and colleagues. Instead, we simply comment on the inefficacy of the lexicon to yield significant features in our study.

The last, but still important, contribution of our work is the development of a simple online application to give users insight into their depression signals. There is much to gain in helping people help themselves, individually and in aggregate. Far too often, help arrives after damage has already been done, as it takes those who suffer much courage to fight stigma and self denial. If our application can one day help even a single person seek help before the storm, we will be glad to have made the effort.

6.2 Future Work

We present the first version of our application, despite its being considerably bare, with the commitment to continue iterating over it, so as to make the tool more and more useful. This and extending support to other social media platforms seem to be the low-hanging fruit. Nevertheless, a closer look suggests that there is some hidden complexity behind the latter; that is, supporting other social media such as Instagram goes beyond the scope of our work, as it brushes upon computer vision rather than textual content. And that is the first suggestion we make for future work. Employing computer vision methods to study mental illness in Brazil, as it is being done in more developed countries.

It follows that another interesting way to extend our work is by proposing, validating and testing features particular to Brazilians that may correlate to depression. It is needless

to say that it will take an interdisciplinary effort to achieve robust results. However, we believe there can be much to discover about the signs Brazilians leave behind. One side note here; it is utterly important that computer scientist and so-called data scientists engage with other domain areas outside of the "hard" sciences. It is not hard to imagine a world in which tasks have been fully automated, and humans become the core of our attention. As we have much to contribute with our precision, methods, tools and proofs, "soft" sciences will be ever more present in our domain of research.

Another way that our line of work can be made even more relevant is if it can be used to better understand social trends with more granularity than usual, and allow for more effective actions and policies by governments when it comes to awareness, prevention and treatment of mental illness. Furthermore, it need not be limited to guidance of public agents, but it can also be useful for physicians and care-takers to better understand the needs of their patients. In fact, there is an ever growing number of startups trying to provide mental health services online, and a couple of them have been successful thus far. Nonetheless, there is little to no oversight on how these services are built, which can lead to abuse of privacy, data security issues, and misuse of data, as discussed in Chapter 5. Fortunately, every challenge is also an opportunity and here is where academia can help. By providing an open source application that focus on mental health, any interested person can become a contributor or verify compliance with the rights of users. This is indeed what we hope our application can one day become, a public service maintained by the community for the community.

Finally, we suggest that more sophisticated learning algorithms, both supervised and unsupervised, be used to achieve better classification results. Moreover, forecasting is a task that has been untouched in this work, and yet it could be as useful, if not more useful, for physicians and care-takers than classification itself. Artificial neural networks have been shown to boost performance in uncountable domains of application, it should be no different this time. We suggest exploring long short-term memory (LSTM) recurrent neural networks (RNN) for it has the ability to hold representations of relevant observations much further in the past than other RNNs. It seems that it could be used to track user signals going through a depressive episode, so that it learns how to forecast future ones.

6.3 Closing Remarks

With that, we reach the end of this work. It has been an immense opportunity to learn about and contribute to a cause that far exceeds the intricacies of individualism. We believe mental illness is still seen as the problem of a particular mind, perhaps due to

genetics or an abusive upbringing, but contained to and experience by particular individuals. However, we have learned that this idea could not be further from the truth. Mental illness bears a burden in all of society. It is us who lose when we fail to care, it is us who lose when somebody decides to end their own life. There is no stereotype, no target group; there is no favorite color, race, creed or gender. Mental illness is a problem of us all, and it is up to us, and only us, to solve it.

In the future, we hope to see much more research in mental health and data mining in our home country. Regrettably, far too often technology is taken for granted or disingenuously employed. One of the major driving forces of innovation in behaviour modeling and prediction — *profit maximization* — is not usually aligned with the well-being of the public at large. Despite the diversion, public interest has been shifting towards a more open (and equally difficult) discourse about privacy and individual rights. In the scope of individual rights, we argue that mental health and supportive care are intrinsic to having a fulfilling journey through life; thus, individual rights.

Bibliography

[1] World Health Organization: *Depression and other common mental disorders: Global health estimates*. https://bit.ly/30iFz52, 2017. 1, 8

[2] Greenberg, Paul E, Andree Anne Fournier, Tammy Sisitsky, Crystal T Pike, and Ronald C Kessler: *The economic burden of adults with major depressive disorder in the united states (2005 and 2010)*. The Journal of clinical psychiatry, 76(2):155–162, 2015. 1

[3] De Choudhury, Munmun, Michael Gamon, Scott Counts, and Eric Horvitz: *Predicting depression via social media*. In *Seventh international AAAI conference on weblogs and social media*, pages 0–10, Cambridge, USA, 2013. AAAI conference on weblogs and social media. 1, 6, 8, 27, 28, 37

[4] Coppersmith, Glen, Mark Dredze, and Craig Harman: *Quantifying mental health signals in twitter*. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 51–60, Baltimore, USA, 2014. Association for Computational Linguistics. 1, 6, 8, 29, 37

[5] Rabiner, Lawrence R and Biing Hwang Juang: *An introduction to hidden markov models*. ieee assp magazine, 3(1):4–16, 1986. 2

[6] Reagan, Andrew J, Christopher M Danforth, Brian Tivnan, Jake Ryland Williams, and Peter Sheridan Dodds: *Sentiment analysis methods for understanding large-scale texts: a case for using continuum-scored words and word shift graphs*. EPJ Data Science, 6(1), Oct 2017. `http://dx.doi.org/10.1140/epjds/s13688-017-0121-9`. 2

[7] Reece, Andrew G., Andrew J. Reagan, Katharina L. Lix, Peter Sheridan Dodds, Christopher M. Danforth, and Ellen J. Langer: *Forecasting the onset and course of mental illness with twitter data*. Scientific reports, 7(1):13006, 2017. 2, 6

[8] Rude, Stephanie, Eva Maria Gortner, and James W. Pennebaker: *Language use of depressed and depression-vulnerable college students*. Cognition & Emotion, 18(8):1121–1133, 2004. 6, 28

[9] Williams, Keith L. and Renée Galliher: *Predicting depression and self-esteem from social connectedness, support, and competence*. Journal of Social and Clinical Psychology - J SOC CLIN PSYCHOL, 25(8):855–874, October 2006. 6, 28

[10] Bollen, Johan, Huina Mao, and Alberto Pepe: *Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena.* In *Fifth International AAAI Conference on Weblogs and Social Media*, volume 2011, pages 450–453, Palo Alto, USA, 2011. AAAI Conference. 6

[11] Moreno, Megan A, Lauren A Jelenchick, Katie G Egan, Elizabeth Cox, Henry Young, Kerry E Gannon, and Tara Becker: *Feeling bad on facebook: Depression disclosures by college students on a social networking site.* Depression and anxiety, 28(6):447–455, 2011. 6

[12] Park, Minsu, Chiyoung Cha, and Meeyoung Cha: *Depressive moods of users portrayed in twitter.* In *Proceedings of the ACM SIGKDD Workshop on healthcare informatics (HI-KDD)*, volume 2012, pages 1–8, Philadelphia, USA, 2012. ACM SIGKDD. 6, 28, 37

[13] Resnik, Philip, William Armstrong, Leonardo Claudino, and Thang Nguyen: *The university of maryland clpsych 2015 shared task system.* In *2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 54–60, Baltimore, USA, January 2015. 6

[14] Pennebaker, James W., Matthias R. Mehl, and Kate G. Niederhoffer: *Psychological aspects of natural language use: Our words, our selves.* Annual review of psychology, 54(1):547–577, 2003. 6

[15] Bradley, Margaret M. and Peter J. Lang: *Affective norms for english words (ANEW): Instruction manual and affective ratings.* The Center for Research in psychophysiology, 30(1):25–36, 1999. 6

[16] Nascimento, Rodolpho S., Pedro Parreira, Gabriel N. Santos, and Gustavo P. Guedes: *Identifying signs of depressive behaviour on social media (identificando sinais de comportamento depressivo em redes sociais).* In *7º Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2018)*, pages 0–6, Porto Alegre, Brazil, 2018. SBC. 7, 22, 29, 38

[17] Kristensen, Christian Haag, Carlos Falcão de Azevedo Gomes, Alice Reuwsaat Justo, and Karin Vieira: *Brazilian norms for the affective norms for english words.* Trends in Psychiatry and Psychotherapy, 33(3):135–146, 2011. 7, 8, 22, 29, 38

[18] Nuzzo, Regina: *Scientific method: statistical errors.* Nature News, 506(7487):150, 2014. 8

[19] Reinhart, Alex: *Statistics done wrong: The woefully complete guide.* No starch press, 2015. 8

[20] Agrawal, Rakesh, Tomasz Imieliński, and Arun Swami: *Mining association rules between sets of items in large databases.* SIGMOD Rec., 22(2):207–216, June 1993, ISSN 0163-5808. `http://doi.acm.org/10.1145/170036.170072`. 10

[21] Fielding, Roy T and Richard N Taylor: *Architectural styles and the design of network-based software architectures*, volume 7. University of California, Irvine Doctoral dissertation, 2000. 15

[22] Sartorius, Nornam, T. Bedirhan Üstün, Yves Lecrubier, and Hans Ulrich Wittchen: *Depression comorbid with anxiety: results from the WHO study on psychological disorders in primary health care.* The British journal of psychiatry, 168(S30):38–43, 1996. 18

[23] Burcusa, Stephanie L. and William G. Iacono: *Risk for recurrence in depression.* Clinical psychology review, 27(8):959–985, 2007. 18

[24] Farhangfar, Alireza, Lukasz Kurgan, and Jennifer Dy: *Impact of imputation of missing values on classification error for discrete data.* Pattern Recognition, 41(12):3692–3705, 2008. 19

[25] Jansson-Fröjmark, Markus and Karin Lindblom: *A bidirectional relationship between anxiety and depression, and insomnia? a prospective study in the general population.* Journal of Psychosomatic Research, 64(4):443 – 449, 2008, ISSN 0022-3999. `http://www.sciencedirect.com/science/article/pii/S0022399907004114`. 22, 28

[26] Rabkin, Judith G. and Elmer L. Struening: *Life events, stress, and illness.* Science, 194(4269):1013–1020, 1976. 28

[27] Radloff, Lenore Sawyer: *The CES-D scale: A self-report depression scale for research in the general population.* Applied psychological measurement, 1(3):385–401, 1977. 29

[28] Beck, Aaron T., Robert A. Beck, and Gregory K. Brown: *Manual for the beck depression inventory-ii.* Psychological Corporation, 78(2):490–498, 1996. 29

[29] Conway, Mike: *Ethical issues in using twitter for public health surveillance and research: developing a taxonomy of ethical concepts from the research literature.* Journal of medical Internet research, 16(12):e290, 2014. 35

[30] Sunyaev, Ali, Tobias Dehling, Patrick L Taylor, and Kenneth D Mandl: *Availability and quality of mobile health app privacy policies.* Journal of the American Medical Informatics Association, 22(e1):e28–e33, 2014. 35

[31] Selinger, Evan and Woodrow Hartzog: *Facebook's emotional contagion study and the ethical problem of co-opted identity in mediated environments where users lack control.* Research Ethics, 12(1):35–43, 2016. `https://doi.org/10.1177/1747016115579531`. 35

Appendix A

Images

Figure A.1: Outliers in Ames Housing Dataset

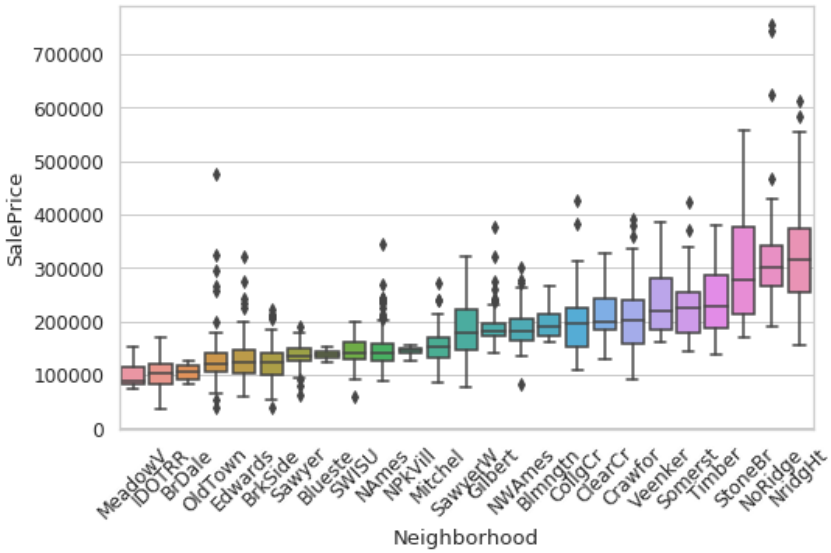


Figure A.2: Example of K-Means clustering

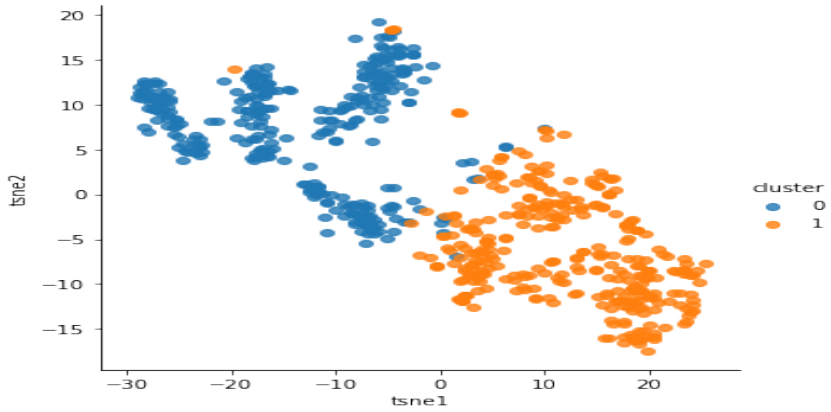
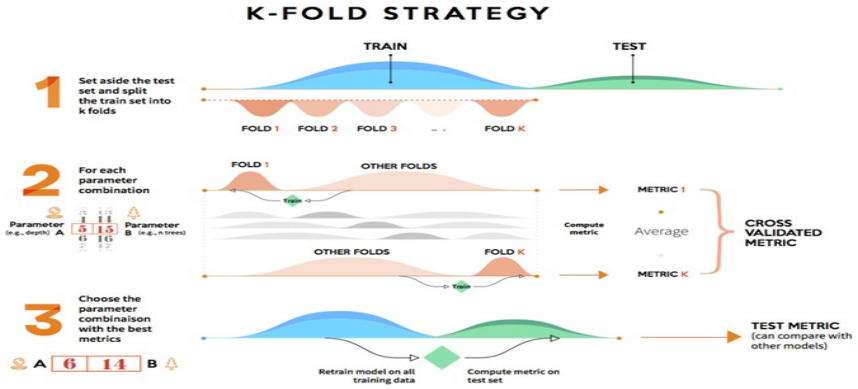


Figure A.3: Full confusion matrix.

Source: https://en.wikipedia.org/wiki/Precision_and_recall

Figure A.4: Instructions for the K-fold technique



Source: https://shorturl.at/pINU5

Figure A.5: Front (left) and back (right) end structure.

```
./src/
├── assets
│   └── img
├── components
│   └── charts
├── firebase
├── misc
├── router
├── store
│   ├── app
│   ├── authentication
│   ├── products
│   └── sensitive_data
├── theme
└── views

14 directories
```

```
./
├── assets
│   └── __pycache__
├── methods
│   ├── logs
│   └── __pycache__
├── __pycache__
└── templates

7 directories
```

Appendix B

Application Dependencies

BRIEF INTRO

B.1 Web Client Dependencies

Built with Vue.Js and:

```
    "axios": "^0.19.0",
    "bootstrap": "^4.3.1",
    "bootstrap-vue": "^2.0.4",
    "core-js": "^3.3.6",
    "firebase": "^6.6.2",
    "highcharts": "^7.2.1",
    "is_js": "^0.9.0",
    "lodash": "^4.17.15",
    "pwacompat": "^2.0.9",
    "register-service-worker": "^1.6.2",
    "vue": "^2.6.10",
    "vue-head": "^2.1.2",
    "vue-router": "^3.1.3",
    "vuex": "^3.1.1"
```

B.2 API Dependencies

```
Built with Python and:
    CacheControl==0.12.5
    cachetools==3.1.1
    certifi==2019.9.11
    chardet==3.0.4
    Click==7.0
    entrypoints==0.3
    firebase-admin==3.1.0
    flake8==3.7.8
    Flask==1.1.1
    google-api-core==1.14.3
    google-api-python-client==1.7.11
    google-auth==1.6.3
    google-auth-httplib2==0.0.3
    google-cloud-core==1.0.3
    google-cloud-firestore==1.5.0
    google-cloud-storage==1.20.0
    google-resumable-media==0.4.1
    googleapis-common-protos==1.6.0
    grpcio==1.24.3
    gunicorn==19.9.0
    httplib2==0.14.0
    idna==2.8
    itsdangerous==1.1.0
    Jinja2==2.10.3
    MarkupSafe==1.1.1
    mccabe==0.6.1
    msgpack==0.6.2
    numpy==1.17.3
    oauthlib==3.1.0
    pandas==0.25.2
    protobuf==3.10.0
    pyasn1==0.4.7
    pyasn1-modules==0.2.7
    pycodestyle==2.5.0
    pyflakes==2.1.1
```

```
PySocks==1.7.1
python-dateutil==2.8.0
pytz==2019.3
requests==2.22.0
requests-oauthlib==1.2.0
rsa==4.0
six==1.12.0
tweepy==3.8.0
Unidecode==1.1.1
uritemplate==3.0.0
urllib3==1.25.6
Werkzeug==0.16.0
xlrd==1.2.0
```