Master thesis on Intelligent Interactive Systems
Universitat Pompeu Fabra

Anorexia Nervosa Detection and Recommendation System

Miguel Fagundez

Supervisor: Ana Freire

Co-Supervisor: Ludovico Boratto

July 2019

Universitat Pompeu Fabra
Barcelona

Master thesis on Intelligent Interactive Systems
Universitat Pompeu Fabra

Anorexia Nervosa Detection and Recommendation System

Miguel Fagundez

Supervisor: Ana Freire

Co-Supervisor: Ludovico Boratto

July 2019

Universitat Pompeu Fabra
Barcelona

Contents

Dedication

I would like to dedicate this work to My Family and Friends and everyone who has supported me.

Acknowledgement

I would like to express my sincere gratitude to:

- My supervisor: Ana Freire.
- My co-supervisor: Ludovico Boratto.
- My family and Friends for all the support.

Abstract

Anorexia Nervosa is a fatal illness that affects a large number of people. This master thesis proposes an approach to contribute to the recovery process for the Anorexia Nervosa patients. A Recommender System is modeled in order to recommend recovered users to anorexic users to encourage them to start their journey to recovery, using collected tweets and the topics or categories that the users talk about in them. Two classification models were created to compare how precise they were classifying anorexic users and recovered users respectively. A pair of topic detection approaches were applied in order to determine the topics or categories present in the tweets. Results have shown that LDF and NMF algorithms are not good for this task. But, using a third-party library allowed to get a useful set of categories and how much the users talk about them. The recommender system was done based on these categories and using the cosine similarity measure. The validation process used to evaluate the recommendation was comparing if the results obtained are more similar to the interests of users in process of recovery (based on the users they follow) than what a selected group of random users is. The results obtained show that a recommender could be built and AN users might benefit from it to start or help their recovery process.

Keywords: Recommender System; Eating Disorders; Social Networks; Classification Algorithms; Topic Detection.

Chapter 1

Introduction

Eating Disorders are fatal illnesses where there are disturbances in people's eating behaviors and their entire set of feelings and emotions related to eating [1]. Anorexia Nervosa (AN) is an Eating Disorder in which people see themselves overweight when they are actually underweight. There are several key symptoms: They weigh repeatedly, they control in a non-healthy way what they eat, they will not maintain a normal or healthy weight, they do not eat enough to be healthy and they force themselves to vomit or they just simply use laxatives. It is important to remark that is the Eating Disorder with the highest mortality rate [1]. Eating Disorders also involve the highest mortality rate of the psychiatric disorders [2].

Social networks are crowded with personal information and that is regularly used for banal purposes. However, some users use these platforms to express themselves about some personal issue and find support even from anonymous people. For example, users with AN tend to create profiles where they frequently use some keywords in order to be identified by other users in the same condition. Terms such as *#proana, #ana, #mia, #promia, cw, gw, Thighgap, Bikinibridge, Thinspo or Bonespo,* are frequent in this kind of profiles. *Thighgap* refers to the gap formed between thighs, *Bikinibridge* refers to the space between pelvic bones and a lot of acronyms like *GW/CW/UGW* mean Goal Weight, Current Weight, Ultimate Goal Weight, etc.

Detecting anorexic users is key to be able to address the issue in the society. In this thesis we aim to detect anorexic users, as well as users recovered from anorexia, and make a recommender system between them. This idea is supported in [3] and it is evident that people with a group of support will always be more likely to recover from a psychological illness.

1.1 Related Work

Anorexia and Eating Disorders can be reflected in tweets [2] and it has been studied that is possible to determine if an account is a PRO-ED (Eating - Disorders) account based on a set of keywords. The connection that exists between the body image, social networks and the Eating Disorders is evident, given the amount of content that is available in a free and easy way [4]. It is also possible to see that the online communities that exist display and promote this kind of behavior encouraging it, as it was researched in [3]. There have been past studies as stated in [5] focused on how these online health communities interact (In the case of the present project it is important to know how the pro-anorexia community interact). In [5] it is also displayed how there are two clearly separated communities where one of them encourages and promotes disordered eating behaviours and there is another one that actually tries to fight against the Eating Disorders. If these two groups are considered as clusters, it is possible to say that the main communication is intra-cluster. The latter community will be helpful to analyze the recovery of the anorexic patients which is also a topic of the present work. Those users will be recommended to the anorexic ones.

According to [6] there is a need to better understand how the content on social media provides us a way to actually prevent problems for the people suffering anorexia where it is specified that 90% of them were girls and 77% were younger than 19 years old giving too much importance to their body weight in a non-healthy way. A girl this young needs help and the present work tries to dig deeper into the evaluation of one approach to this problem. The problem of encouraging Eating Disorders is growing because it is self encouraging as it is explained in [2]. There are too many online communities but there is a difference established between online communities that look for a weight loss in a healthy way and the communities which are only centered on the weight loss regardless of the self-harm they are doing to themselves [7] [8]. Based on all these ideas, this research will try to leverage all the knowledge and help, by using a recommender system, the users that suffer from this Eating Disorder.

In the present work, the intention is to enable the communication from a recovered user to an anorexic one with the hope of creating a recovery network using the interest similarities of the users. This similarity will be measured using a topic detection approach which has already been tested in [9].

Twitter is not the only social network where the community that promotes anorexia is present, another social network that is highly known for its pro-ana community is Tumblr [10]. In Tumblr the set of keywords is similar to the list of keywords to be used in the present project. The pro-ana community has demonstrated to be consistent with the way to expose itself. Instagram is another social network where despite the message is not as explicit as in Twitter, a lot of admiration is given to super skinny celebrities which could lead to more pro anorexic users trying to idolize them. On the other hand, Reddit is also a social network where the ED are encouraged [11] and establishes that a further research on how to intervene (or affect in a positive way) online to prevent future problems.

It is important to highlight that apparently healthy communities might devolve in a self harm community that can portray a negative image on the body image of the viewer [12]. The present work aims to combine the classification system and analysis of the tweets in order to recommend the best recovered users (considering its similarity) to the anorexic users. To the best of our knowledge, this thesis is the first work proposing a recommender system for Anorexic users in social media with the purpose of preventing future problems for them.

1.2 Objectives

The goal of the present work is to elaborate a system that can improve people's (suffering from anorexia) lives by helping them to overcome the anorexia condition. In order to achieve this, the following steps will be followed:

- Develop a classifier for anorexic users using a Machine Learning Approach.
- Develop a classifier for recovered users using a Machine Learning Approach.
- Build a recommender system in order to match a recovered user to an anorexic user where the users are alike. Since it is not an option to deploy the system to production (this would mean that Twitter would have to get user's account recommendation from the present work), another metrics will be used in order to measure the quality of the recommendation by analyzing the characteristics of each one of the users on both sides.

1.3 Outline

The present work is organized in 6 different chapters. In Chapter 2 all the background is described including all the concepts needed to understand the process and results. In Chapter 3 the description of the proposal to achieve the objectives of this master thesis will be explained. In Chapter 4 the results are discussed and described. Chapter 5 summarizes the results and concepts. In Chapter 6 future work and research that can be done using this work as a baseline are described.

Chapter 2

Background

This section is addressed to those readers unfamiliar with classification algorithms and recommender systems.

2.1 Classification Algorithms

In order to address our objectives, we will build two set of classification algorithms and a recommeder system, described in Chapter 3. In this section we describe the machine learning methods and tools used for building our algorithms.

2.1.1 Methods Used

K - Nearest Neighbors

The K-Nearest Neighbors algorithm is possible to be implemented to be applied to both classification and regression problems[13]. The main idea is to find the nearest points to the new point and classify the new one as the older(but nearest) have been classified.

In order to find what is "near" for a point, different metrics can be used as the distance, three examples of those metrics can be: Minkowski Distance, Manhattan Distance and Euclidean Distance.

Random Forest

Random forest is based on a decision tree learning algorithm where a graph is used to model the conditionals that defines the output label of the input. It is tightly closed to a set of rules to apply to give the label of the element. In this case, a variation that avoids overfitting is Random Forest that is an ensemble method that works by training several decision trees being each one of them independent from the others. Those trees have an intrinsic randomness property in order to learn different aspects of the data and when they are a ensemble, they are expected to return a better output that a common decision tree[14].

Multinomial Naive Bayes

Multinomial Naive Bayes is a probabilistic machine learning model based on the bayes theorem that can be seen in equation 2.1:

$$P(A \mid B) = \frac{P(B \mid A)\,P(A)}{P(B)} \tag{2.1}$$

The 'Naive' word appears because it is often used (as a simplifying assumption) in cases where the 'effect' variables are not actually conditionally independent given the cause variable[15]. On the other hand, the 'Multinomial' is added when the frequencies of events can be generated by a multinomial distribution.

SVM

Support Vector Machine, also abbreviated as SVM, is a classifier model (discriminative), formally defined by a hyperplane which tries to separate as best as possible the classes by maximizing the distance between them. In the implementation, the algorithm finds the solution to a quadratic programming formulation of the problem, This found solution represents the hyperplane with maximum margin to its closest points (these are called support vectors)[16].

Logistic Regression

The logistic regression algorithm is based upon a logistic function (in its basic form) which tries to model a binary dependent variable. This logistic function was invented in the 19th century in order to describe the populations growth and the course of autocatalytic chemical reactions[17]. Logistic regression is a statistical method developed by David Cox in 1958. It is used to predict a binary outcome given a set of independent variables. The algorithm's goal is to fit the data to a logistic function to predict the probability of occurrence for a given event [18].

Multilayer Perceptron

Multilayer Perceptron, also known as an Artificial Neural Network, is a learning algorithm that is based on how the human brain works. It tries to create a set of layers of neurons to run an input through all of them and get an output. It is worth highlighting that it must have at least three layers. The input layer where all the input gets into the network, then a set of hidden layers (at least 1) to process the input and output layer showing the output [19].

XGBoost

XGBoost is a boosting algorithm designed to be scalable[20] for billions of examples using as little resources as possible. This tree bosting method is based on gradient tree boosting algorithms using several algorithmic optimizations that make it possible to be scalable to billions of examples on a single machine.

AdaBoost

AdaBoost is a boosting algorithm created by Freund and Schapire [21] that tried to solve many of the boosting algorithms difficulties on its beginning. It is based on some basic principles as calling a weak or base learning algorithm repeatedly in a series of rounds. One key idea is to maintain a distribution or set of weights over the training set and each weight is updated when there are incorrectly classified examples. In order that the weak learner is forced to focus on the hard examples in the training set.

2.2 Techniques used for Training and Validation

To ensure proper training and correct validation of what is a good fit, some statistics concepts and libraries were used, the below concepts will describe how they work.

2.2.1 SMOTE (Synthetic Minority Over-sampling TEchnique)

SMOTE is a technique of over-sampling used to deal with imbalanced datasets in which the minority class is going to be over-sampled using "synthetic" examples instead of the usual over-sampling with replacement. The specifics of the SMOTE approach is that the way the "synthetic" examples are generated is done in the feature space instead of the data space using the nearest neighbors[22].

2.2.2 Cross Validation

Cross-validation works by splitting a dataset into a number N of folds, getting N-1 folds as the data for training data and the the rest of the data as the test set. This process is to be repeated N times (using all the folds) in order for each bucket to be used one time as validation [23].

2.3 Natural Language Processing

In order to analyze the contents of the tweets it is needed to be specified how to handle natural language information. The important concepts used are described below.

2.3.1 TF-IDF

Term Frequency-Inverse document frequency (TF-IDF) is a method to handle the importance of a term to a document in a collection of documents (which can be called corpus). It can be split in two important terms, Term frequency (TF): assigns a weight to each term in a document equals to the number of times that a term *Term* appears in document *Doc*. A con of Term Frequency, by itself, is that considers all terms equally important. Words highly common but not important, such as pronouns, will be assigned a high weight no matter what they are actual relevance really is. Inverse document frequency (IDF): It tries to handle the issue just mentioned above by adding a penalty to all the words that appear in a lot of documents [24].

2.3.2 Topic Detection

As stated in [9] the Topic Detection in Twitter Data is possible with a set of algorithms, this was the approach taken in the present work. Particularly, 3 algorithms were used: Non-Negative Matrix Factorization (NMF), Latent Dirichlet Allocation (LDA) and K-Means.

Non-Negative Matrix Factorization (NMF)

One of the approaches to topic detection is using Matrix Factorization Techniques where the issue of finding negative values in the factorized matrices is present [9]. As a solution to this issue, the technique of Non-Negative Matrix factorization (NMF) was proposed. In order to achieve this, given a matrix M, it is factorized in the product of matrices A and B, where the matrix A represents the coefficients of each document and the matrix B represents the basis (in each row).

Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words [25].

2.4 Recommender Systems

A recommender system is a software tool to provide possible suggestion to a user which they can leverage upon to [26].

It is easy to find several examples like:

- Item recommendation to buy.
- Movies that the user is expected to like.
- Places wherever the user probably will like to go.

The purpose of identifying useful items for a user is vital in a recommender system. The recommender system must predict that the item should be worth recommending. For example, for a song recommender system a really simple approach is just to recommend the most popular song, in this way the likelihood of the user liking a song that a lot of people like is expected to be high [26].

There are several classes of recommendation approaches:

- Content-Based: Recommends items similar to ones users have liked in the past.
- Collaborative Filtering: Recommends items that users with similar tastes have liked in the past.
- Demographic: It is based on demographic niches where people from different niches should get different recommendations.
- Knowledge-Based: Recommendations are based on specific domain knowledge, it requires a function to estimate how is the match between the needs of the user and the actual available recommendations.
- Hybrid Recommender System: Recommends based on two or more of the approaches defined above.

The Content-Based approach learns how to recommend items that are similar to others in which the users have shown interest in the past. This *similarity* is calculated using some of the features that are associated with the respective items. An example of this approach is when a user has already liked a reggae song in the past and based on this, the system recommends similar songs from this genre. Another adequate example for a set of users is: Considering a user A and the topics of discussion of the users followed by user A, it is possible to recommend to user A, other users to follow, by calculating the topics and the users that discussed in the most similar way about those topics. This approach (Content-Based) is the one to be used in the present work.

In order to achieve this match the recommender system needs a measure metric.

2.4.1 Cosine Similarity

Cosine similarity is a similarity measure used to compare two documents with respect to a given vector of query words. Let x and y be two vectors for comparison, these two vectors can be very sparse since they are gotten as a term frequency vector and in this kind of vectors the cosine similarity is really appropriate since it is a measure for numeric data that ignores zero-matches and handles properly the terms that the documents do actually have in common [27].

Chapter 3

Proposal

The goal of the present work is to be able to classify anorexic users as well as recovered users. As well as implementing a recommender system to provide recovered users to the anorexic users as a mechanism to help their recovery. A set of keywords will be used to retrieve the data from Twitter. We filtered out retweets, removed emojis and hid the user's Twitter screen name (In order to make the tagging completely anonymous) and the URLs. After the retrieval of the data, a file will be uploaded to Amazon Mechanical Turk in order to be tagged as Anorexic Users and not Anorexic Users. The same will happen for users recovered from anorexia. After this tagging is done, a recommender system will be implemented in order to match the anorexic users with the users recovered from anorexia. This is done with the goal of getting them together, so the anorexic user can be inspired and overcome their eating disorder. It is important that the social network data is leveraged in order to help people improve their life.

Amazon Mechanical Turk is a crowdsourcing tool to outsource processes and jobs to a paid workforce who can perform these tasks virtually. It can include from data validation to more subjective tasks like content moderation [28]. Amazon mechanical Turk was the tool chosen to label (tag) the users that were retrieved. In this case, the task created was to tag a set of tweets for a user. There were several options of classes in which a user could be tagged as seen in the Table 2 and Table 3.

3.1 Anorexic Users

3.1.1 Retrieval

The following approach will be used for retrieving the Twitter Data: The retrieval of the set of tweets from Twitter will be done by filtering the data using a set of keywords for the anorexic users as it can be seen in Table 1.

Table 1: Keywords Table

Anorexic Users Retrieval	**Recovered Users Retrieval**
anorexic	eating disorder recovery
ana	anarecovery
bulimic	chooserecovery
proana	healthy recovery
Pro-ED	pro recovery blog
promia	reasons to recover anorexia
ednos	recovery fighter anorexia
wannarexic	recovery food
thighgap	recovery intake anorexia
bikinibridge	recovery record anorexia
thinspo	recovery tips anorexia
bonespo	recoveryisworthit anorexia
GW	recoverywarriors anorexia
CW	road to recovery anorexia
LW	self recovery anorexia
UGW	
calorieApril	
ABCdiet	
laxies	

3.1.2 Tagging

A set of labels were designed to be used for the tagging purpose of the possible anorexic users, the Table 2 specifies them:

Table 2: Anorexic Labels Tagging

Anorexic Users Labels
Accounts giving Information about Anorexia
Unrelated
User Against Anorexia
User Promoting Anorexia
User Recovered From Anorexia
User with Anorexia

The tagging will be done using Amazon Mechanical Turk with a set of 3 workers which will help the project to get the proper labels for the user without displaying any of the information of the user and only showing text that are not URLs or mentions.

3.1.3 Classify Users With Anorexia Nervosa

The set of tweets from anorexic users, prior to classification, will have a label describing whether is an anorexic user or not, per user id. Then, the classifier will work with this binary set of labels. Cross Validation and Grid Search approaches will be followed in order to ensure the best quality for the classifiers. It is also worth highlighting that the classification algorithms used are described in 2.1.1.

3.2 AN Recovered Users

3.2.1 Retrieval

The following approach will be used for retrieving the Twitter Data: The retrieval of the set of tweets from Twitter will be done by filtering the data using a set of keywords for the recovered as it can be seen in Table 1.

3.2.2 Tagging

Table 3: Recovered Labels Tags Table

Recovered Users Labels
Anorexic user talking about recovery
Media Account giving information about anorexia
User in process of Anorexia Recovery (not recovered yet)
Other
User recovered from anorexia
User recovered from other disorders (not anorexia)
User that encourages other users to recover from anorexia

The tagging will be done using Amazon Mechanical Turk with a set of 3 workers which will help the project to get the proper labels for the user without displaying any of the information of the user and only showing text that are not URLs or mentions.

A set of labels were designed to be used for the tagging purpose of the possible recovered users that could provide help to the anorexic suffering patients, the Table 3 specifies them.

3.2.3 Classify Users Recovered From Anorexia Nervosa

The set of tweets for recovered users, prior to classification, will have a label describing whether is a recovered (or recovery helpful) user or not, per user id. Then, the classifier will work with this binary set of labels. Cross Validation and Grid Search approaches will be followed in order to ensure the best quality for the classifiers. It is also worth highlighting that the classification algorithms used are described in 2.1.1.

3.3 User Recommendation

The goal of the recommender system is to recommend Twitter accounts of users recovered from anorexia to anorexic users based on how similar they are. Similarity will be measured by checking the values for the categories on their tweets. For example if a user usually talks about movies, the chances of connecting with another account that talks about movies too is greater in comparison to other account where sports is the main subject described by the tweets. Summarizing (and simplifying) this process, if a user talks 80% about sports and 20% about politics it will be more probable to be recommended to users that talk in the same percentage about these topics than to users that do not talk about these topics at all and talk about, entirely, different topics.

The definition and calculation of the categories will be done using a topic detection technique or a third-party library that allows to detect a numeric value for a set of categories given a document. The choice between these two options will be made by observing and analyzing the results.

The similarity measure that is going to be used will be the cosine similarity. This similarity is determined by the equation 3.1, in this case it is worth highlighting that the vectors A and B will be the set of values for the categories of two different users.

$$\cos(\mathbf{A},\mathbf{B}) = \frac{\mathbf{AB}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} \mathbf{A}_i \mathbf{B}_i}{\sqrt{\sum_{i=1}^{n} (\mathbf{A}_i)^2}\sqrt{\sum_{i=1}^{n} (\mathbf{B}_i)^2}} \tag{3.1}$$

For a clearer picture take a look at the illustration of the Figure 1.

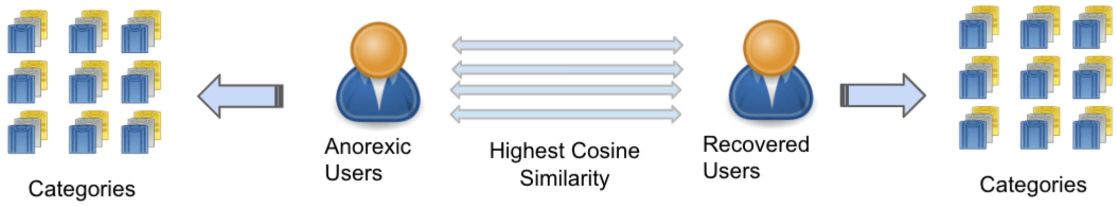

Figure 1: Illustration of the Recommender.

3.3.1 User Recommendation Validation

In order to validate the results, the users that are in process of recovery are taken and a set of users followed by them will be used for comparison. Following the idea that once anorexic users start their recovery process, they are expected to look more similar in terms of users followed to the users in process of recovery.

A baseline of random users will be selected to compare the categories they talk about with the categories of the users mentioned above. The same process is done for the recovered users that have been recommended. The expected behavior is for the recommended users to be more similar to the users mentioned above than the random users.

In this topic it is important to define how will the random users be obtained and the approach taken will be using the trending topics of the Twitter social network, enabling the system to capture users that talk about different topics and that are more colorful and less homogeneous in order to make a better comparison. It is worth highlighting that the number of random users to be used in the comparison can not be small since we do not want to get possible outliers in the comparison but rather a mean or median of the similarity values for the random users. Take a look at the Figure 2, to see a clearer illustration of the process.

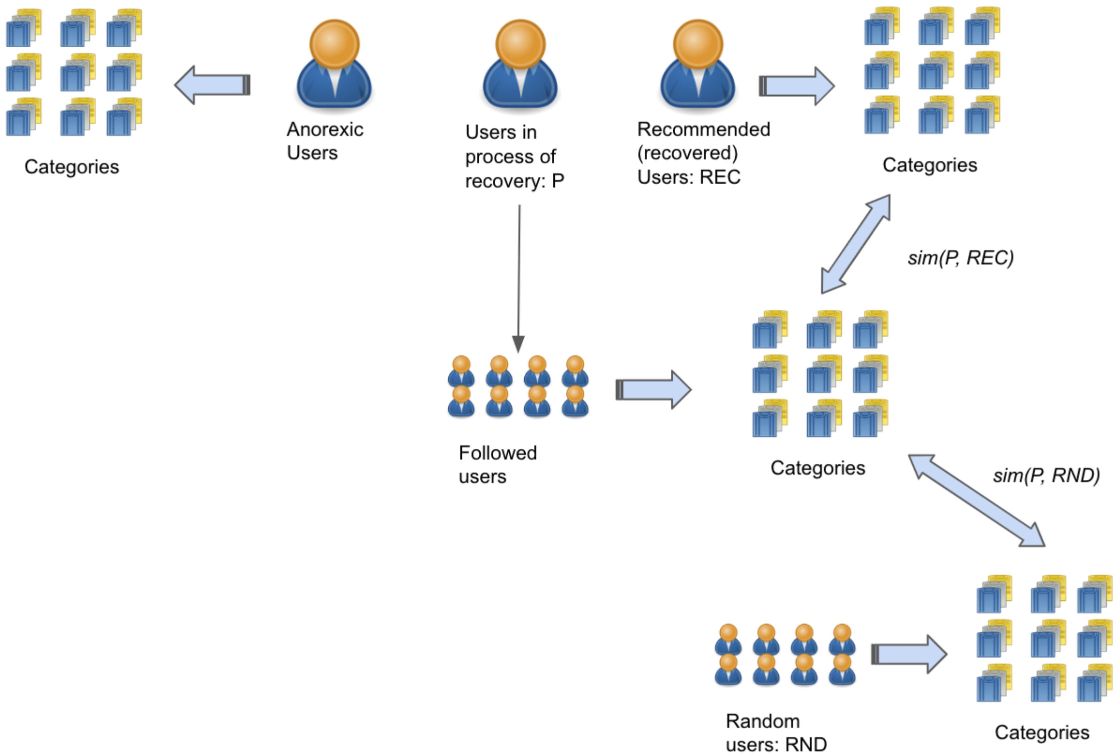

Figure 2: Illustration of the Recommender Validation

Chapter 4

Results

In this chapter the results will be shown for the datasets of the Anorexic Users and the Recovered from anorexia users regarding the possible recommendations and classification results.

4.1 Dataset Details

In the process of the tweets retrieval several issues were found:

- The data was not relevant.
- There were a lot of spam-related tweets.
- The username in twitter mentions was displayed in the data.
- There were public URLs displayed.
- There were searches irrelevant because they were using one of the short keywords (cw, gw) inside another words with entirely another intention.

After having stated the general issues found, there were some specific issues found relative to the set of tweets for anorexic users and recovered from anorexia users that will be explained in the following section.

4.1.1 Anorexic Users Retrieval

The amount of information found for the anorexic-users set of keywords was huge(only 3 weeks of data was needed) and it was possible to actually leverage the relevance of each of the tweets. A blacklist of keywords was used since the first run returned some results that were not entirely relevant. Some of them even stated political tweets mentioning the president of the United States (Donald Trump) or some others mentioning subjects that were more spam-related than actually worth for the task at hand.

Filters

The filters used for excluding the tweets retrieved for the anorexic users were:

- There were more than 2 mentions in the tweet.
- There were more than one URL in the tweet.
- There was, at least, one of the blacklisted words: IOTA, SOTA, GMO, porn, Trump, xxx.
- It was a retweet.

Data Cleaning/Hiding

There were some measures taken into account in order to clean/hide the data displayed in a tweet:

- URLs were replaced by the string "<URL_hidden>".
- Mentions were replaced by the string "@XXXX".
- Emojis were removed.

After retrieving the data from twitter for the anorexic users using the *Tweepy*[1] Python Library, there were users with a set of tweets and these tweets were grouped

[1] https://www.tweepy.org/

by their user and this way then the tagging in Amazon Mechanical Turk could be done by user (since the worker could read the whole set of tweets anorexia-related for the user) instead of per tweet which grants a more useful result.

It is important to mention that the user identification was hidden for the labeling in order not to disclosure the user information but making the tagging anonymous.

The dataset for anorexic users was tagged as follows in table 4 for the 1081 instances (users) keeping in mind that a minimum of three workers per user was required. Tables 2 specified the instances and the Table 4 specifies how many users were tagged in each label.

Table 4: Anorexic Users Label Instances

Label	**Instances No.**	**Instances (%)**
Accounts giving Information about Anorexia	26	2.41
Unrelated	328	30.37
User Against Anorexia	62	5.74
User Promoting Anorexia	37	3.34
User Recovered From Anorexia	25	2.31
User with Anorexia	603	55.83
Total	**1081**	**100**

As it is evident in Table 4 the percentage of anorexic users was 55.83 % against non-anorexic users that was 44.17 %. Having seen this, the classes were relatively balanced (It is worth to remark that a new set of labels was created as anorexic and non-anorexic.

4.1.2 Recovered Users Retrieval

The amount of information for the recovered user was not as big as the one it was obtained for the anorexic users. The keywords from above went from a little bit more flexible ones to the ones we see above in Table 1. That is because the recovery data obtained was too broad. It was not specific to the anorexia as this work intended. For the retrieval of relevant data it was necessary to look for the tweets in the last

3 months and as it is easily interpreted there were a lot of keywords that required the "anorexia" suffix in order to stay relevant. It is worth highlighting that keeping the relevance of this set of tweets (for the problem at hand) was really important given that the term “recovery” can be used for anything that harms people.

Filters

The filters used for the tweets retrieved for the recovered from anorexia users were not necessary, just using the keywords stated in table 1 was enough since the suffix made the response as restrictive and relevant as possible.

Data Cleaning/Hiding

There were some measures taken into account in order to clean/hide the data displayed in a tweet (similar to what it was done for anorexic users):

- URLs were replaced by the string "<URL_hidden>"
- Mentions were replaced by the string "@XXXX"
- Emojis were removed

After retrieving the data from twitter for the recovered users using the *Tweepy* Python Library, there were users with a set of tweets. The tweets were grouped by their user and this way then the tagging in Amazon Mechanical Turk could be done by user (since the worker could read the whole set of tweets recovery-related for the user) instead of per tweet which grants a more useful result as specified above for the anorexic users.

It is important to mention that the user identification was hidden for the labeling in order not to disclosure the user information but making the tagging anonymous.

The dataset for recovered users was tagged as in the table 5 for the 1819 instances (users) keeping in mind that a minimum of three workers per user was required. Table 3 specified the instances and Table 5 specifies how many users were tagged in each label.

It is worth highlighting that in this case a set of master workers were required to improve the quality of the tagging from Amazon Mechanical Turk.

Table 5: Recovered Users Label Instances

Label	**Instances No.**	**Instances (%)**
Anorexic user talking about recovery	59	3.24
Media Account giving information about anorexia recovery	203	11.16
Other	1106	60.8
User in process of Anorexia Recovery (not recovered yet)	127	6.98
User Recovered From Anorexia	123	6.76
User recovered from other disorders (not anorexia)	25	1.37
User that encourages other users to recover from Anorexia	176	9.68
Total	**1819**	**100**

As it is evident in the dataset the percentage of *Other* user tagged is really high: 60.8%. The other tags represent less than the 50% and the tags that are going to be used as the accounts to be recommended to the anorexic users are the following: User Recovered From Anorexia, User recovered from other disorders (not anorexia), Media Account giving information about anorexia recovery and User that encourages other users to recover from Anorexia, together they add up to a 28.97% of the dataset. This result requires a balancing of the data to be done since there are more users non-recovered (71.03%) than recovered-helpful (28.97%) and this could bias the classification models. The approach used to balance the data was SMOTE (Synthetic Minority Over-sampling Technique).

The number of users that were identified as accounts to be recommended to the anorexic users was 527 and the other set was of 1292 users. Using SMOTE the over-sampling was done in order to achieve 1292 data samples for each of the two classes above.

The library used for the SMOTE approach was *imbalanced* [2] from python together with the *scikit-learn* [3] library.

4.2 Classification Results

In the present work the tweets retrieved using the anorexia keywords and recovery keywords were fit to a model for classification (using *scikit-learn*). Different approaches were taken into account. Here it follows a list of the classification algorithms used (implemented in *scikit-learn*):

In the following sections it will be shown the results of the accuracy as well as F1-Score and F2-Score. It is important to remark that the accuracy was calculated as the percentage of properly classified instances and that in order to get the best possible values not only for accuracy but for F1-Score and F2-Score, a Grid Search Cross Validation was done for the majority of the algorithms. Here in table 6 are listed the parameters in which the search was done for the different classification algorithms:

Table 6: Classification Grid Search Cross Validation Table

Classification Algorithm	**Parameters Search**
Random Forest	Estimators: [70, 100, 143, 200, 250]
Multinomial NB	Alpha: [0.05, 0.1, 0.2, 0.4, 0.5]
SVM	Alpha: [0.1, 0.01, 0.001] Iterations: [3, 4, 5, 7, 9, 11, 13]
Logistic Regression	Tolerance: [10, 1, 0.1 0.01]
MLP	Activation: ['relu', 'logistic', 'tanh']
AdaBoost	Estimators: [10, 20, 50, 70]

As seen in Table 6 there is no entry point neither for K-NN nor XGBoost since the Grid Search wasn't applied over those algorithms.

The Grid Search was applied to the classification algorithms in order to find the best

[2]https://imbalanced-learn.readthedocs.io/en/stable/
[3]https://scikit-learn.org/stable/

accuracy values by using the different set of parameters given (described in Table 6). The Grid Search implementation from *scikit-learn* was used with a cross validation of three different sets in order to ensure the quality of the output.

For the classification purpose a set of words was chosen from the dataset that did not add any value and they were used as a blacklist of words to remove from the document previous to the TF-IDF processing. Here is the blacklist used: *https*, *http*, *rt*, *co*. In addition to this list, a set of stopwords used for the English language from the *stop-words*[4] library was used too.

4.2.1 Anorexic Users

The following table 7 shows results for anorexic users classification.

Table 7: Anorexic Classification Models Table

Algorithm	**Parameters**	**F1**	**Acc.**	**F2**
K-NN	Neighbors: 5 Metric: Manhattan	80.48%	77.42%	81.85%
Random Forest	Estimators: 100	90.98%	89.86%	90.98%
Multinomial NB	Alpha: 0.1	83.63%	79.26%	89.7%
SVM	Alpha: 0.01 Iterations: 11 Loss Function: Hinge	90.84%	89.4%	**92.38%**
Logistic Regression	Tolerance: 1	**91.8%**	**90.78%**	91.8%
MLP	Activation: Logistic	82.49%	79.26%	85.07%
XGBoost		91.53%	**90.78%**	89.7%
AdaBoost	Estimators: 20	84.26%	82.95%	82.36%

As seen above the obtained results are good, the majority of them are above 80% in terms of accuracy and all of them are above 80% when considering F1-Score and F2-Score.

[4] https://pypi.org/project/stop-words/

Worst results are the ones obtained for the K- Nearest Neighbors classifier, the accuracy does not achieve the 80% and it only gets to 77.42% while it is also the worst in terms of F1-Score and F2-Score, having 80.48% and 81.85% respectively.

Considering the best values: SVM, Logistic Regression, Random Forest and XG-Boost offer some good values. Overall, the best value is for Logistic Regression except for the F2-Score which is the highest for the SVM. It is also worth to highlight that the XGBoost model also has the same accuracy (the highest one) as the Logistic Regression model.

It is easily seen that for this dataset the more convoluted methods are not necessarily the best ones. AdaBoost does not get near to the best values in neither of the three metrics: F1-Score, F2-Score, Accuracy. Also, the Multilayer Perceptron regardless of taking a lot of processing time for the anorexic dataset does not return better results than the approaches mentioned previously.

It is important to highlight that Random Forest also offers good results without reaching the best results obtained for the Logistic Regression Model, having an accuracy value of 89.86%, F1-Score: 90.98% and F2-Score of 90.98%.

The results from above are not exactly the same for the recovery users dataset that will be explained in the next section.

4.2.2 Recovered From Anorexia Users

The following table 8 show results for recovered users classification.

For the recovery dataset, again, the K-NN model offers the worst results, having the following scores: F1-Score is 72.64%, Accuracy is 76.98% and the F2-Score is 65.34 %. The best performing algorithm in this case is the Multilayer Perceptron one with the following values: F1-Score: 89.62%, Accuracy: 88.97% and F2-Score: 92.97%, it is worth noting that the activation function picked is the Relu function.

The other algorithms also showed good results higher than 80%. In the case of this dataset (recovered users) the results were not as good as with the anorexic dataset.

Table 8: Recovery Classification Models Table

Algorithm	**Parameters**	**F1**	**Acc**	**F2**
K-NN	Neighbors: 5 Metric: Manhattan	72.64%	76.98%	65.34%
Random Forest	Estimators: 143	87.52%	87.04%	89.63%
Multinomial NB	Alpha: 0.05	86.74%	85.69%	90.84%
SVM	Alpha: 0.001 Iiterations: 11 Loss Function: Hinge	87.05%	86.65%	88.75%
Logistic Regression	Tolerance: 1	86.25%	85.88%	87.73%
MLP	Activation: Relu	**89.62%**	**88.97%**	**92.97%**
XGBoost		85.71%	85.11%	87.97%
AdaBoost	Estimators: 10	84.6%	83.95%	86.82%

Since the data for recovery users was bigger and mixed because of the reasons explained in 4.1.2 where it is not only one kind of account to be analyzed, it is expected to be harder to analyze from a semantically to a syntactic point of view and making it more complex to fit a classification model.

4.3 Recommender System

4.3.1 Retrieval of Twitter Timelines

After the retrieval of the tweets based on the keywords, the list of anorexic(and recovered) users was stored to retrieve its timelines and be able to find the topics they talk about in order to make relevant recommendations. The goal of getting the timelines was achieved also by using the Tweepy library, with the restriction around 3200 tweets per user timeline.

Anorexic Users

For the 603 anorexic users found it was only possible to retrieve the timelines for 492 users since some accounts were deleted in the process. Given 492 users a total of 810001 tweets were retrieved, averaging around 1600 tweets per user.

Recovered From Anorexia Users

For the 527 recovered users it was only possible to retrieve the timeline for 507 users because other accounts might have been deleted. Using the 507 users it was possible to collect 1104493 tweets. Averaging around 2000 tweets per user.

4.3.2 Timelines Analysis

Given the anorexic and recovered users, the timelines of the users were retrieved considering the Twitter limitations (around 3200 tweets). Two different approaches were taken to group the possible topics of each user considering their timelines. The first step was to concatenate all the user tweets and only considering the timelines with more than 30 tweets, considering this restriction the anorexic users was down to 470 different timelines. For the recovered user the number of timelines with more than 30 tweets was 497.

After having retrieved all the timelines the purpose was to detect a common topic between users of each category (anorexic and recovered ones) in order to match them. The topic detection approach at first was done using the two approaches defined in 2.3.2 and 2.3.2. These two algorithms did not achieve a good set of topics regardless of the executions including from 1-gram to 5-gram. There were a lot of repeated words and the algorithms chose words that did not involve a concept but rather swearing words and anorexic references, which are not the expected topic since it is clear that the anorexic users will use anorexic words. Since the purpose was to detect the topics outside anorexia that the anorexic (and recovered) users use, a different approach was taken using the *Empath*[29] and *TextBlob* libraries [5].

[5]https://textblob.readthedocs.io/en/dev/

Empath is a text analysis tool that enable its users to validate new categories and create new ones on demand starting from a small set of terms. It has around 200 of categories built-in [29]. This library is helpful to find categories (or topics) that are being discussed in a text which is one of the steps needed in the present work in order to be able to achieve the recommender system.

TextBlob is a library that processes text in order to analyze it. In the present work, it is being used to find the text polarity and check if the text has a positive or negative connotation.

4.3.3 User Recommendation

The user recommendation is often based on its similarities. The users will be compared using the categories they talk about. These categories will be found using the *Empath* library which uses 200 general categories as base. Some of the categories could be an issue if taken into account considering that the purpose is to recommend a user to an anorexic user it would not be appropriate to recommend it using as a similar category sadness or some bad emotion. Here it is the list of categories that were not taken into account since it could cause the opposite effect: *contentment*, *restaurant*, *social_ media*, *death*, *injury*, *trust*, *emotional*, *swearing_ terms*, *envy*, *ugliness*, *suffering*, *sadness*, *negative_ emotion*, *positive_ emotion*, *eating*, *pain*, *violence*, *body*, *hate*, *cooking*, *appearance*, *health*, *exercise*, *nervousness*, *shame*.

Recovered users which were a set of 497 with more than 30 tweets were also filtered by the ones having a positive polarity (using *TextBlob*) in order to recommend only the users which usually post positive tweets, since it is not the purpose to provide people with non-positive tweets. After this filtering, the recovered users were down to 494, only 3 recovered users were usually tweeting non-positive tweets.

The crossed calculation of the cosine similarity will be done between 494 recovered user and 470 anorexic users. For each of the anorexic user, let it be *X*. And a recovered user, *Y*, will be the one with the highest cosine similarity out of all the recovered users calculated with the *X* anorexic user. That means, for example: If anorexic-user 1 has a cosine similarity of 0.2 with recovered-user 1 and anorexic-user 1 has a cosine similarity of 0.3 with recovered-user 2 and there are no more users, it is easy to check that the recovered-user 2 will be the one to recommend to anorexic-user 1. This is the approach that was taken in order to recommend a recovered user to all the anorexic users. An illustration of the process is found in Figure 1.

After having done all the recommendations the results were that from the 494 recovered users a set of 84 (17%) users was recommended to the anorexic users. These 84 users are the ones needed to stay relevant and validated in future steps.

4.4 Users Recommendation Validation

One of the main focus of any recommender system is how to actually validate that the recommendation was appropriate. In environments like Twitter this gets even harder since, for the moment, it is impossible to deliver this implementation to the production environment, ergo it is impossible to execute a testing approach which could track if the recommendations done by the system are actually followed by the anorexic user or any approach that involves the production release of the product.

The followed approach will be based on the assumption that the anorexic users should follow people similar to what the users in process of recovery follows, since the purpose of the present work is for the anorexic users to start the process of recovery (and intrinsically be more similar to those users). In this work, the users labeled as Users in Recovery process will be leveraged in order to analyze the users they follow and see how similar are they to the users the system is recommending.

The number of users in process of recovery are 127 as shown on Table 5. Analyzing all the users they are following is not neccesary, a subsample of the followed users is enough. A limit of 50 followed users per each of the 127 users in the process of recovery will be retrieved.

After retrieving of the mentioned above users, 5324 unique users were obtained. Having this number of users, the timelines of them will be retrieved limiting the number of tweets to 100 per user. The total number of tweets retrieved was 504646 tweets, which averages to 95 tweets per user.

After having this information for the 5324 users, the set of categories they talk about was calculated using the same approach than for the Timelines Analysis (see Section 4.3.2) leveraging upon *Empath*.

For the actual validation, the set of 84 categories corresponding to the 84 different recovered users was compared to the 5324 different categories of the followed users by the Users In Process of Recovery. The average cosine similarity was calculated, taking into account all the combinations between recommended users and the newly added categories mentioned above, and the result obtained in average was 0.6502.

The comparison to be done was with a set of random users. The process to pick the random users was leveraged upon the capability of getting a random trending topic and picking one of the last tweets (also randomly) in that trending topic and analyzing the user (who published the tweet) to get its categories. This a way to obtain an actual random user since the trending topics are colorful in regards to the different users.

It is logical that similar users that are labeled in the same category have a higher cosine similarity, for example an anorexic user would probably get the higher cosine similarity with another anorexic user but actually recommending an anorexic user to follow another anorexic user would probably not help in their path to recovery. The same happens for User in Process of Recovery, they are users who might fail in their journey or that are barely just starting it. That is why it is not the wisest choice to recommend them to an anorexic user in order to help them, regardless of the possibility of them having a high cosine similarity. Nevertheless the goal is to achieve the higher cosine similarity possible between an anorexic user and another user that could actually help them in their recovery journey as a recovered user which should definitely in average have a higher cosine similarity (with the anorexic user) than another random user.

This put in an equation form is:

$$sim(A, A) > sim(A, P) > sim(P, REC) > sim(P, RND)$$

Given that purpose, it is possible to obtain the value of the similarities between 1000 categories from random users and the 5324 already obtained before from the followed users by the in process of recovery users. After making this comparison the obtained mean similarity was 0.395 which is actually a lower value than 0.6502 proving than in average the recommendation being done is better than a random recommendation. The median comparison was also done corresponding to a value of 0.329 for the random users and 0.67 for the recommended users getting an even bigger difference.

This process is illustrated in the Figure 2 except the initial comparison of the recovery users with the anorexic users since that is the actual recommendation and the purpose in this section is to illustrate how to validate that recommendation.

Chapter 5

Conclusion

In this work, different approaches were taken in order to be able to recommend recovered users to currently anorexic ones. As well as different classification models that performed above the 80% threshold in average for F1-Score, F2-Score and Accuracy. There were topic detection approaches that were also used in another papers like NMF and LDA. Other topic detection techniques were based on Third-Party libraries like *Empath* achieving more logical results.

The first step was to retrieve potentially anorexic users based on a list of keywords. After the retrieval, the purpose was to get the users tagged between a set of possible labels where the interesting ones were the users tagged as *User with Anorexia*, where the process of tagging was done using the help of workers from Amazon Mechanical Turk. The same process was followed for the recovered users where the main difference was that there were a set of labels that were interesting instead of only one like *User recovered from Anorexia*, *Media Account giving information about anorexia*, etc.

After having all the users properly tagged the following step in the process of the present work was to match an Anorexic user with a Recovered user, there were several approaches to achieve this. The first approach taken was to use Topic Detection Techniques (NMF and LDA) to find what concepts the people talk about. This approach was unsuccessful since the topics found were not relevant. The second approach to achieve a characterization of the user was based on the *Empath* library. This library can offer a score of how related to a category is a document, where the document is created based on the timelines of the users. Having over 200 categories this library offered a different approach to the problem.

With all the categories for both anorexic users and recovered users the cosine similarity between them was calculated and the match between the users across categories with the maximum similarity was the one provided to that anorexic user. A set of different recovered users was recommended (84 recovered users).

The last step was the validation and relevance of the recommended users. The decision taken for this step was to check the users that the *Users in Process of Anorexia Recovery* follow. After the proper analysis about this new set of users, it was shown that the recommended users had a bigger similarity measure with this set that random users by calculating the cosine similarity between the recommended users and the users followed by the users that already started their recovery and getting the mean, this mean was higher than the mean obtained using random users obtained from twitter (in average).

This validation was done based on the fact that the purpose is to recommend users similar (taking into account categories they talk about and calculating the cosine similarity) to what the users that are in process of recovery follow. This way, it is expected that the anorexic users start following (by being recommended to them) the recovered users which are similar to what the users that already started that journey follow, since it would be ideal that the anorexic users start their recovery process.

The present work was done with the purpose of analyzing the Twitter data in order to look for a way to help people by leveraging the available information and recommend another user that can improve their life by helping them overcome an eating disorder as anorexia.

Chapter 6

Further Research

For the present work, different approaches were taken but there are many more of interest for the research community.

First, the dataset could be leveraged for several purposes even by creating a simpler dataset as only the tweet and a binary label that would describe if the user is anorexic or not (could be useful as baseline for different research in classification). Another approach is to leverage users talking about anorexia and users talking about different topics.

The classification could be improved by testing different approaches and more features in order to find better fits, given that the dataset is not big the only approach that does not seem to be as valuable is having a deep neural network because it could be an overkill for the problem resulting in an overfitting.

The recommendation part can be done considering an entirely different set of topics. In this work, a library was used for that given the time limitations for the work but different and better techniques of topic detection for the given tweets can be applied and this could actually improve the value of the users recommended. Also, not only the way the categories or topics are being obtained is possible for future research but the different similarity measures that exist and that can be applied in order to improve the similarity between users.

Another approach for validation is to actually do a temporal split of the time for testing for the system if it could be deployed to production on Twitter. In this way it would be possible to actually evaluate in real life how well the suggestion of the users would be in *Twitter* since it would be possible to see the users that the anorexic user started following in a point of the time. It is worth highlighting that in order to achieve this a lot of joint work with Twitter to get metrics of user suggestions to the specific case of possibly detected users inside the network. This would be a very complex work since a lot of alignment with Twitter (and even privacy issues might come up) would be required.

List of Figures

List of Tables

Bibliography

[1] Eating disorders. URL `https://www.nimh.nih.gov/health/topics/eating-disorders/index.shtml`.

[2] Arseniev-koehler, A. *et al.* # Proana : Pro-Eating Disorder Socialization on Twitter. *Journal of Adolescent Health* **58**, 659–664 (2016). URL `http://dx.doi.org/10.1016/j.jadohealth.2016.02.012`.

[3] Tech, G., Tech, G. & Tech, G. Recovery Amid Pro-Anorexia : Analysis of Recovery in Social Media 2111–2123 (2016).

[4] Juarascio, A. S., Shoaib, A. & Timko, C. A. Pro-eating disorder communities on social networking sites: A content analysis. *Eating Disorders* **18**, 393–407 (2010).

[5] Wang, T., Brede, M., Ianni, A. & Mentzakis, E. Social interactions in online eating disorder communities : A network perspective 1–17 (2018).

[6] Cavazos-Rehg, P. A. *et al.* just want to be skinny.A content analysis of tweets expressing eating disorder symptoms. *PLoS ONE* **14**, 1–11 (2019).

[7] Tiggemann, M., Churches, O., Mitchell, L. & Brown, Z. Tweeting weight loss: A comparison of #thinspiration and #fitspiration communities on Twitter. *Body Image* **25**, 133–138 (2018). URL `https://doi.org/10.1016/j.bodyim.2018.03.002`.

[8] Talbot, C. V., Gavin, J., van Steen, T. & Morey, Y. A content analysis of thinspiration, fitspiration, and bonespiration imagery on social media. *Journal of Eating Disorders* **5**, 1–8 (2017).

[9] Ibrahim, R., Elbagoury, A., Kamel, M. S. & Karray, F. Tools and approaches for topic detection from Twitter streams : survey. *Knowledge and Information Systems* **54**, 511–539 (2018).

[10] McLaughlin, M., Park, M. & Sun, Y. Sharing information promoting unhealthy behavior through social media: Examination of the users and the content using the case of pro-anorexia in tumblr. *Proceedings - 2015 International Conference on Computational Science and Computational Intelligence, CSCI 2015* 800–803 (2016).

[11] Sowles, S. J. *et al.* A content analysis of an online pro-eating disorder community on Reddit. *Body Image* **24**, 137–144 (2018). URL `http://dx.doi.org/10.1016/j.bodyim.2018.01.001`.

[12] Tiggemann, M. & Zaccardo, M. Strong is the new skinny: A content analysis of #fitspiration images on instagram. *Journal of Health Psychology* **23**, 1003–1011 (2018). URL `https://doi.org/10.1177/1359105316639436`. PMID: 27611630, `https://doi.org/10.1177/1359105316639436`.

[13] Altman, N. S. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician* **46**, 175–185 (1992).

[14] Breiman, L. Random forests. *Machine Learning* **45**, 5–32 (2001). URL `https://doi.org/10.1023/A:1010933404324`.

[15] Russell, S. & Norvig, P. *Artificial Intelligence A Modern Approach Third Edition* (2010). `9809069v1`.

[16] Cortes, C. & Vapnik, V. Support-vector networks. *Machine Learning* **20**, 273–297 (1995). URL `https://doi.org/10.1007/BF00994018`.

[17] Cramer, J. The Origins of Logistic Regression. *Ssrn* (2002).

[18] Cox, D. R. The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)* **20**, 215–242 (1958).

[19] *Artificial Neural Networks* (McGraw-Hill Education (India) Pvt Limited). URL `https://books.google.com/books?id=bEpt-eBZ1JYC`.

[20] Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System (2016). URL `http://arxiv.org/abs/1603.02754{%}0Ahttp://dx.doi.org/10.1145/2939672.2939785`. `1603.02754`.

[21] Freund, Y. & E Schapire, R. A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence* **14**, 771–780 (1999).

[22] Chawla, N.V., Bowyer, K.W., Hall, L.O., K. W. SMOTE: Synthetic Minority Over-Sampling Technique. Journal of Artificial Intelligence Research. *Journal of Artificial Intelligence Research* **16**, 321–357 (2002). `1106.1813`.

[23] Stone, M. Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society. Series B (Methodological)* 111–147 (1974).

[24] Manning, C. D., Raghavan, P. & Schütze, H. *Introduction to Information Retrieval* (Cambridge University Press, New York, NY, USA, 2008).

[25] Blei, D. M. *et al.* technique...Latent Dirichlet Allocation. *Journal of Machine Learning Research* **3**, 993–1022 (2003). `1111.6189v1`.

[26] Ricci, F., Rokach, L. & Shapira, B. *Recommender Systems Handbook* (2015). URL `http://link.springer.com/10.1007/978-1-4899-7637-6`.

[27] Jiawei Han, Micheline Kamber, J. P. *Data Mining Concepts And Techniques* (2011). `arXiv:1011.1669v3`.

[28] Amazon mechanical turk. URL `https://www.mturk.com/`.

[29] Fast, E., Chen, B. & Bernstein, M. Empath: Understanding Topic Signals in Large-Scale Text (2016). URL `http://arxiv.org/abs/1602.06979{%}0Ahttp://dx.doi.org/10.1145/2858036.2858535`. `1602.06979`.