



Norwegian University of  
Science and Technology

# Classification of Pro-Eating Disorder Users on Twitter

**Ingrid Nelson Giæver**

Master of Science in Computer Science

Submission date: June 2018

Supervisor: Björn Gambäck, IDI

Norwegian University of Science and Technology  
Department of Computer Science



# Abstract

Pro-eating disorder (pro-ED) refers to online communities endorsing engagement with disordered eating behaviour. On the microblogging site Twitter, members of pro-ED communities share messages (tweets) that glorify extreme thinness or portray unhealthy weight control methods as lifestyle choices rather than symptoms of a mental illness. The goal of this thesis was to achieve automatic detection of users taking part in pro-eating disorder communities on the Twitter platform.

For the purpose of this study, 7096 users and 10.7M tweets were collected from Twitter and manually annotated. The data set included users taking part in pro-ED communities and users whose tweets were either recovery-oriented or unrelated to eating disorders. Analysis of the data set revealed differentiating characteristics in the users' tweets and profile information, with respect to emoji use, presence of URLs and user mentions, and references to eating disorders and related topics.

Based on the established differences, groups of features, such as tweet  $n$ -grams and emojis, were extracted and used to train a series of supervised classifiers. Four machine learning models were explored; a Support Vector Machine, a Naïve Bayes model, a Logistic Regression model and a Random Forest. The highest  $F_1$ -score (0.98) was achieved both when using an SVM and when using an ensemble approach trained on weighted feature groups with emphasis on unigrams from tweets.

# Sammendrag

Pro-eating disorder (Pro-ED) refererer til online subkulturer som oppfordrer til spiseforstyrrelser. Miljøene sprer omstridt innhold, som eksempelvis glorifiserte bilder av anorektiske kropper, usunne dietter eller tips til hvordan man kan skjule ekstreme spisevaner fra venner og familie. Denne masteroppgaven tok for seg pro-ED miljøer på nettsamfunnet Twitter, og siktet mot å oppnå automatisk klassifisering av pro-ED brukere på denne plattformen.

For dette formålet ble 7096 brukere og totalt 10.7 millioner av deres delte meldinger (tweets) nedlastet. Brukerne i datasettet inkluderte både medlemmer av pro-ED miljøer, brukere som diskuterte spiseforstyrrelser med fokus på mental og fysisk helse, samt brukere som var urelatert til problematikken. Gjennom analyse av tweets og profinformasjon hos brukerprofilene i datasettet, ble flere differensierende kjennetegn oppdaget, blant annet innen bruk av emoji'er, URLer og referanser til spiseforstyrrelser og relatert tematikk.

Basert på denne analysen ble flere sett med «features» dannet og benyttet til opptrening av ulike maskinlæringsalgoritmer. Oppgaven testet ut fire ulike modeller; en Support Vector Machine, en Naïve Bayes modell, en logistisk regresjonsmodell og en Random Forest. Den beste  $F_1$ -scoren ble oppnådd ved bruk av en SVM eller en kombinasjonsmodell trent opp på vektete feature-sett hvor unigrams fra tweets ble gitt størst påvirkningskraft.

# Preface

This thesis was written during the spring of 2018, as a part of my Master of Science (MSc) degree in Computer Science at the Department of Computer Science (IDI) at the Norwegian University of Science and Technology (NTNU).

I would like to thank my wonderful supervisor, Björn Gambäck, for his thorough feedback and enthusiasm for my choice of topic for this project. I am also grateful to the control annotators for volunteering to help me. In addition, I would like to thank my amazing flatmates of five years for making my time in Trondheim so special.

Ingrid Nelson Giæver  
Trondheim, 11th June 2018

# Contents

## 1. Introduction

1.1	Motivation .....	1
1.2	Twitter .....	2
1.3	Pro-Eating Disorder .....	3
	1.3.1 Eating Disorders .....	3
	1.3.2 Pro-ED Communities .....	4
	1.3.3 Thinspiration .....	4
	1.3.4 Initiatives Against Pro-ED Communities .....	5
1.4	Project Goal and Thesis Description .....	7
1.5	Research Methods .....	8
1.6	Contributions .....	9
1.7	Considerations .....	9
1.8	Thesis Structure .....	10

## 2. Background

2.1	Machine Learning for Text Classification .....	11
	2.1.1 Naïve Bayes Classifiers .....	11
	2.1.2 Logistic Regression .....	12
	2.1.3 Decision Trees and Random Forests .....	13
	2.1.4 Support Vector Machines .....	15
	2.1.5 Classification Scoring Metrics .....	17
2.2	Text Representations .....	18
	2.2.1 <i>N</i> -grams .....	19
	2.2.2 Bag of Words .....	19
	2.2.3 Term Frequency - Inverse Document Frequency .....	20
	2.2.4 Stop Word Removal .....	22
	2.2.5 Part-of-Speech Tagging .....	22
2.3	Manual Annotation and Inter-Annotator Agreement .....	23
	2.3.1 Cohen's kappa .....	24
	2.3.2 Fleiss' kappa .....	24
2.4	Tools .....	24
	2.4.1 Twitter API .....	24

	2.4.2 Scikit-Learn .....	26
	2.4.3 Natural Language Toolkit .....	26
<b>3. Related Work</b>		
3.1	Studies on Pro-ED .....	27
3.2	Twitter User Classification .....	29
	3.2.1 Mental Health .....	30
	3.2.2 Subcultures .....	31
3.3	Relevant Approaches to Building a Classifier .....	32
	3.3.1 Data Collection .....	32
	3.3.2 Pre-Processing .....	33
	3.3.3 Feature Extraction .....	34
	3.3.4 Models .....	36
<b>4. Data</b>		
4.1	Characterising «Pro-Eating Disorder» Users .....	37
4.2	Pro-Recovery & Unrelated users .....	40
4.3	Data Collection Procedure .....	42
4.4	Data Set .....	43
4.5	High-Level Filtering and Pre-Processing .....	44
4.6	Annotations .....	46
4.7	Characteristics .....	48
	4.7.1 Twitter Functionality and Internet Terms .....	48
	4.7.2 Emojis .....	49
	4.7.3 Locations .....	52
	4.7.4 Tweet length .....	53
	4.7.5 Part-of-Speech .....	54
	4.7.6 References to Eating Disorders and Related Topics .....	54
<b>5. Architecture</b>		
5.1	Pre-Processing Continuation .....	59
	5.1.1 Tweet Aggregation .....	60

5.1.2	Pre-Processing in Scikit-Learn.....	61
5.2	Feature Extraction .....	61
5.2.1	Feature Extraction in Scikit-Learn .....	62
5.2.2	Feature Groups.....	62
5.3	Classifiers .....	63
<b>6. Experiments &amp; Results</b>		
6.1	Experimental Setups .....	65
6.2	Interpretation of Results.....	66
6.3	Feature Study .....	66
6.3.1	Unigrams from tweet text.....	67
6.3.1	Bigrams from tweet text .....	69
6.3.3	Emojis .....	70
6.3.4	Biographies.....	72
6.3.5	Names .....	74
6.4	Combining Feature Groups .....	75
6.5	Test Set Results.....	78
<b>7. Discussion &amp; Conclusion</b>		
7.1	Evaluation of Results .....	81
7.2	Evaluation of Research Questions.....	82
7.3	Ethical Considerations .....	84
7.4	Future work and Limitations of the study .....	85
<b>References .....</b>		<b>89</b>
<b>Image references .....</b>		<b>97</b>
<b>A. Pro-Eating Disorder Terms .....</b>		<b>99</b>
<b>B. Data</b>		
B.1	Data Collection .....	101
B.2	Keywords for Data Collection.....	101
B.3	Part-of-Speech .....	102
B.4	Most Popular Locations .....	103
<b>C. Architecture</b>		
C.1	Stop Words .....	105
C.2	Vocabulary sizes .....	106



## List of Figures

1.1	Examples of thinspiration content. ....	5
1.2	Search results on Instagram, Tumblr and Pinterest, given the keyword <i>anorexia</i> . ....	6
2.1	Decision tree example. ....	13
2.2	Example of a Random Forest Classifier ....	14
2.3	Example of a Support Vector Machine with data points in a two-dimensional feature space ....	16
2.4	Confusion Matrix ....	17
2.5	Word unigrams, bigrams and trigrams from a song lyric ....	19
2.6	Part-of-Speech example ....	23
2.7	Example Twitter user with data fields ....	25
4.1	Distribution of total users and tweets in the data set ....	44
4.2	Distribution of Internet and Twitter terms ....	48
4.3	Presence of popular emojis in tweets. ....	49
4.4	Presence of popular emojis in users' tweet sets ....	50
4.5	Presence of popular emojis among pro-ed users ....	51
4.6	Presence of popular emojis among pro-recovery users ....	51
4.7	Presence of popular emojis among unrelated users ....	51
4.8	Average tweet lengths in number of words ....	53
4.9	Average tweet lengths in number of characters ....	53
4.10	Average distribution of POS-tags in tweets. ....	54
4.11	References in tweets. ....	56
4.12	References in biographies. ....	57
4.13	References in usernames or display names ....	58
5.1	Classification system ....	60
6.1	Contribution to the $F_1$ -score of the pro-ed class ....	76
6.2	Contribution to the $F_1$ -score of the not pro-ed class ....	76

6.3	Confusion matrix for the voting classifier based on 5-fold cross-validation .....	78
6.4	Confusion matrix for the voting classifier on the test set.....	79

## List of Tables

1.1	Examples of restricted hashtags in popular social networks .....	6
2.1	Example vocabulary .....	20
3.1	Arseniev-Koehler <i>et al.</i> 's (2016) ED Reference Codebook .....	28
3.2	Top 10 most indicative image tags on Flickr according to Yom-Tov <i>et al.</i> (2012) .....	29
2.5	Support Vector Machine example .....	16
2.6	Confusion Matrix .....	17
2.7	<i>N</i> -grams example .....	19
2.8	POS tagging example .....	22
2.9	Twitter user example .....	25
4.1	Example tweets and associated categories .....	41
4.2	Example of keywords used in searches .....	42
4.3	Raw data set .....	43
4.4	Placeholders .....	46
4.5	Pairwise Cohen's kappa .....	47
4.6	Fleiss' kappa .....	47
4.7	Codebook .....	55
6.1	Unigrams: Classification results based on 5-fold cross-validation ...	68
6.2	Unigrams with highest weights .....	69
6.3	Unigrams with lowest weights .....	69
6.4	Bigrams: Classification results based on 5-fold cross-validation .....	69
6.5	Bigrams with highest weights .....	70
6.6	Bigrams with lowest weights .....	70
6.7	Emojis: Classification results based on 5-fold cross-validation .....	70
6.8	Emoji features with highest weight coefficients .....	71
6.9	Emoji features with lowest weight coefficients .....	71
6.10	Biography: Classification results based on 5-fold cross-validation ...	72

6.11	Biography features with highest weights . . . . .	73
6.12	Biography features with lowest weights . . . . .	73
6.13	Names: Classification results based on 5-fold cross-validation . . . . .	72
6.14	Highest weighted name features . . . . .	74
6.15	Lowest weighted name features. . . . .	74
6.16	Combination: Classification results based on 5-fold cross-validation	75
6.17	Weighted Combination: Classification results based on 5-fold cross -validation . . . . .	77
6.18	Ensemble Weighted Combination: Classification results based on 5 -fold cross-validation. . . . .	78
6.19	Classification results on unseen test data . . . . .	79
A.1	Medical terms and diagnoses . . . . .	99
A.2	Explanations of Pro-ED terms and phrases . . . . .	99
B.1	Sampling tag words . . . . .	101
B.2	Popular locations . . . . .	103
C.1	Stop words . . . . .	105
C.2	Unigram vocabulary size . . . . .	107
C.3	Bigram vocabulary size . . . . .	108
C.4	Biography features vocabulary size. . . . .	109
C.5	Name features vocabulary size . . . . .	110





# 1. Introduction

Pro-eating disorder (pro-ED) refers to online communities endorsing engagement with disordered eating and whose members use social media to share content that encourages dangerous behaviour and body ideals. This thesis' area of interest is the pro-eating disorder communities on the microblogging site, Twitter. Pro-ED users' tweets and profile information were investigated and classifiers were built for the purpose of detecting pro-ED users on the platform.

This introductory chapter presents the motivation behind the study, and includes an introduction to the microblogging site, Twitter, and the phenomenon of online pro-eating disorder communities. It also discusses the current restriction policies used by different social media to limit the presence of the content shared by members of such communities. The final sections of this chapter present the goal and research questions for the study, research methods, contributions and important considerations.

## 1.1 Motivation

Social media sites, such as Twitter, have provided unique opportunities for online communication and are increasingly used to share ideas, opinions, information and personal messages. For people suffering mental illnesses, such online platforms can create conducive environments to get in touch with others who share similar difficulties, provide access to emotional support and advice, and work against stigmatisation (Betton *et al.*, 2015). However, social media also connect people in ways that could amplify the destructiveness of some mental illnesses.

Pro-eating disorder (pro-ED) are online movements supporting engagement with an eating disorder lifestyle. Previous research has shown that exposure to pro-eating disorder content is associated with increased body dissatisfaction (Bardone-Cone & Cass, 2007), eating disorder identity reinforcement (Giles, 2006) and acquisition of unhealthy

## 1. Introduction

weight reducing methods (Wilson *et al.*, 2007; Ransom *et al.*, 2010). Furthermore, it is assumed that the content has detrimental effect on the treatment progress of individuals with eating disorders, could trigger disordered eating behaviour and contribute to relapse among people who have suffered eating disorders in the past (Reel, 2013, p. 366-368).

Popular social media, including Instagram<sup>1</sup>, Pinterest<sup>2</sup> and Tumblr<sup>3</sup>, have tried to limit the amount of pro-eating disorder content by banning related hashtags or by displaying advisory content in response to searches. However, a study of the aftermath of Instagram's censoring policy showed that the communities had adopted lexical variant tags, and concluded that the content moderation had mostly been ineffective at decelerating the dissemination of pro-eating disorder content (Chancellor *et al.*, 2016).

Identification of users taking part in pro-ED communities could provide useful information in the understanding of eating disorders and online behaviour. Moreover, better insight into these topics could guide both health officials and social media sites to find methods to tackle the issues related to the online communities. Detection of pro-eating disorder users could also be of interest for potential filtering strategies on social media platforms.

### 1.2 Twitter

Twitter<sup>4</sup> is an online micro-blogging site, attracting millions of users on daily basis. As of April 2018, Twitter has 336 million monthly active accounts (Twitter, 2018), varying from average citizens to celebrities and from charity organisations to large companies.

Twitter allows users to post short messages known as «tweets». From Twitter's start-up in 2006, tweets were limited to include a maximum of 140 characters; however, this limit

---

<sup>1</sup> <https://www.instagram.com/about/us/>

<sup>2</sup> <https://about.pinterest.com/en>

<sup>3</sup> <https://www.tumblr.com/about>

<sup>4</sup> <https://about.twitter.com/>



was increased to 280 characters in November of 2017<sup>5</sup> (Rosen, 2017). Tweets form the basis of the interactions on Twitter, and users follow other profiles in order to keep up with their feeds of tweets. Twitter also offers functionality to forward other users' tweets, referred to as re-tweeting. By using hashtags, i.e., words prefixed by #-sign, users can categorise tweets based on topic, connecting the tweet to a larger online conversation.

The platform's ability to let people express themselves through tweeting and re-tweeting, discover and follow other users and attract an audience, makes it a suitable online space for communities of like-minded individuals. Although Twitter has the potential to expose users to diverse voices from all over the world, many users have a tendency to seek intragroup communication and choose not to follow user profiles with contradicting views or opinions. Hence, Twitter also creates favourable conditions for creations of isolated communities and intensification of group beliefs, a phenomenon often referred to as echo chambers (Carr, 2017).

### **1.3 Pro-Eating Disorder**

The expression, pro-eating disorder (pro-ED), refers to the promotion of disordered eating behaviour and is used to describe online communities that support engagement with an eating disorder lifestyle (Arseniev-Koehler *et al.*, 2016). This section gives an overall description of this phenomenon.

#### **1.3.1 Eating Disorders**

Eating disorders are serious and complex illnesses that are characterised as severe disturbance to a person's eating behaviour. Individuals with an eating disorder will typically experience having an obsessive relationship to food, body shape and weight. The most common forms of eating disorders include anorexia nervosa, bulimia nervosa and binge eating disorder (The National Institute of Mental Health, 2016).

Eating disorders may lead to major health problems, including osteoporosis, infertility, organ failure and heart and brain damage. Anorexia nervosa is responsible for the highest mortality rate among mental illnesses (The National Institute of Mental Health, 2016).

---

<sup>5</sup> With the exception of tweets written in Chinese, Korean and Japanese.

## 1. Introduction

### 1.3.2 Pro-ED Communities

Online pro-ED communities have flourished since the early 2000s, as the internet and social media have grown in scope and popularity. The most popular pro-ED communities are the *pro-ana* community, supporting engagement with the eating disorder anorexia nervosa, and *pro-mia*, likewise referring to the community endorsing bulimia nervosa. The two communities are often simply referred to as *ana* and *mia*.

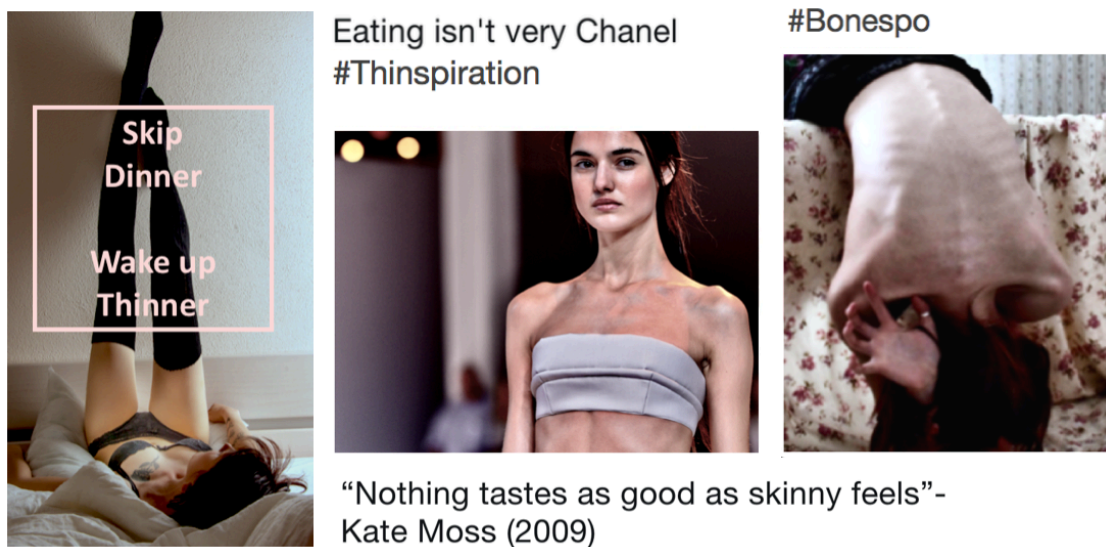
The content posted in pro-ED communities usually glorifies extreme thinness, or portrays unhealthy weight control methods as lifestyle choices rather than symptoms of a mental illness. Members of the communities might share extreme diets, advices on how to suppress hunger or how to hide symptoms from friends and family. Motivational quotes and mantras are often used to endorse eating disorders as an accomplishment of self-control, and to create the sense of a common identity surrounding the illness. However, individuals who identify as a part of such communities differ in their stance. While some claim there is no such thing as an eating disorders, others openly share their struggles with their mental illness. Many users also seek the pro-ED communities for emotional support and a place not to be judged.

### 1.3.3 Thinspiration

Thinspiration (a combination of «thin» and «inspiration») signifies thin-ideal media content and is often seen in combination with, or as a part of, pro-ED communities. The term is commonly used as a hashtag for images, often with descriptions encouraging weight loss, starvation or providing food guilt (Boepple & Thompson, 2015). On the social media platform Instagram, the majority of images tagged with #thinspiration (or #thinspo), or the more recent tag #bonespiration (or #bonespo), features thin, female bodies. Images tagged with #bonespiration show fewer muscles and more bone protrusions, indicating that it may represent an exaggerated form of thinspiration (Talbot *et al.*, 2017).

Thinspiration can also refer to textual content that aims to inspire thinness, including song lyrics, celebrity quotes, etc. Figure 1.1 displays examples of thinspiration content.

**Figure 1.1:** Examples of thinspiration content<sup>6</sup>



#### 1.3.4 Initiatives Against Pro-ED Communities

Ever since pro-ED websites started to gain publicity, efforts to eradicate such sites have been made. The first to ban pro-ED websites was Yahoo, back in 2001 (Holahan, 2001). However, the attempts to eliminate the presence of the content online did not succeed, and as social media grew larger, so did the pro-ED communities. In 2015, the French government put restrictive legislations in place, making the act of «provoking people to excessive thinness by encouraging prolonged dietary restrictions that could expose them to a danger of death or directly impair their health» a criminal offence that could carry a sentence of up to one year's imprisonment and a fine of €10.000 (Saul, 2015).

In 2012, the social networking sites Tumblr, Pinterest and Instagram, all updated their terms of service in order to take action against the controversial content (Tumblr, 2012; Pails, 2012; Hasan, 2012). As of March 2018, their restriction policies involve suspending users, banning certain hashtags, or providing advisory content in response to user searches. Table 1.1 displays examples of pro-eating disorder related hashtags that currently have restricted access on popular social media platforms, and figure 1.2 shows

<sup>6</sup> All photos licensed under CC-BY 2.0 (See image references)  
<https://creativecommons.org/licenses/by/2.0/>

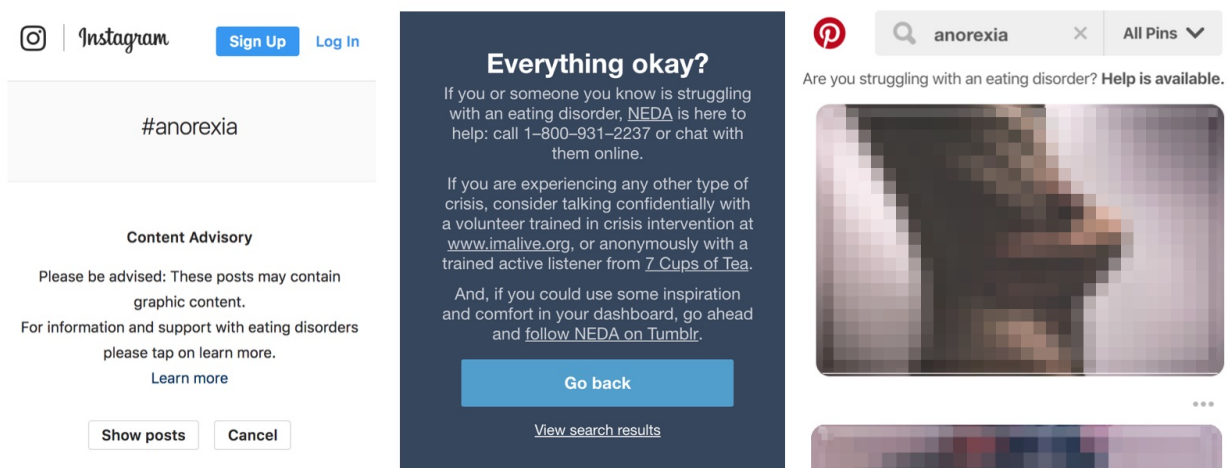
## 1. Introduction

the advisory content presented to users searching for such hashtags on Instagram, Tumblr and Pinterest. The former warns the user about potential graphic content, and all three refers the user to the National Eating Disorders Association (NEDA). For explanations of popular pro-ED terms and hashtags, see appendix A.

**Table 1.1:** Examples of restricted hashtags in popular social networks.<sup>7</sup>

Platform:	Banned	Advisory
Instagram	#loseweight #proana #proanorexia #promia #probulimia #thinspiration #thinspo #thighgap	#abcdiet #anabuddy #anamia # <b>anorexia</b> #bulimia #collarbones #dyingtobethin #eatingdisorder #ednos #hipbones #meanspo #purge #thygap #starve
Pinterest	-	#ana # <b>anorexia</b> #bulimia #mia #proana #promia
Tumblr	-	#ana # <b>anorexia</b> #bulimia #mia #proana #promia #skinny #starvation #thin #thinspo #thinspiration #thighgap
Twitter	-	-

**Figure 1.2:** Search results on Instagram, Tumblr and Pinterest, given the keyword *anorexia*<sup>8</sup>.



<sup>7</sup> As of March 2018

<sup>8</sup> Screenshots taken March 2018. (See image references)

Twitter regards glorifying self-harm, including eating disorders, as violations of their general guidelines and policies<sup>9</sup>. They state that users who repeatedly violate their policy will be considered suspended. However, Twitter has not banned, or restricted the access to, any specific pro-ED related hashtags.

## 1.4 Project Goal and Thesis Description

The goal of this thesis was to research how to achieve automatic detection of users taking part in pro-eating disorder communities on Twitter using machine learning.

**Goal:** *Identify pro-eating disorder (pro-ED) users on Twitter.*

Given a Twitter user's profile information and past tweets, this work aimed to classify the user as either a participant in a pro-ED community or not. The following research questions were addressed:

**RQ1:** *How is the Twitter platform used by members of pro-ED communities, and what criteria should be used in annotation of such users.*

In order to collect and annotate a representative dataset for the purpose of the thesis, it was essential to gather information about how members of pro-ed communities use social media and establish a way of evaluating whether a user is considered to take part in such a community or not.

**RQ2:** *What does previous research establish as useful methods and features for classification of user generated, textual data with respect to mental health or online subcultures?*

To understand how accurate classification of users could be achieved, previous studies on related topics were examined. A review of such studies would gather relevant information on what kind of machine learning methods and tools to use, and challenges to be taken into account, when building a user classifier.

---

<sup>9</sup> Twitter's general guidelines and policies: «Glorifying self-harm and suicide»  
<https://help.twitter.com/en/rules-and-policies/glorifying-self-harm>

## 1. Introduction

**RQ3:** *What characterises the tweets and profile information of users taking part in pro-ED communities on Twitter?*

For the sake of differentiating pro-ED users from the general population on Twitter, an understanding of what aspects of text and profile data that characterises the pro-ed community members was needed.

**RQ4:** *What methods and features are useful for the classification of pro-eating disorder users on Twitter.*

Finally, in order to achieve the goal of the study, different features and machine learning methods were evaluated for the sake of designing an efficient classification system for detection of pro-eating disorder users on Twitter.

### 1.5 Research Methods

To achieve the goal presented above, several methodologies were used. In order to answer the first two research questions, a study of relevant literature concerning pro-ED communities and previous work in the field of user classification on Twitter was conducted. Based on the results from the literature review, a set of inclusion criteria was defined in order to evaluate whether a user is considered to take part in a pro-eating disorder community or not. These criteria were then used to collect and annotate a set of Twitter users.

To answer the third research question, regarding characteristics of pro-ED users, data analyses were conducted on the training proportion of the data set, and the results were used to decide on feature groups for the training of the classification models. Four different machine learning models were explored, including a Support Vector Machine, a Logistic Regression model, a Naïve Bayes model and a Random Forest. The informativeness of different feature groups was also considered. The best performing models, including an ensemble model, were finally tested on unseen data in order to evaluate the performance of the system.

## 1.6 Contributions

This work contributes to the discussion of online pro-eating disorder culture, by offering efficient methods, using machine learning approaches, to automatically identify users taking part in pro-ED communities on Twitter. It also provides insight into the characteristics of profile information and the shared content of pro-ED users.

Additionally, it contributes to further research with a data set of 7096 Twitter users and 10.7 million tweets, annotated as either pro-ED, pro-recovery (i.e., discussing eating disorders and related topics with a recovery-oriented focus) or unrelated to eating disorders.

## 1.7 Considerations

Twitter is a widely used public site, especially among adolescents, and censorship of pro-ED content is a heated and much-discussed topic. As stated above, similar sites have conducted actions to minimise the exposure to this type of content. However, it is argued that censorship might not be the best solution to the problem (Casilli *et al.*, 2013; Chancellor *et al.*, 2016). Restrictions may lead to further spreading of the content onto other online platforms, making it more difficult for physicians, families and charities to reach out to the pro-ED users. It could potentially also lead to increased stigmatisation of those suffering eating disorders. This work does not take a stand in this discussion, and is not to be taken as a statement towards censorship of pro-ED content.

It is important to emphasise that the classification of a user as «pro-eating disorder» is not a diagnostic claim of eating disorder.

Please note that the thesis does contain some graphic quotations and images for exemplary purposes. Example tweets have been modified for the sole purpose of removing potential personally identifiable information.

## 1.8 Thesis Structure

This thesis is structured as follows:

**Chapter 2** covers relevant background theory, used or references in this thesis, regarding pro-eating disorder communities, text-processing, machine learning and tools.

**Chapter 3** presents related work on pro-ED communities in social media and classification of Twitter users with respect to mental health and participation in sub-cultures.

**Chapter 4** describes the process of collecting data and presents results from a data analyses conducted on the found data set.

**Chapter 5** covers the architecture of the classification system.

**Chapter 6** describes the experimental set-up and investigates the informativeness of feature groups and the performance of different machine learning models.

**Chapter 7** concludes the thesis by discussing the found results, ethical considerations, limitations of the study and potential future work.



## 2. Background

This chapter covers background theory and information relevant to this project, and serves as an introduction to the terminology, regarding Machine Learning (ML) and Natural Language Processing (NLP), that are used in this thesis. First, section 2.1 presents popular machine learning methods and classification metrics, while the following section covers relevant theory in the field of natural language processing. Section 2.3 then describes the concept of manually annotations. Finally, section 2.4 gives a brief overview of some popular technological tools and resources related to Twitter data collection and text classification.

### 2.1 Machine Learning for Text Classification

Text classification aims to identify to which of a set of pre-defined classes a document belongs, and is the core problem in many applications, such as spam detection systems or sentiment analysis. Most modern text classification systems incorporate some form of supervised machine learning (ML), that is, learning to classify instances from a set of already classified training samples. A supervised machine learning algorithm analyses the training examples and produces a function that later can be used for classification of new unseen documents.

This section covers relevant theory from the field of supervised machine learning, and briefly introduces different methods referenced in this thesis or in reviewed literature. The final subsection presents metrics to evaluate the performance of a text classification system.

#### 2.1.1 Naïve Bayes Classifiers

Naïve Bayes (NB) classifiers are simple probabilistic classifiers based on Bayes' Theorem, which describes the probability of an event based on prior knowledge. The method was first introduced in the early 1960s, but remains popular in automated text classification.

## 2. Background

A Naïve Bayes model (naively) assumes that each input feature of a document is conditionally independent of all the other. More formally, if a document,  $d$ , belonging to some class,  $c$ , is represented as a vector of its features,  $\mathbf{x} = (x_1, \dots, x_n)$ , the Naïve Bayes model assumes the following holds while calculating the probability of observing this feature vector:

$$p(x_1, \dots, x_n | c) = \prod_i^n P(x_i | c)$$

As a result of this assumed conditional independence, the probability of a document with a feature vector,  $\mathbf{x}$ , belonging to a class  $c$ , can be written as:

$$p(c | x_1, \dots, x_n) = P(c) \prod_i^n P(x_i | c)$$

For the actual class prediction, the Naïve Bayes model returns the class option that maximises the product of the conditional probabilities:

$$c_{\text{NB}} = \arg \max_{c_j \in C} P(c_j) \prod_i^n P(x_i | c_j)$$

In reality, the assumption of conditional independence rarely holds, as features tend not to be completely independent of each other. However, with appropriate pro-processing, Naïve Bayes classifiers can work well for many problems, even for those where the independence assumption is not true (Russell & Norvig, 2010, p.499).

### 2.1.2 Logistic Regression

Another popular algorithm for classification is Logistic Regression (LR), sometimes also referred to as Maximum entropy modelling or MaxEnt for short. Similarly to the Naïve Bayes classifiers, logistic regression works by extracting features and combining them linearly in order to detect what class label is suitable (Russell & Norvig, 2010, p.725). The algorithm takes the log of each input feature and combines them by weighting and adding the features up.

The biggest difference to Naïve Bayes classifiers is the fact that logistic regression is discriminative. Naïve Bayes classifiers are generative classifiers, meaning that they are based on the conditional probability of the target value, given an observation. On the other hand, discriminative classifiers such as logistic regression, model the probability of a document belonging to a class directly without modelling the joint distribution.

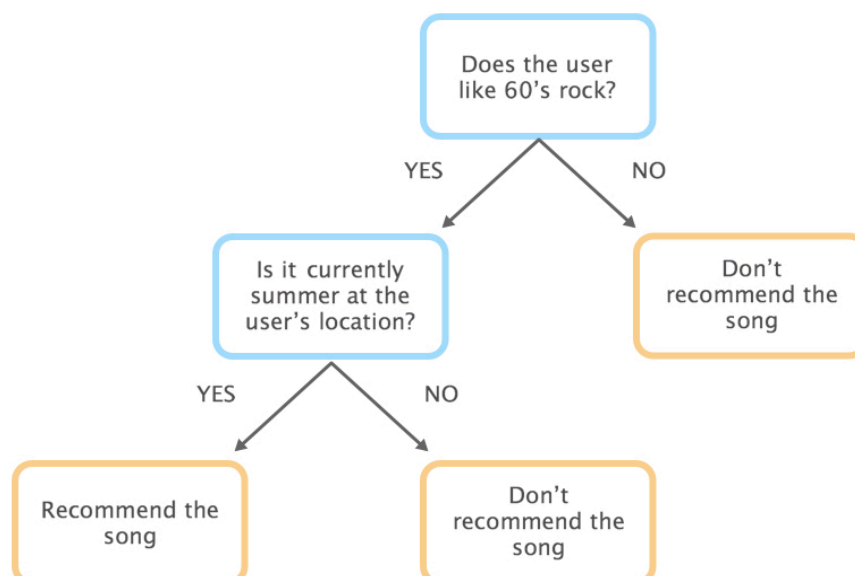
Given an input feature vector,  $\mathbf{x} = (x_1, \dots, x_n)$ , the algorithm defines a mapping to the class label:  $c_{LR} = f(\mathbf{x}; \mathbf{w})$ . For binary classification problems, the logistic regression algorithm can use Maximum Likelihood Estimation (MLE), an iterative approach used to find optimal values for the weights,  $\mathbf{w}$ , by adjusting them repeatedly until there is no additional improvement in the algorithm's ability to predict the class variable.

### 2.1.3 Decision Trees and Random Forests

Decision tree induction is a simple form of machine learning, mostly used for classification purposes. The decision tree learning algorithm aims to build a tree-structured model that can predict the class of an object, based on several input features. The produced model has a flowchart-like structure, where each internal node splits the data based on the value of a certain input feature, and leaf nodes represent the classes. Decision trees reach their decision on what class to predict by going through the chart, performing tests on features for each internal node, and returning the decision when they reach a leaf node (Russell & Norvig, 2010, p.697).

To illustrate, consider a music streaming system evaluating whether it should recommend the song «Summer in the City» by The Lovin' Spoonful to a specific user. Figure 2.1 illustrates a potential decision tree for this example.

**Figure 2.1:** Decision tree example.



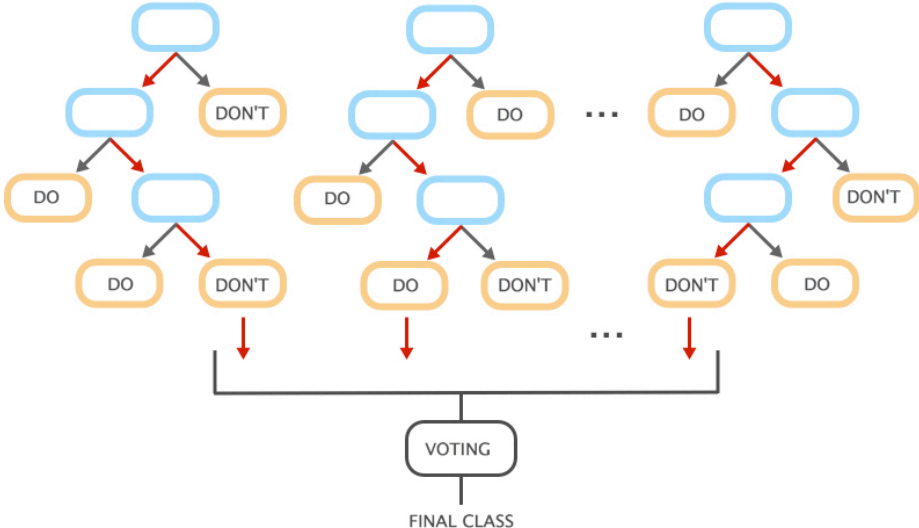
2. Background

In order to construct a decision tree model, the learning algorithm adopts a greedy divide-and-conquer strategy, choosing to test on the most important attributes first, i.e., the features that create the best split of data. To measure the importance of a feature, measures like Gini impurity or information gain (Entropy) are typically used (Raileanu & Soffel, 2004).

Decision trees are useful in situations where the data resources are limited and when it is desirable to have an easily interpretable model. Traditional decision trees are prone to overfit the training data, i.e., having low bias, but a high variance, as they easily pick up on irregular patterns in large feature sets.

In order to tackle the problem of overfitting, Random Forests (RF) are ensemble methods that combine multiple weaker models in order to construct a larger model with greater performance. Random Forests work by utilizing multiple decision trees, with controlled variance, trained on different parts of the training data, and letting each have a say in the classification of an instance. This way the method corrects for single decision trees' tendency to overfit the training data set. Figure 2.2 displays what a simplified random forest model could look like.

Figure 2.2: Example of a Random Forest Classifier



### 2.1.4 Support Vector Machines

Support Vector Machines (SVMs) are supervised machine learning algorithms widely used for binary classification purposes. The key idea behind SVMs is to plot labeled training data items in a finite-dimensional feature space and design an optimal hyperplane that differentiates the data points belonging to each class.

Each of the  $n$  data points in the set can be written as  $(\mathbf{x}_i, y_i)$  for  $i = \{1, \dots, n\}$ , where  $\mathbf{x}_i$  is a  $p$ -dimensional feature vector and  $y_i$  indicates what class the item belongs to ( $y_i = -1$  for the negative class,  $y_i = +1$  for the positive class). Each of the  $p$  dimensions corresponds to a feature of the training data. In text classification, this could for instance be the frequency of a specific term.

Support vector machines aim to find a  $p - 1$  dimensional hyperplane that divides the data points into the two different classes. An example can be found in figure 2.3, where the data points are represented with two-dimensional vectors, and the separating hyperplane is one-dimensional, thus a line.

In general terms, the hyperplane can be expressed as points satisfying:

$$\mathbf{w} \cdot \mathbf{x} - b = 0,$$

Where  $\mathbf{w}$  is a weight vector normal to the hyperplane and  $\frac{b}{\|\mathbf{w}\|}$  represents the offset from the origin.

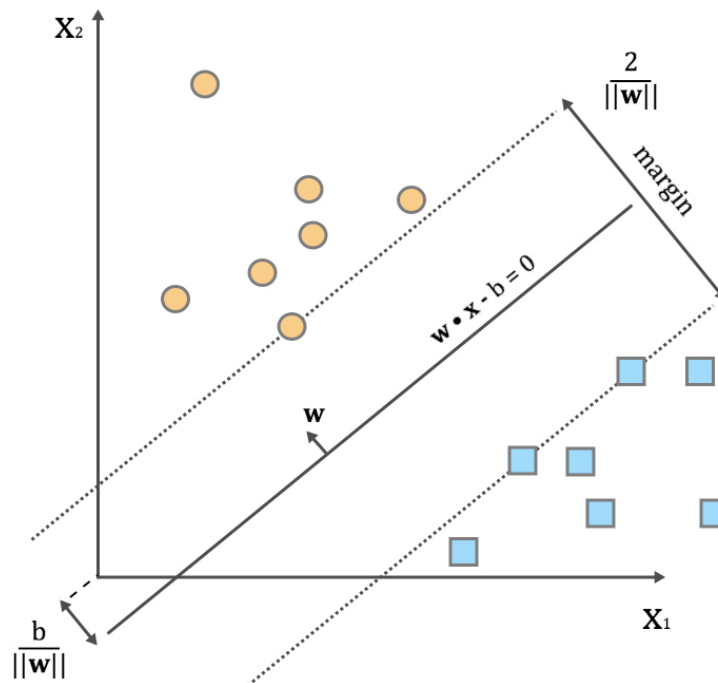
Assuming the training data is linearly separable, there will be an infinite number of possible hyperplanes that separate the points belonging to each class; however, not all of them will generalise well to new data. Intuitively, the optimal hyperplane is the one that has the largest distance to its nearest data points, as this will lower the probability of misclassifying unseen data points and hence reduce the generalisation error.

The distance of separation is called the margin (see figure 2.3), and the nearest training data points to the optimal hyperplane are referred to as support vectors. Unlike models like linear regression, where all points are taken into account, only those closest to the decision boundary, i.e., the support vectors, influence the location of the optimal hyperplane.

## 2. Background

The area between the two trenches, i.e., the support hyperplanes parallel to the optimal hyperplane, can be regarded as no-mans land, and no training data points should fall into this space. The algorithm has to make sure each training data point is located on the correct side of the margin. The SVM therefore seeks to maximise the size of the margin while honouring this constraint. What is left is a constrained, quadratic optimisation problem, solvable by the Lagrangian multiplier method and guaranteed to have a unique optimum.

**Figure 2.3:** Example of a Support Vector Machine with data points in a two-dimensional feature space.



In practice, many problems are not linearly separable. Support vector machines deal with this in two ways. Either by incorporating slack variables to allow for some mistakes when fitting the training data, or by using kernel functions.

For the slack variable approach, often called the soft margin classifier, there is a trade-off between the size of the margin and the misclassifications on the training data set. A loss function is introduced to penalise errors proportional to the distance from their location to the separating hyperplane. Besides this, the SVM will treat the problem as if it were

linearly separable. This approach is also applied to problems that are, in fact, linearly solvable, but where the «hard margin» approach might not yield the model with the lowest generalisation error.

The kernel approach involves projecting the data onto a higher dimensional space, with some transformation function,  $\varphi(\mathbf{x}_i)$ , in order to make it linearly separable. The optimisation problem described earlier will depend on the pairwise inner product of the feature vectors. With the kernel approach, this product is replaced by a kernel function. E.g., the linear kernel is simply the inner product of the transformed data points. Other kernel functions such as Polynomial, Gaussian radial basis function or Hyperbolic tangent are also applicable. By computing the kernel function, it is not necessary to compute the transformations of each feature vector,  $\varphi(\mathbf{x}_i)$ . This will often decrease the algorithm's computational costs and is referred to as the «kernel trick». Kernels can be defined over vectors, graphs, text, images and data sequences, making support vector machines useful in many situations.

### 2.1.5 Classification Scoring Metrics

In order to evaluate the performance of a classifier, different scoring metrics can be calculated based on the classification system's predictions. The most common metrics to use in evaluation of text classification are precision, recall and  $F_1$ -score. The calculation of these depends on the number of *true-positives* ( $tp$ ), *true-negatives* ( $tn$ ), *false-positives* ( $fp$ ) and *false-negatives* ( $fn$ ), where true positives and negatives are the number of correctly classified examples as respectively positive and negative, and false positives and negatives similarly refer to the numbers of falsely classified examples. Figure 2.4 illustrates the values' location in a confusion matrix.

**Figure 2.4:** Confusion matrix

		Predicted	
		Positive	Negative
True	Positive	$tp$	$fn$
	Negative	$fp$	$tn$

To illustrate, consider the music streaming system again, this time trying to figure out what songs to recommend for a playlist named «Sound of Summer». If the system returns

## 2. Background

the ill-suited «All I Want For Christmas Is You» as a candidate song, this would be counted as a false-positive. On the other hand, if the system also considered the song to be inappropriate for the playlist, it would be counted as a true-negative.

**Precision** is a measure of how relevant the predicted positives are, i.e., how many of the detected positives that truly are positive. It is defined as the number of true positives over the sum of false and correct positive classifications. **Recall**, on the other hand, is a measure of how many of the positive instances that were picked up by the classifier. Recall, sometimes also referred to as sensitivity, is given by the fraction of true positives over the total amount of truly positive instances in the data set, i.e., the sum of true positives and false negatives.

$$Precision = \frac{tp}{tp + fp} \qquad Recall = \frac{tp}{tp + fn}$$

In the example with the summer playlist, precision would measure how many of the recommended songs that are indeed suited for the playlist, while recall would measure how many of the well-suited summer songs that were recommended by the system. Both metrics should be examined under evaluation of a model's effectiveness. Unfortunately, the two metrics are often in tension as improving precision will lower the recall score and vice versa.

Precision and recall are often combined using the F-scores, representing the weighted means of the two metrics. The most popular F-score to use in evaluation of classification systems is the  $F_1$ -score, with  $\beta = 1$ , representing the harmonic mean of precision and recall.

$$F_\beta = (1 + \beta) \cdot \frac{precision \cdot recall}{\beta^2 precision + recall} \qquad F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$



## 2.2 Text Representations

The field of Natural Language Processing (NLP) is concerned with the interactions between computers and human (natural) language, including enabling computers' ability to process and derive information from texts written by humans. In order to utilise the machine learning methods presented in the last section for classification of textual data, it is necessary to represent such texts in an efficient way.

This section presents an introduction to relevant methods and concepts used in text representations for natural language processing.

### 2.2.1 *N*-grams

*N*-grams are sequences of  $n$  consecutive units in a text, e.g., words (word  $n$ -grams) or characters (character  $n$ -grams), and are popular in text mining and different NLP tasks. When  $n = 1$ , the items are referred to as unigrams, which corresponds to single words/characters. Similarly, bigram is used to describe a sequence of length 2, trigram for a sequence of length 3, and so on. To illustrate, figure 2.5 displays how the first line of «Summer Nights» could be turned into sets of word unigrams, bigrams and trigrams.

*N*-grams can be used to develop features for supervised machine learning models. They can also serve as useful tools in spelling correction, text summarisation and speech recognition.

**Figure 2.5:** Word unigrams, bigrams and trigrams from a song lyric

Unigram ( $n=1$ )	summer	lovin'	had	me	a	blast
Bigram ( $n=2$ )	summer lovin'	lovin' had	had me	me a	a blast	
Trigram ( $n=3$ )	summer lovin' had	lovin' had me	had me a	me a blast		

### 2.2.2 Bag of Words

The bag-of-words (BOW) model is a simple method to represent textual documents in natural language processing, in which each document is represented as a numerical feature vector, disregarding grammar and word ordering. The representation is often

## 2. Background

used in text classification where the frequency of each term is used as numerical feature to train classifiers.

The process of creating feature vectors for documents with the bag-of-words model consists of two main steps; first, creating a vocabulary of all unique terms in the collection of documents, and then, creating a feature vector per document and updating it by counting the occurrences of each term in the vocabulary.

As an example, image four documents,  $d_1$ ,  $d_2$ ,  $d_3$  and  $d_4$ , from a collection of lyric lines from summer hit singles. Table 2.1 shows the vocabulary, while  $\mathbf{x}_{d_i}$  represents the feature vector of document  $d_i$ .

$d_1 = \text{«back in the summer of sixty-nine»}$

$d_2 = \text{«after the boys of summer have gone»}$

$d_3 = \text{«in the summer, in the city»}$

$d_4 = \text{«boys, boys, boys, let the summertime roll»}$

$\mathbf{x}_{d_1} = [0\ 1\ 0\ 0\ 0\ 0\ 1\ 0\ 1\ 1\ 0\ 1\ 1\ 0\ 1]$

$\mathbf{x}_{d_2} = [1\ 0\ 1\ 0\ 1\ 1\ 0\ 0\ 0\ 1\ 0\ 0\ 1\ 0\ 1]$

$\mathbf{x}_{d_3} = [0\ 0\ 0\ 1\ 0\ 0\ 2\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 2]$

$\mathbf{x}_{d_4} = [0\ 0\ 3\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 1\ 0\ 0\ 1\ 1]$

**Table 2.1:** Example vocabulary

Key	Word		
0	after	7	let
1	back	8	nine
2	boys	9	of
3	city	10	roll
4	gone	11	sixty
5	have	12	summer
6	in	13	summertime
		14	the

### 2.2.3 Term Frequency - Inverse Document Frequency

Term Frequency - Inverse Document Frequency (tf-idf) is a numerical statistic used to measure the importance of terms in a document, calculated by comparing the number of occurrences in a single document to its usage in the larger document collection. Tf-idf is used as a weighting scheme in many of today's text classification systems.

Considering the example given in the preceding subsection, we see that the term «the» is present in all documents and hence might not carry as much information as words such as «city» or «boys». The idea behind tf-idf weighting is that instead of using the raw

frequencies of occurrence as weights, the impact of the frequently used tokens is scaled down. A common way to calculate tf-idf is to use the following formulas:

$$tf-idf(d, t) = tf(d, t) \times idf(D, t)$$

$$idf(D, t) = \log\left[\frac{n}{df(D, t)}\right]$$

Here,  $tf(d, t)$  is the term frequency, i.e., the number of occurrences the term has in the document. The term frequency count is usually normalised to prevent bias towards longer documents. The inverse-document frequency,  $idf(D, t)$  is the scaling factor calculated as the log of total number of documents,  $n$ , divided by the document frequency, i.e., the number of documents that contain the term.

Suppose we want to calculate the tf-idf for the term «boys» for document  $d_4$  in the previous example. From the bag-of-words feature vector we get that the term «boys» occurs 3 times in  $d_4$ . If we choose to normalise the term frequency ( $\hat{tf}$ ), the raw frequency ( $tf$ ) is divided by the maximum possible frequency, i.e., the number of words in the document. Then, using the formula above with base 10 logarithm, we get the following:

$$\begin{aligned} tf-idf(d_4, "boys") &= \hat{tf}(d_4, "boys") \times idf(D, "boys") \\ &= \frac{tf(d_4, "boys")}{\sum_{t' \in d_4} tf(d_4, t')} \times \log\left[\frac{n}{df(D, "boys")}\right] \\ &= \frac{3}{7} \times \log\left[\frac{4}{2}\right] = 0.13 \end{aligned}$$

If we were to similarly calculate the tf-idf weight for the term «the», the  $idf(D, "the")$  would return 0, since the term is present in all document in the collection. Hence, the tf-idf would be 0 as well. In order to not entirely ignore all terms that occur in all documents in a training set, some implementations, such as the one by Scikit-learn (section 2.4.2), choose to add «1» to all idf-scores (smoothing).

## 2. Background

### 2.2.4 Stop word removal

In natural language processing, stop words are words that are removed from a text before or after the processing of the data. Usually, stop words refer to the most frequently used words in a language, such as «the» or «is» in English. The idea is that these words are too common to have strong distinguishing power for the task at hand, and are therefore excluded from the vocabulary entirely. Sometimes these words even contribute to noise. Removing terms from the vocabulary reduces the number of posts the system has to store and could make the representation of text more efficient for the machine learning algorithms. However, removing stop words could in some cases remove too much meaningful information, thus weaken the system. Stop word lists should either be kept minimal or customised to the domain of the problem.

There is no universal list of stop words, but many tools, such as the NLTK libraries for Python (section 2.4.3) offer different stop word corpora. Considering the example with song lyrics from subsection 2.2.2, using the NLTK's list of stop words would result in removal of the words; «after», «have», «in», «nine», «of», «sixty», and «the», thus reducing the size of both the vocabulary and the feature vectors.

### 2.2.5 Part-of-Speech Tagging

Part-of-speech (POS) tagging is the process of marking each token of a sentence with its contextual part of speech category, such as verb, noun, adjective and adverb, depending on the token's definition and context. Figure 2.6 presents an example of a part-of-speech tagged sentence. POS-tagging can serve as a useful tool in characterisation of context and in determining authorship.

The tagging process could be performed by hand, but in the field of natural language processing it is usually performed automatically by trained algorithms. Many of the available POS-tagging systems are trained on news corpora with a formal and precise language, which could lead to difficulties with text from social media, which often happens to be short and contain misspellings, slang and abbreviations.

**Figure 2.6:** Part-of-Speech example

Input	«I'm walking on sunshine»				
Part-of-Speech tags	I	'm	walking	on	sunshine
	<b>PRP</b>	<b>VBP</b>	<b>VBG</b>	<b>IN</b>	<b>NN</b>
Explanations	Personal pronoun	Verb, non-3rd person singular present	Verb, gerund/present participle	Preposition	Noun, singular or mass

### 2.3 Manual Annotation and Inter-Annotator Agreement

In order to solve the problem of supervised classification, there is a need for a representative corpus of examples to train the model on. In cases where the available data is unlabelled with respect to the classification, the task of annotation has to be done either automatically or by hand.

A popular method to label data entities in a data set, although cost expensive, is to use manual annotation. Manual annotation is a methodology where a human annotator reviews a document at some level and adds meta-data to it. For classification tasks, the annotator is free to decide on a single class from multiple pre-defined classes per document. It is desirable that the annotation process is conducted by multiple annotators, to reduce mistakes and to make sure the annotations are not biased. Working with a manually annotated dataset, it is desirable to have a metric on how certain the annotations are.

Inter-annotator agreement measures how well the annotations from different annotators match. This measure can reveal the reliability of the annotations, and whether the class conditions and boundaries are clear enough. Several inter-annotator agreement measures have been proposed in literature. For nominal data, where the classes are considered to be unordered, two popular metrics are Cohen's kappa (Cohen, 1960) and Fleiss' kappa (Fleiss, 1971).

## 2. Background

### 2.3.1 Cohen's kappa

Cohen's kappa is a widely used metric for assessing agreement between two annotators. It takes the possibility that the annotators agree by chance into account, by considering the individual class distribution. Cohen's kappa is given by the following formula:

$$k = \frac{(P_o - P_e)}{(1 - P_e)}$$

Where,  $P_e$  is the probability of agreement by chance if the annotators were to answer randomly according to the observed distribution, and  $P_o$  is the observed agreement.

### 2.3.2 Fleiss' kappa

Fleiss' kappa is an extension of Cohen's kappa, suitable for more than two annotators. While Cohen's kappa assumes the annotators have evaluated the exact same set of documents, not every annotator needs to classify each document in order to calculate Fleiss kappa. However, each document must be evaluated by a fixed number of annotators.

For both Cohen's and Fleiss kappa, the score ranges from 0 to 1. A score of 0 indicates no agreement above what is expected by chance, and 1 represents perfect agreement. Negative values are possible, if the observed agreement is less than the expected chance agreement, and are usually treated as 0.

## 2.4 Tools

This section introduces some popular tools and resources available for use in collection of Twitter data, natural language processing and text classification. For the sake of brevity, it only presents an overview of the tools and refers the reader to the associated documentations for further information.

### 2.4.1 Twitter API<sup>1</sup>

The rapid growth of social media in the last decade has resulted in enormous amounts of data available to use in a large variety of studies. Twitter provides a REST API to access

---

<sup>1</sup> Twitter Developer Documentation:  
<https://developer.twitter.com/en/docs>

and download users' past tweets and profile information. Each user profile on Twitter has available data at both user- and tweet-level. Key data fields at both levels are displayed in figure 2.7.

**Figure 2.7:** Example Twitter user with data fields<sup>2</sup>



At user-level, the only string data field that is guaranteed existence is the username. The username is required to be unique and has a length of maximum 15 characters. Display name and biography are optional, and can each have a maximum length of respectively 50 and 160 characters. Twitter also allows users to define a location or a homepage for the account. Other available data includes the date the account was created, number of followers, number of tweets, and numbers of accounts the profile is following.

At tweet-level the most obvious piece of data is the tweet text itself. The tweet field holds the textual content and often includes hashtags, mentions of other users, URLs, etc. Each tweet has an associated timestamp, representing the Coordinated Universal Time of the tweet's creation. Tweets also have numerical features available, such as number of retweets and likes.

<sup>2</sup> Twitter profile presented in compliance with Twitter's terms of fair use, including their display requirements:

<https://help.twitter.com/en/rules-and-policies/fair-use-policy>  
<https://developer.twitter.com/en/developer-terms/display-requirements.html>

## 2. Background

### 2.4.2 Scikit-Learn<sup>3</sup>

Scikit-learn (sklearn) (Pedregosa *et al.*, 2011) is an open source machine learning library for the programming language Python<sup>4</sup>. It provides easy access to implementations of various classification, regression and clustering algorithms, including all methods presented in section 2.1. The Scikit-learn API is uniform and streamlined, making it easy to switch between and explore different machine learning algorithms. As well as providing implemented methods and customisation tools, it offers an extensive online documentation.

Scikit-learn is licensed under a BSD licence<sup>5</sup>, encouraging both academic and commercial usage of the library.

### 2.4.3 Natural Language Toolkit<sup>6</sup>

The Natural Language Toolkit (NLTK), is a collection of libraries for natural language processing for Python. NLTK includes more than 50 corpora and resources, along with libraries for language detection, tokenisation, POS-tagging, stop word removal, etc.

NLTK is licensed under the Apache 2.0 licence<sup>7</sup>, a permissive license which allows for modifications, distribution and commercial usage.

---

<sup>3</sup> Scikit-learn Documentation:  
<http://scikit-learn.org/stable/documentation.html>

<sup>4</sup> Python Documentation:  
<https://www.python.org/doc/>

<sup>5</sup> Scikit-Learn License  
<https://github.com/scikit-learn/scikit-learn/blob/master/COPYING>

<sup>6</sup> Natural Language Toolkit Documentation:  
<https://www.nltk.org/>

<sup>7</sup> Apache 2.0 license:  
<https://www.apache.org/licenses/LICENSE-2.0>



## 3. Related Work

This chapter presents previous research considered relevant to the objective of this work. The first section presents a review of some studies related to pro-ED communities and their use of language, while section 3.2 presents studies on automatic classification of Twitter users based on mental health and subcultures.

The final section of this chapter describes in more detail how the relevant studies, presented in the two preceding sections, approached the different steps of constructing a classifier including data collection, pre-processing, feature extraction and model building.

### 3.1 Studies on Pro-ED

Pro-eating disorder websites gained publicity and growth in the early 2000s, and with the advent of social media, pro-ED communities are now to be found easily on such platforms. The body of research related to pro-ED communities is large and varied, including analyses of themes (Borzekowski *et al.*, 2010) and imagery (Boepple *et al.*, 2015; Borzekowski *et al.*, 2010; Ghaznavi & Laramie, 2015; Talbot *et al.*, 2013), and research on community members' experiences regarding motivation (Yeshua-Katz & Martins, 2017), awareness (Wilson *et al.*, 2006) and identity (Bates, 2015; Giles, 2006). Studies have also been conducted to evaluate the influence this type of content has on community members (Ransom *et al.*, 2010; Yeshua-Katz & Martins, 2017) and the general public (Bardone-Cone & Cass, 2007; Talbot, 2010). Despite this, the availability of related work considering linguistic characteristics and traits of online pro-ED content and its authors, is limited.

Arseniev-Koehler *et al.* (2016) explored pro-anorexia socialization on Twitter. The purpose of their study was to investigate pro-ana users' textual references to eating disorders in tweets, and the online social connections between the community members. They collected a data set consisting of 45 Twitter users considered to take part in the pro-

### 3. Related Work

ana community, along with their tweets and a selection of 100 random followers. Further, Arseniev-Koehler *et al.* developed a codebook to examine the users' tweets and profile information for references to eating disorders. The codebook consisted of keywords from a clinical screening questionnaire, and from literature on eating disorders and pro-ED communities. Example words from their ED reference codebook are shown in table 3.1. For explanations of hashtags and terms related to pro-ED communities, see appendix A.

Arseniev-Koehler *et al.* (2016) found that a mean of 35.7% of the pro-ana profiles' tweets contained at least one word from the codebook, and 86.7% of the pro-ana users displayed an ED reference in either username, display name or biography.

**Table 3.1:** Arseniev-Koehler *et al.*'s (2016) ED Reference Codebook

<b>Eating Disorder</b>	anorexic, ana, bulimic, proana, proed, promia, ednos, eating disorder, wannarexic
<b>Body Image &amp; Weight</b>	overweight, obese, fatty, skinnier, skeleton, emaciated, hipbones, backbone, bones, collarbone, thighgap, bikinibridge, hipbones, thighs, hips, thinspo, bonespo, perfection, weight, scale, gw, cw, lw, ugw, bmi, pound
<b>Food and Meals</b>	calorie, food, breakfast, dinner, meals, eating, eat, ate, appetite, starve, hunger, diet, skip, fasted, calorieapril, projectthin, rg, abcdiet, binge, bloated
<b>Compensatory behaviour</b>	laxies, laxatives, vomited, throwup, puke, purge, workout, abs, jog, elliptical, exercise, miles, gym, treadmill

Studies have also been carried out on image descriptions on Instagram, in order to study the lexical and orthographic variations of pro-ED terms, adopted to differentiate the communities from outsiders or to circumvent the platform's restrictions. Examples include the usage of hashtags like #thygap instead of #thighgap, and #anorexiaa for #anorexia. Chancellor *et al.* (2016) found that the pro-ED communities on Instagram used orthographic variation more frequently over time, and that the variations kept diverging from the original term.

Stewart *et al.* (2017) also compared the use of lexical variations found on Instagram to a sample set of 4043 pro-ED tweets from Twitter. Their results showed that only 15.0% of the pro-ED tweets contained a variant hashtag, compared to 51.9% of the image

descriptions on Instagram. The variations used in tweets were also closer to their original form. This demonstrated that the practice of using lexical variations is more common on Instagram, and likely a result of Instagram’s restriction policy.

Yom-Tov *et al.* (2012) compared pro-anorexia users’ activities on the photo sharing site Flickr<sup>1</sup> to those performed by users with a recovery-oriented perspective on eating disorders. They extracted photos from 491 users on the site, and compared the interactions and characteristics of pro-anorexia and pro-recovery users. Interestingly, they found that pro-recovery users employed similar words to the pro-anorexia users when describing their photos. Yom-Tov *et al.* suspected that the explanation for this was that the pro-recovery users wanted to expose their content to the pro-ana users.

Yom-Tov *et al.* (2012) also computed the most indicative tags for each class by calculating the ratio between the probabilities of users from both classes (pro-ED vs. pro-recovery) utilizing each tag word. The tags with the highest probability of usage by the pro-anorexia users, compared to those with a recovery-oriented perspective, and vice versa, are presented in table 3.2 The researchers highlighted the fact that the recovery-oriented users had a more varied tag set, compared to the pro-ana users whose top tags seemed to refer mostly to body images.

**Table 3.2:** Top 10 most indicative image tags on Flickr according to Yom-Tov *et al.*(2012)

<b>Pro-anorexia</b>	doll, thinspo, skinny, thin, cigarette, sexy, landscape, legs, abstract, long
<b>Pro-recovery</b>	home, sign, selfportrait, glass, cars, plants, building, mother, sunshine, bird

### 3.2 Twitter User Classification

Twitter user profile classification aims to identify to which of a set of pre-defined class labels a Twitter user belongs. Previous examples of Twitter profile classification include identification of ethnicity, gender, occupation and political orientation (Pennacchiotti & Popescu, 2011; Rao *et al.*, 2010).

<sup>1</sup> <https://www.flickr.com/about>

### 3. Related Work

With the exception of Yom-Tov *et al.*'s (2012) work, few previous studies have explored the differences between members of pro-ED communities and users considered to either have a recovery-oriented perspective on eating disorders or considered unrelated to the matter at hand. To the best of the author's knowledge, there are no previous studies on classification of pro-ED users on Twitter or similar social networking sites. This section presents some previous research on Twitter user classification that presumably share similarities to the objective of identifying pro-ED profiles.

#### 3.2.1 Mental Health

Based on the obvious connection between the pro-ED communities and eating disorders, it was considered relevant to review previous work done in the field of identifying symptoms and cases of mental illnesses from textual data sources. There has been a relatively large amount of studies on the usage of texts from medical records or clinical settings to detect mental illnesses (Jackson *et al.*, 2017; Tran & Kavuluru, 2017), including eating disorders (Bellows *et al.*, 2014). However, the task of utilizing user generated texts from social media in relation to mental health has not been subject for equivalent amount of research and shared tasks, likely due to the privacy concerns and the searchability of the content.

Coppersmith *et al.*'s (2015) shared task for the Computational Linguistic and Clinical Psychology (CLPsych) conference from 2015, aimed to detect users diagnosed with depression or post-traumatic stress disorder (PTSD) on Twitter. Due to ethical concerns, the organisers anonymised the entire data set and required all participants to sign a privacy agreement.

The CLPsych 2015's training data set contained 1145 user profiles, labeled as either depressed, PTSD or unrelated (control set). The shared task focused on the three binary classification problems of identifying depression users versus the unrelated users, PTSD vs. unrelated users, and depression vs. PTSD users.

CLPsych 2015 had several participants, including the University of Maryland (Resnik *et al.*, 2015), the University of Pennsylvania (Preotiuc-Pietro *et al.*, 2015), the University of Minnesota Duluth (Pedersen, 2015), and a small team («MIQ») composed of Microsoft, IHMC, and Qntfy (Coppersmith *et al.*, 2015). On all three binary classification problems, the University of Maryland's approach, using a support vector machine trained on lexical

features discovered through topic modelling, outperformed the competitors on average precision.

Besides this shared task, studies have been conducted to discover evidence of depression in single tweets with varying results (Burnap, Colombo & Scourfield, 2015; Homan *et al.*, 2014; Yazdavar *et al.*, 2017). Mowery *et al.* (2016 & 2017) aimed to classify depression-related tweets into different sub-classes, such as «depressed mood», «disturbed sleep», etc. They compared the performance of different machine learning approaches, and later conducted a feature study exploring the performance of support vector machines trained on different types of features from the tweets. Their best performance was achieved using an SVM trained on unigram features.

#### 3.2.2 Subcultures

Although there is a strong connection between eating disorders and pro-ED communities, it is important to stress the fact that identification of users taking part in a pro-ED community is not the same as identification of users suffering from an eating disorder. The classification problem's core is to detect users taking part in an online subculture. With this in mind, Balasuriya *et al.*'s (2016) study considering identification of American street gang members on Twitter was included as a relevant study. Gang members and pro-ED users may not have many common characteristics, but the two classification problems do. Both are aiming to detect users whose tweets often surround specific topics (drugs, money, weapons vs. body shape and food) and include words, abbreviations and phrases, whose usage changes with time and whose meanings are unfamiliar to the general public. The two communities both represent a very small population of the hundreds of millions of users on Twitter.

Balasuriya *et al.* (2016) extracted 400 gang members', and 2865 non-gang members', user profiles and tweets. The researchers explored using different machine learning models, trained on a large variety of different features. Their best proposed classifier, a random forest model trained a combination of features, achieved an  $F_1$ -score of 0.78 over the gang-member profiles.

### 3.3 Relevant Approaches to Building a Classifier

The relevant studies presented above yield insight to how one can proceed to achieve the intended goal of this study. As with most supervised classification systems, the procedure can be broken down into four main steps; data collection, pre-processing, feature extraction and model building.

#### 3.3.1 Data Collection

The task of data collection refers to the process of retrieving and annotating a relevant data set for training and testing the classifier. Before starting the actual process of extracting data from Twitter, one needs to establish what users should be considered pro-ED profiles. Yom-Tov *et al.* (2012) defined a pro-anorexia user as «*one who is actively involved in the creation and dissemination of content that takes a positive and encouraging attitude towards eating disorders*».

In an attempt to concretise such notions, Arseniev-Koehler *et al.* (2016) presented inclusion criteria to evaluate whether a user profile was considered «pro-ED». Profiles that had tweeted the hashtag #proana at least once, and displayed a positive pro-ED attitude in profile information or recent tweets, were identified as pro-ED. A *positive pro-ED attitude* was defined as:

- (1) self-identifying as pro-ED and/or having an eating disorder alongside being anti-recovery, or
- (2) expressing a desire for emaciation, or
- (3) ascribing to a pro-ED event, such as participating in collective diets or competitions to restrict calorie intake.

For the actual retrieving process, both studies first used related keywords, such as «proana», «thinspo» and «thinspiration» in searches, and then manually labeled the users appearing in the results. None of the pro-ED related studies included a set of unrelated control users. The CLPsych shared task (Coppersmith *et al.*, 2015) estimated the age and gender of the users in their data set based on their language usage, and utilised this information to construct a demographically matched control set. In Balasuriya *et al.*'s (2016) study of gang members, they included 2000 random, American, profiles to work as a control set against the 400 gang users. Additionally, in order to capture local

language of friends and family of gang members, and ordinary people living in areas where the gangs operate, they introduced 865 non-gang users who utilised similar hashtags.

For the CLPsych 15 shared task (Coppersmith *et al.*, 2015), most information at user-level was removed during the anonymisation, keeping only the number of followers, followed accounts, favourites and the time the account was created. Balasuriya *et al.*'s (2016) study, on the other hand, found usage of many of the available data fields, using the location field in pre-processing, and extracting features from biographies, avatars and tweet text.

#### 3.3.2 Pre-Processing

Pre-processing consists of various steps to prepare the collected data for further analysis and feature extraction. Usually this phase includes filtering out unwanted data entities and removing parts that do not contribute useful information. In the CLPsych 15 shared task's data set (Coppersmith *et al.*, 2015), users with less than 25 tweets were removed in order to make sure each user had a sufficient amount of data. They also used a language detector to ensure that 75% of all included user tweets were in English. Arseniev-Koehler *et al.* (2016) similarly removed profiles tweeting in languages other than English during their manual review of the users. In Balasuriya *et al.*'s (2016) gang-member study, they excluded all Twitter profiles whose location was unspecified or outside the United States. Although, an efficient method to exclude foreign profiles, this would usually cause a large reduction of data, as many users choose to leave the location field empty or fill in locations that are either imprecise, non-physical or clearly wrong, such as «in your dreams» or «Antarctica».

In data sets collected from Twitter, another common trait is the presence of mentions, hashtags and URLs (Uniform Resource Locator). How to handle such Twitter/Internet specific terms differs, and some studies have chosen to remove all of them reasoning that they do not contribute with differentiating information for the problem at hand. The most popular solution is to replace such entities with placeholder tags, e.g., replacing «@PunlimitedPuns» with «@mention», and «<https://imgur.com/a/8PKUI>» with «URL». This method captures the presence of links and mentions, without having to deal with the large collection of unique entities. Balasuriya *et al.* (2016) on the other hand, used URLs

### 3. Related Work

to detect links to gangster rap videos on YouTube, as this happens to be a characteristic of the online culture of street gang members. Emojis and emoticons are often either removed or converted to an equivalent unigram, e.g., the Unicode name of the emoji or emoticon.

Other pre-processing methods are to remove non-alphanumeric symbols and stopwords, as they are often too common to bear valuable information for the classification. Some studies have also performed stemming (Balasuriya *et al.* 2016) or lemmatisation (Resnik *et al.* 2015) on tweets and biographies in order to reduce the number of features while keeping the main essence of the words.

#### 3.3.3 Feature Extraction

Section 2.2 described different features often used in the field of text processing in general. However, not all features seen in NLP literature are suitable for Twitter applications, due to the informal and short nature of tweets. This sub-section presents a brief overview of the representations and features that were used in the studies presented above.

As described in 2.4.1, each tweet, and associated author, contains multiple data fields with information such as the tweet text, time stamps, popularity measures and a collection of user related data. Most previous work within classification on Twitter tends to primarily focus on the textual content of tweets. Since each user typically has multiple tweets, one need to decide on how to represent this in the classification system. A possible method is to treat each tweet as a single document and duplicate the user-level data over all tweets per user profile. On the extreme opposite, Balasuriya *et al.* (2016) and Preoțiu-Pietro *et al.* (2015) decided to aggregate all tweets per user together, irrespective of their timestamp, before extracting features. Besides being simple, this method makes sure each user corresponds to one document in the training set. Resnik *et al.* (2015); however, decided on a middle ground of aggregating tweets by the week they were created, such that all tweets from a given week corresponded to one document in the set. Thus each user was represented as many times as weeks they had been tweeting. Their study did, however, not include features at user-level, so they did not have to duplicate this data.



After deciding on what is considered a document, next step is to extract features. From the textual content present in tweets, word n-grams tend to be a popular choice of feature in Twitter classification applications. Balasuriya *et al.* (2016) chose to represent all text from both tweets and biographies through counts of unigrams. In Preoțiu-Pietro *et al.*'s (2015) submission to the 2015 CLPsych shared task, they included unigram features, as well as features from 6 different word clusters and topic models, with the unigram features yielding the best results when used separately. The MIQ submission to the shared task (Coppersmith *et al.*, 2015) went for the more unusual choice of representing tweets as character n-grams of lengths 5 and 6, with 5-grams achieving the best performance.

Some previous work on user classification on Twitter have also included use of emojis in their set of features. Using emojis as features have previously shown promising results on tasks such as sentiment analysis on Twitter (Novak *et al.*, 2015). Balasuriya *et al.* (2016) trained models on features consisting of the frequency of each emoji used across all tweets per user. Their model, trained solely on emoji features, achieved a precision of 0.73 and 0.93 for the gang-members and non-gang members class.

As well as deriving unigram features from text and emojis, Balasuriya *et al.* (2016) also included avatars and YouTube links as input features for their gang-membership classifier. The avatar's features were unigram tags generated by a third party web service using deep learning. However, the image tag features on their own yielded relatively poor results. The YouTube links were converted to unigram features retrieved from video description and top comments on YouTube. Balasuriya *et al.* (2016) reported promising results using YouTube-link features for their classification model.

In Mowery *et al.*'s (2017) feature study for classification of depressive symptoms in single tweets, they experimented with a variety of different features, such as lexical (unigrams), Part-of-speech-tags, emoticons, demographic indicators, sentiment terms, personality traits and LIWC<sup>2</sup> terms for their SVM classifier model. They found that for identification of tweets labeled as «no evidence of depression», «evidence of depression» and «depressive symptoms», removing all non-lexical features resulted in equal or higher  $F_1$ -scores, indicating that most of the features did not contribute any valuable information.

---

<sup>2</sup> Linguistic Inquiry and Word Count:  
<http://liwc.wpengine.com/>

### 3. Related Work

However, for detection of subclasses of the «depressive symptoms» class such as «disturbed sleep» and «fatigue», inclusion of emoticons, sentiment and demographics yielded a positive contribution.

#### 3.3.4 Models

In the relevant studies presented in this chapter, the most popular choice of machine learning model was the Support Vector Machine with either a linear or radial basis function (RBF) kernel. However, some studies chose other approaches, or explored different models in order to find the best suited for their application.

Mowery *et al.* (2016) trained and tested a collection of different supervised machine learning classifiers for predicting which depression related class tweets belonged to, and reported average precision and recall, as well as  $F_1$ -scores for each model. The set of models they explored included decision trees, random forests, logistic regression, SVMs, Linear perceptron and Naïve Bayes. For most of the classification problems, the support vector machines were able to produce the highest  $F_1$ -scores. The researchers hypothesised that the reason behind this was the SVMs' ability to tolerate a large number of features. In a continuation of this work, studying features for classification of depressive symptoms (Mowery *et al.*, 2017), they chose to base all experiments on support vector machines.

In the CLPsych (Coppersmith *et al.*, 2015) shared task submission by Preoțiuc-Pietro *et al.* (2015), the researchers explored using two regularised linear classifiers; logistic regression and linear SVMs. For all of the three binary classification problems, both models yielded good and similar results. Balasuriya *et al.* (2016) explored using four different approaches; a Naïve Bayes model, a Logistic Regression model, a Random Forest and a Support Vector Machine. All models were chosen because of their reputation of performing well on textual data. The best reported  $F_1$ -score was achieved using the Random Forest classifier, although all, except the Naïve Bayes model, scored reasonable well.

## 4. Data

With the intended goal of using supervised machine learning to identify pro-ED users on Twitter, it was essential to have a fair amount of representative training data. For the purpose of this study, 7096 users were extracted from Twitter and manually annotated.

This chapter presents the details of the data set used in this work. The first section presents the criteria used to establish which users that were regarded as pro-ED community members in this thesis, while the following section discusses the choice of dividing the non pro-ED users into two classes, *pro-recovery* and *unrelated*. Section 4.3 describes the procedure of identifying and collecting user data, and section 4.4 presents an overview of the distribution of the collected data set. Further, section 4.5 presents some high-level filtering and pre-processing techniques that were applied. A discussion on the reliability of the annotations are to be found in section 4.6. Finally, section 4.7 presents analyses of the found data set considering usage of Twitter specific terms and emojis, tweet lengths, part-of-speech, locations and references to eating disorders in tweets, biographies and names.

All data collected for the purpose of this work was publicly accessible on Twitter and made available through the official Twitter API. The complete set of Twitter users, and associated data, was collected between October of 2017 and January of 2018<sup>1</sup>.

### 4.1 Characterising «Pro-Eating Disorder» Users

Pro-eating disorder users are considered to be users that take part in a pro-ED community online. As previously mentioned, pro-ED communities usually post content that glorifies disordered eating habits and idolises thinness. Users seek the communities for tips and inspiration, but also to find support through a community of likeminded

---

<sup>1</sup> For more technical information about the data collection, see appendix B.

#### 4. Data

users. Establishing that all pro-ED communities share one coherent outlook would be an oversimplification. Nevertheless, in order to build a corpus of Twitter users for the purpose of this thesis, it was essential to have unambiguous criteria to evaluate whether a user was participating in such a community or not.

For this purpose, it was established that pro-ED users were user profiles that at least once, either in their tweets, retweets or profile information<sup>2</sup>, displayed a positive pro-ED attitude.

This work's definition of a *positive pro-ED attitude* was based on the set of inclusion criteria presented in Arseniev-Koehler *et al's* (2016) paper, with an additional forth eligibility criterion added. Satisfying *one* of the criteria, was considered a display of a positive pro-ED attitude, thus causing the user to be labeled as pro-ED in the data set.

Tweets, retweets or profile information were considered displays of positive pro-ED attitude, if they:

- (1) Included a self-identification as pro-ED, or
- (2) Expressed a desire for emaciation, or
- (3) Ascribed a pro-ED event, or
- (4) Encouraged extreme weight control methods.

##### **(1) Self-identifying:**

The first inclusion criterion was satisfied if the user either stated being pro-eating disorder, or stated having an eating disorder alongside being anti-recovery. The criterion did also include the usage of explicit pro-ED hashtags<sup>3</sup> in a supporting manner.

Table 4.1 displays examples of tweets satisfying the different inclusion criteria and tweets that are not considered to be pro-ED. Example no. 3 used the hashtag #proana in an approving manner, and thus was considered a display of a positive pro-ED attitude. On the other hand, example No. 12 displays a negative charged statement towards the hashtag #proana, and did not satisfy the inclusion criterion.

---

<sup>2</sup> Profile information refers to a user's biography, username and display name.

<sup>3</sup> Pro-ED hashtags were here defined to include: #proana, #promia, #proanorexia, #probulimia, #proed

As stated in previous chapters, the members of pro-ED communities display different views on what the communities involve. It seemed that some of the users on Twitter, sharing content such as extreme diets, thinspiration, etc., did not identify as being part of a specific pro-ED community. In this work, multiple users declaring, for instance, to be «not proana» in their profile biographies or tweets, were still categorised as pro-eating disorder if their tweets or profile information satisfied one of the remaining three criteria.

**(2) Expressing a desire for emaciation:**

This criterion included all tweets or biographies that stated an aspiration to become extremely thin. This did also include sharing of thinspiration or bonespiration content. Other examples could be textual references to desires of having protruding bones, experiencing symptoms as a consequence of emaciation, or romanticising hospitalisation.

**(3) Ascribing to a Pro-ED event:**

Pro-eating disorder events are happenings where pro-ED community members take part in a collective competition, diet or fast. This often involves tweeting one's calorie intake during the day. The goal of such events is usually to minimise daily calorie intake or to stay below a given threshold that is decided by the organisers.

An example of a seasonal event is #Skinny4Xmas (see example No. 4 in table 4.1), drifted by an anonymous user profile on Twitter. What distinguish these tweets from other fitness and diet related posts are their unhealthy nature. For some weeks, #Skinny4Xmas encouraged a daily intake of 400 calories, i.e., less than 20% of the energy requirement for an average teenage girl (National Health Service, 2015).

This third inclusion criterion was satisfied if a tweet or biography stated taking part in such events, or used an event's hashtag, e.g., #Skinny4Xmas, in a supportive manner.

**(4) Encouraging extreme weight control methods:**

Sharing or encouraging such pro-ED competitions, fasts and extreme diets also satisfied the last criterion of encouraging extreme weight control methods. This final criterion did also include support of abnormal food restrictions or compensatory behaviour such as vomiting, misuse of laxatives or overtraining.

## 4.2 Pro-Recovery & Unrelated users

Some Twitter users dedicate much of their profile either to document a recovery process from an eating disorder, to criticise pro-ED communities, or to participate in anti-eating disorder campaigns. Previous research found that users, outside the pro-ED communities, with a recovery-oriented perspective on eating disorders, employed similar wordings to those used by pro-ED members when describing photographs on Flickr (Yom-Tov *et al.*, 2012).

Based on a hypothesis that users with a recovery-oriented perspective on eating disorders would be challenging to tell apart from the pro-eating disorder users, the users that did not satisfy any of the above criteria, and thereby considered being «non pro-eating disorder» were split in two classes; *Pro-recovery* and *Unrelated*.

Users that discussed eating disorders, or related topics, such as body dissatisfaction, with a recovery or health-related focus, were considered as «pro-recovery» in the data set. Pro-recovery users included doctors, activists, people who either had recovered or were ongoing recovery, their dependents, and organisations working for awareness or prevention of eating disorders. Remaining users were labeled as «unrelated».

Although the main focus of this work was to identify pro-eating disorder users, the idea was that the division between pro-recovery and unrelated users might contribute with interesting information about differences and similarities between these as well.

**Table 4.1:** Example tweets and associated categories

No.	Examples:	Inclusion criteria satisfied:	Category:
1	«Fasting is not a big deal. You shouldn't be proud of yourself for fasting that long, keep going. You didn't lose enough»	(4)	Pro-ED
2	«Pretty girls don't eat»	(4)	Pro-ED
3	«Bones are not scary, fat is scary #proana #skinny #bonespo»	(1) , (2)	Pro-ED
4	«Total intake: 450 calories, Exercise: -210 calories, Net intake: 240 calories #Skinny4Xmas»	(3)	Pro-ED
5	«Nothing will make you happier than feel your bones under your skin #askanamia #thinspiration#proana»	(1) , (2)	Pro-ED
6	«maybe if you'd stop shoveling shit into your mouth you'd be pretty and thin»	(4)	Pro-ED
7	«You don't need food. You need water, green tea and a flat stomach #proana #meanspo»	(1) , (4)	Pro-ED
8	«If you don't want your legs touching when you sit on a chair, stop eating #proana»	(1) , (2) , (4)	Pro-ED
10	Food should make you feel nourished, energised and thankful. #eatingdisorders #ed #edrecovery #prorecovery #positivevibe	-	Pro-Recovery
11	Today I ate two biscuits at work and didn't die. Teeny steps, right? #vegan #edrecovery #prorecovery	-	Pro-Recovery
12	Yo @twitter please ban the use of the hashtag #proana. It's promoting dangerous things to impressionable people :(	-	Pro-Recovery
13	Are you supporting someone with an eating disorder? We have an online chat this evening taking place on our website #support #eatingdisorders	-	Pro-Recovery
14	«Whoever invents the equivalent opposite of a microwave is gonna be super rich»	-	Unrelated

**Note:** Some tweets have been modified.

### 4.3 Data Collection Procedure

In order to identify pro-ED users, well-known pro-ED hashtags and terms, such as #proana, were used in searches on Twitter. Table 4.2 lists some of the keywords that were used during this step<sup>4</sup>. While going through the search results, users with profile information that displayed a positive pro-ED attitude were scraped and labeled as pro-ED. For each tweet that satisfied at least one of the criteria above, the author, and any users who had retweeted it, were also downloaded and labeled as pro-ED. Some of the labeled profiles were also used to find more pro-ED tweets and users.

Similarly, pro-recovery profiles were found by searching for hashtags and terms in relation to eating disorders and recovery, such as #EDrecovery. For the unrelated class, searches were conducted on a wide variety of hashtags and phrases unrelated to eating disorders. Some of the keywords were, however, assumed to be semantically close, such as #nutrition and #fitness.

**Table 4.2:** Examples of keywords used in searches

Domain:	Keywords:
Pro-ED	#Proana , #Promia, #Thinspiration, #Thinspo, #Meanspo, #Bonespiration, #Bonespo
Pro-Recovery	Eating Disorder Recovery, #EDRecovery, #RecoveryWarriors, #BeatED, #EatingDisorderRecovery
Unrelated	Stand up to Cancer, #Coffee, #Wedding, #NFL, Museum of Modern Art, #Hadoop,#Audi, #Snapchat, #Fitness, Today, #Funny, #Arsenal, Jimmy Fallon, #Eid, #FreePalestine, #God, #Weed, Business Insider, #PyeongChang2018, #DIY, #Brexit, #Christmas, #Life, #Vlog, #nutrition, The Breakfast Club, Legend of Zelda, #Fashion, #Coke, #Eurovision

For each user collected, the data scraped from Twitter consisted of the user's username, display name, biography, the date the account was created, location and up to 3200<sup>5</sup> of the user's most recent tweets and retweets, along with the publishing time and date for each

<sup>4</sup> For the complete list of keywords, see appendix B.

<sup>5</sup> The Twitter API only returns up to 3,200 of a user's most recent tweets.

[https://developer.twitter.com/en/docs/tweets/timelines/api-reference/get-statuses-user\\_timeline](https://developer.twitter.com/en/docs/tweets/timelines/api-reference/get-statuses-user_timeline)



tweet. Non-textual media elements, such as images, videos or GIFs<sup>6</sup>, were not included in the collected data set. No user profile was scraped more than once and none of the classes had overlapping users.

During this process, before a user was scraped, the user's top most recent tweets were read. If the user seemed to tweet in other languages than English, the user was ignored and not downloaded. Users whose accounts were protected, or had less than 30 public tweets/retweets, were also not included in the data set. For users considered to be unrelated or pro-recovery, as only their top tweets were read, there was no guarantee that they had not posted any tweets with a positive pro-ED attitude in the past. This potential issue is discussed further in section 4.6.

#### 4.4 Data Set

The resulting data set contained 7096 users and 10.7 million tweets. Table 4.3 and figure 4.1 present the distribution of users and tweets in the total raw data set.

**Table 4.3:** Raw Data Set

<b>Label:</b>	<b>#Users:</b>	<b>#Tweets:</b>
Pro-Eating Disorder	2358	2 622 033
Pro-Recovery	802	1 304 887
Unrelated	3936	6 818 140
<b>Total</b>	<b>7096</b>	<b>10 745 060</b>

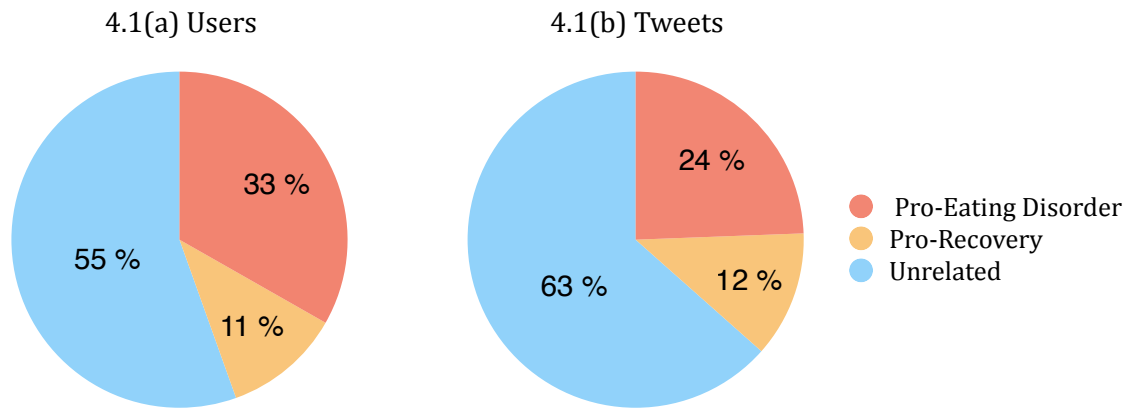
The choice of distribution between the classes in the data set was motivated by the need for a relevant data set, and the accessibility of unique users. In order to represent, to some extent, the real situation on Twitter, the majority of the users was representing the unrelated class. However, if the data set were to accurately reflect a realistic distribution, it would be extremely imbalanced, i.e., the data set would almost completely consist of

<sup>6</sup> Graphics Interchange Format  
<https://www.w3.org/Graphics/GIF/spec-gif87.txt>

#### 4. Data

unrelated users. Thus, this work opted for a middle ground of letting each class have a representation of a certain size, while still reflecting the dominance of the unrelated class.

**Figure 4.1:** Distribution of total users and tweets in the data set.



This choice of distribution does affect the interpretation of the classification results, as the number of false alarms, i.e., pro-ED false-positives, would be expected to be much larger in the actual population on Twitter.

Before any further analyses were conducted, 20% of the users, along with profile information and associated tweets, were randomly selected and put aside for testing purposes. The resulting training data set then consisted of 5676 users and 8.5 million tweets.

### 4.5 High-Level Filtering and Pre-Processing

At this point, the data set included multiple noisy elements, and before starting with any further analysis, some high-level filtering and pre-processing methods were applied to both the test and training data set.

During the data collection phase, the most recent tweets for each user were checked for non-English tweets. However, many users had a history of tweeting in other languages than English, leaving multiple non-English tweets in the data set. In order to detect these

tweets, the Textcat<sup>7</sup> library for the statistical computing environment R<sup>8</sup> was used. The language detector was not very reliable, and many tweets were falsely classified as non-English. Most noticeable was the fact that the language detector classified almost 16% of the tweets to be in Afrikaans, although most of those were in fact in English. As a result, tweets classified as Afrikaans were overlooked in the filtering process and thus not removed from the data set.

Besides this, the issue with mis-classifications seemed most frequent on short tweets. Many tweets were also categorised as languages that were likely not to have a great presence in the dataset, such as Manx Gaelic and Middle Frisian. As a compromise, only tweets with length greater than 35 characters<sup>9</sup> and a detected language other than English, or any of the more rare languages<sup>10</sup> (as they were assumed to be misclassified), were removed from the dataset. Users with less than 20 remaining tweets assumed to be in English were also deleted from the dataset. In total, this resulted in a removal of 3.2% of the tweets and 3.6% of the users.

Some pre-processing steps were also applied at this phase. When the data instances were pulled from the web, all entities were encoded in UTF-8. Hence, text variables in the raw data sets were byte literals rather than strings, and all prefixed by a 'b'. This, alongside with foreign symbols and line separators, were removed from all tweets, biographies, user- and display names.

Specific task relevant entities, such as URLs and mentions, were replaced with placeholders. The reason for this was that most of the URLs and user mentions were unique and thereby considered not to contribute with valuable information for the classification problem. The placeholders, however, capture information related to their

---

<sup>7</sup> Textcat Documentation:  
<https://cran.r-project.org/web/packages/textcat/index.html>

<sup>8</sup> R Documentation:  
<https://www.r-project.org/>

<sup>9</sup> The tweet length was calculated on the plain text in each tweet. URLs, mentions, RTs and hashtags were not included.

<sup>10</sup> Rare languages assumed to be falsely classified:  
Manx Gaelic, Middle Frisian, Romansch and Irish Gaelic

#### 4. Data

presences. In order to study the usage of different emojis, each UTF-8 encoding representing an emoji was converted to a descriptive term, prefixed by 'EMOJI'.

**Table 4.4:** Placeholders

Example	Placeholder
@beatED	MENTION
<a href="https://www.beateatingdisorders.org.uk/your-stories/steps-towards-recovery...">https://www.beateatingdisorders.org.uk/your-stories/steps-towards-recovery...</a>	URL
/XF0/X9F/X98/X82	EMOJISmilingFaceWithTearsOfJoy

### 4.6 Annotations

Due to limited resources, the entirety of the dataset was annotated only by the author. This is not ideal, as it increases the risk of mistakes and subjective answers. To evaluate the reliability of the author annotations, three additional control annotators were asked to classify a subset of 100 users each. For each user, they were given username, display name, Twitter biography and a random selection of 200 tweets. All annotators were asked to follow the same guidelines as described in sections 4.1 and 4.2.

The control annotations were used to reveal inter-annotator agreement. As the identification of pro-ED users is the main focus of this work, agreement was calculated for both the multi-class classification and the binary classification. For the binary case, the pro-recovery and the unrelated class were concatenated in order to generate a single non pro-eating disorder class. The pairwise inter-annotator agreement between each control annotator and the author are presented in table 4.5.

Each pair of control annotators had 10 overlapping users in their subset, resulting in 30 users being rated by three annotators, including the original annotation. These users were used to calculate Fleiss' kappa, between all annotators, shown in table 4.6.

**Table 4.5:** Pairwise Cohen's kappa

<i>Between author and</i>	<b>Cohen's kappa:</b>	
	<b>Multi-class :</b>	<b>Binary :</b>
Control annotator 1	0.85	0.93
Control annotator 2	0.98	1
Control annotator 3	0.96	0.98

**Table 4.6:** Fleiss' kappa

<i>Between all annotators</i>	<b>Fleiss' kappa:</b>	
	<b>Multi-class :</b>	<b>Binary :</b>
	0.88	0.92

For the binary cases the kappa scores indicate almost perfect agreement. The multi-class annotations, naturally, generate slightly lower kappa scores. Overall, the agreement scores show high agreement between all annotators. Although this is no guarantee of a flawless annotation, it seems viable to use the original annotated data set for the purpose of this work.

The control annotators were also used to look for presence of non-English tweets and profiles labeled as non pro-ED having tweets satisfying a criterion of positive pro-ED attitude. In the complete set of 270 users, given to different control annotators, only 7 users included some non-English tweets.

Out of the 178 and 29 users, annotated by the author as respectively; unrelated and pro-recovery, no user was in fact classified by the control annotators as pro-ED. The few discrepancies present in the binary annotations were always the other way around. Possible explanations for this could be that the tweet considered as a display of a positive pro-ED attitude by the author, was not included in the random selection of 200 tweets given to the control annotators, or due to differences in interpretation of the inclusion criteria. Regardless, the presence of users labeled as non pro-ED, with a past of tweeting pro-ED content, appeared to be negligible. As a result of these results, no further work was done to limit the amount of non-English, or hidden pro-ED, tweets.

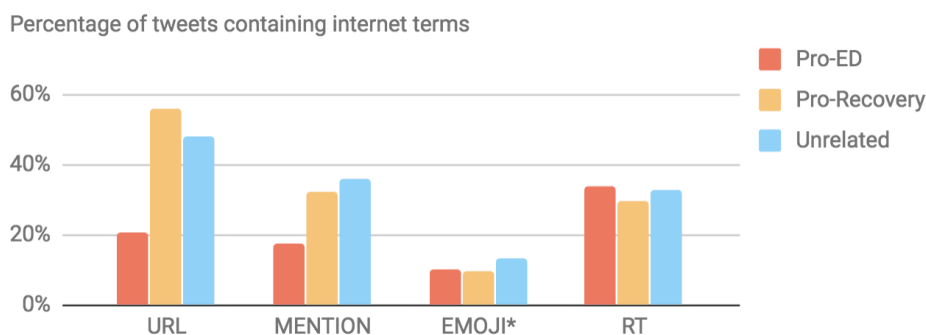
## 4.7 Characteristics

Aiming to answer the third research question regarding characteristics of pro-ED community members on Twitter, this section presents the results from different analyses of tweets and profile information of the 5466 users in the filtered training data set. The conducted analyses cover the presence of specific unigrams and emojis in tweets, users' locations, tweets lengths, part-of-speech and references to eating disorders and related topics. The results identified some distinguishing characteristics of the pro-ED users, and lays the foundation for the selection of features for use in the construction of the classification models.

### 4.7.1 Twitter Functionality and Internet Terms

Compared to traditional textual data, tweets often include some specific unigrams one would not come across outside of the Internet or the Twitter platform. This includes modern Internet entities such as URLs and emojis, and references to Twitter functionality such as user mentions and retweets. Figure 4.2 illustrates the percentages of tweets in the training data set that contained each of these unigrams. The emoji bin includes all the different emojis, while a more detailed study of each is presented in the following section.

**Figure 4.2:** Distribution of Internet and Twitter terms<sup>11</sup>



The histogram showed that the tweets written by pro-ED users had a tendency to include URLs to a considerably lesser extent than the tweets written by both unrelated and pro-recovery users. Also, fewer pro-ED tweets contained mentions of other users. The percentage of tweets being retweets and tweets including an emoji were quite similar for the three classes.

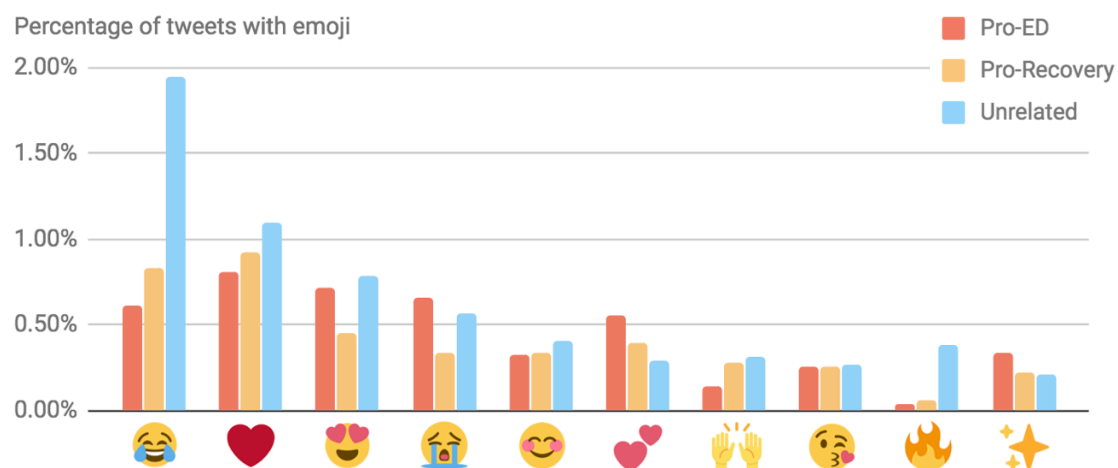
<sup>11</sup> The username representing the original author in a retweet was not counted as a mentions.

### 4.7.2 Emojis

Another trait of modern communication on social media is the usage of emojis. On Twitter, all Unicode emojis<sup>12</sup> are replaced by Twitter's own custom graphics, Twemoji<sup>13</sup>. In the training data set 12.2% of the tweets included an emoji and, as seen in figure 4.2, the percentage of tweets including emojis was somewhat similar for all three classes. Overall, 96% of the users had an emoji in their set of tweets.

For the study of emoji usage, an analysis was first conducted at tweet level, calculating the percentage of tweets from the three classes containing each emoji. For emojis with multiple variants of skin colour, all variants were aggregated and counted as one emoji. Figure 4.3 displays the percentages of tweets including each of the top ten most used emojis<sup>14</sup>.

**Figure 4.3:** Presence of popular emojis in tweets<sup>15</sup>



The histogram illustrates that the overall percentages of tweets including each emoji is low. We observe that the «Smiling face with tears of joy» emoji was more than three times

<sup>12</sup> Complete list of Unicode Emojis:  
<https://unicode.org/emoji/charts/full-emoji-list.html>

<sup>13</sup> Twemoji:  
<https://github.com/twitter/twemoji>

<sup>14</sup> The emojis that were present in the highest number of tweets regardless of class.

<sup>15</sup> Twemoji Graphics are licensed under CC-BY 4.0:  
<https://creativecommons.org/licenses/by/4.0/>

#### 4. Data

more common in tweets from unrelated users, than from pro-ED users. Although, much less frequent, it is also worth noticing the difference in use of the fire emoji.

Further, hoping to detect some less frequently used, yet distinguishing emojis, an analysis was conducted at user-level, calculating the percentage of users that included each emoji in their total set of tweets. The histogram in figure 4.4 displays the percentages of users from each class that included each of the most popular emojis in their tweet set.

**Figure 4.4:** Presence of popular emojis in users' tweets sets

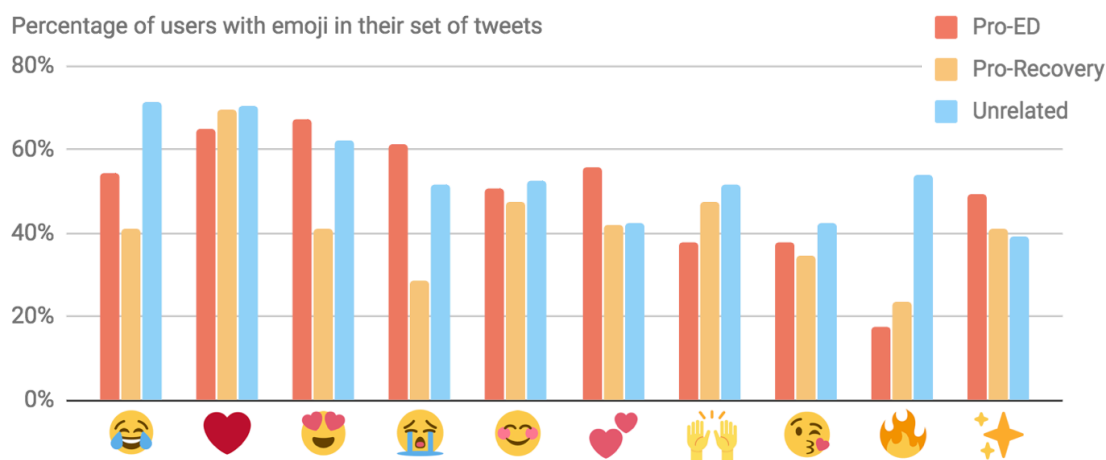
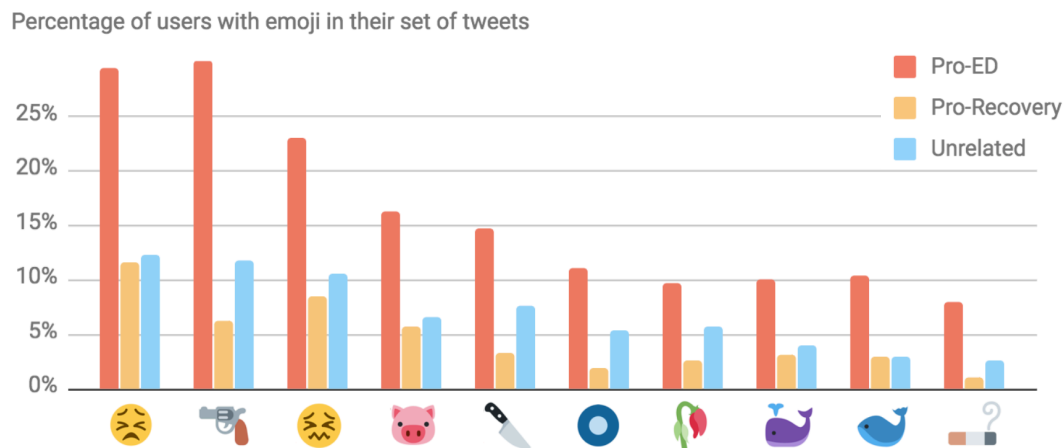


Figure 4.4 shows that the percentage of tweet sets including popular emojis were, for some emojis, quite similar. However, the inclusion of the «Fire» and «Smiling face with tears of joy» emoji were again markedly different, although less prominent than at tweet level. Some of the emojis also have a noticeable gap in usage between the pro-recovery users and users from the other two classes.

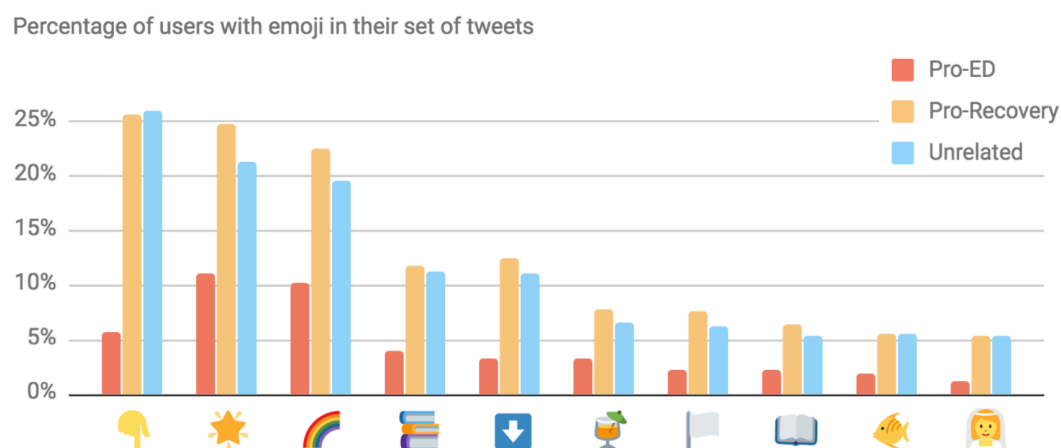
In order to discover more emojis with a high potential of serving as a differentiable feature, the relative difference in percentage was calculated for each emoji. In this step, only emojis present in at least 200 users' tweet set were included. For each class, the ten emojis with the largest relative difference to the two other classes are illustrated in the figures 4.5 - 4.7.



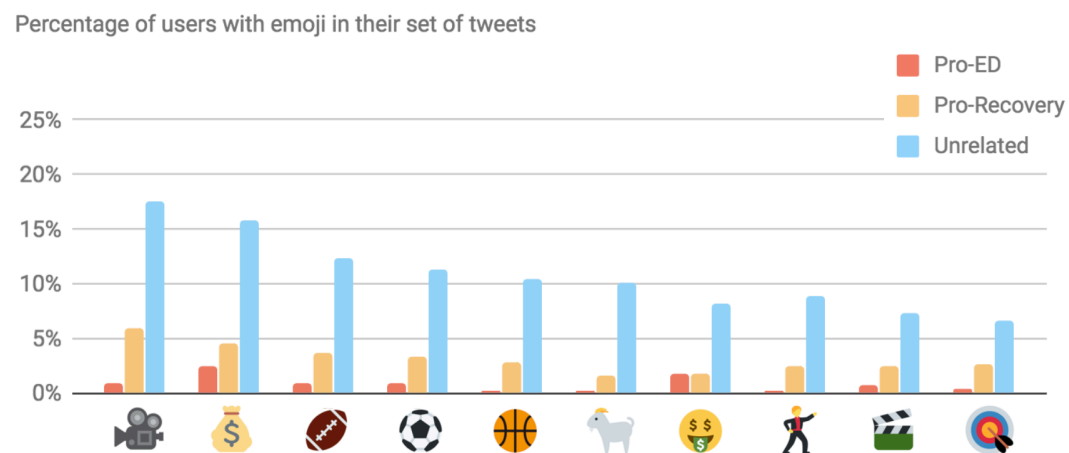
**Figure 4.5:** Presence of popular emojis among pro-ed users



**Figure 4.6:** Presence of popular emojis among pro-recovery users



**Figure 4.7:** Presence of popular emojis among unrelated users



#### 4. Data

According to figure 4.5, the confounded face emoji and the persevering face emoji were clearly more popular among the pro-ED users than the remaining users in the training data set. The pistol and hocho/knife emoji were often seen in context with self-hatred and self-harm, while the pig and the two whale emojis were often used by the pro-ED members as descriptions of body shapes and eating habits, e.g., «Eating like a pig» or «large as a whale».

Interestingly, the emojis with the highest relative positive difference for the pro-recovery class, as presented in figure 4.6, were used approximately the same amount by unrelated users. The calculated difference in percentage was mostly to the pro-ED users. In this set, we find both the down-pointing arrow and the hand pointing downwards, often used to point to links at the bottom of the tweet. This could be put in context to the high percentage of URLs present in tweets from pro-recovery and unrelated users.

The emojis with the highest relative difference for the unrelated class consisted of emojis that generally were included in a lower percentage of the users' tweet sets. Most of the emojis seemed to reference different hobbies, such as sports and movies. Two of the emojis were also related to money. The goat emoji, which was used by 10.1% of the unrelated users, and only 0.2% of the pro-ED users, was, besides alluding to the animal, used instead of GOAT, acronym of «Greatest of All Time», and often used to praise others.

#### 4.7.3 Locations

Each user on Twitter has the opportunity to provide a location to their Twitter account. In reality, this is a string field, and does not necessarily have to carry a physical location. For instance, in the training data set, 2.2% of the pro-ED users had reported «Hell» as their location. According to the GeoText<sup>16</sup> library for Python, only 51% of the non-empty locations in the training set were referencing real cities or countries.

Appendix B.4 contains results from an analysis on the location fields, including the most popular locations and share of null values for each class. Overall, 36% of the users' locations were left empty, and out of the remaining 62% were unique. Because of this, locations were not explored further and were not considered possible features for detection of pro-ED users as they contribute with very little information.

---

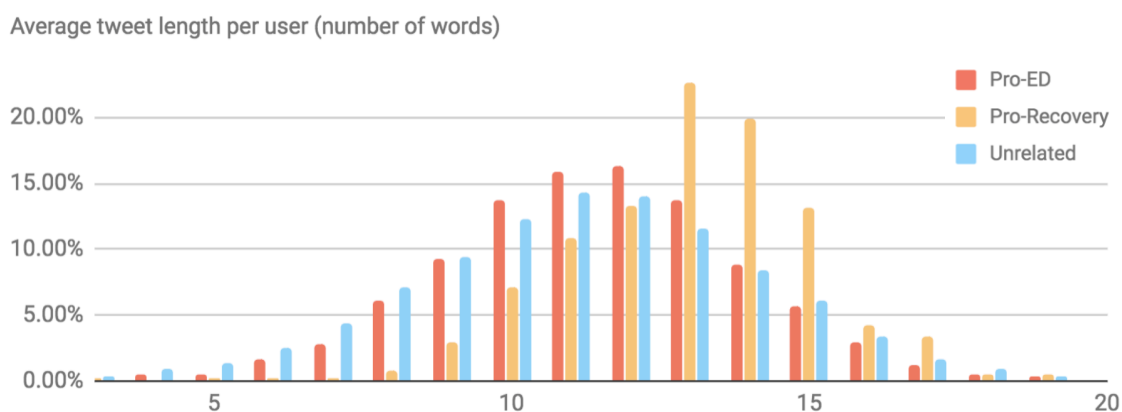
<sup>16</sup> <https://github.com/elyase/geotext>

#### 4.7.4 Tweet length

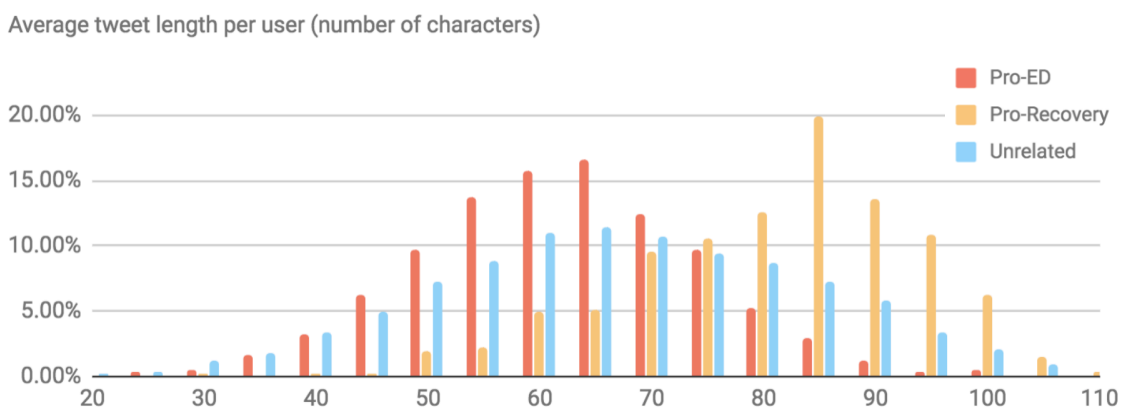
In the search for distinguishing features, average tweet length was considered. Figure 4.8 and 4.9 illustrates the users' average tweet length measured through number of words and characters present in each. The placeholders, i.e., URLs, mentions, RTs and emojis, were not included in this calculation as they do not necessarily correspond to the length of the original tweet.

Both figures also show that the overall average tweet length is longer for the pro-recovery class (12.6 words and 79.3 characters) than the pro-ED (11.4 words and 60 characters) and the unrelated (11.4 words and 66 characters) class. The unrelated users have an average tweet length slightly longer than the pro-ED users; however, the differences are not that large.

**Figure 4.8:** Average tweet lengths in number of words



**Figure 4.9:** Average tweets lengths in number of characters



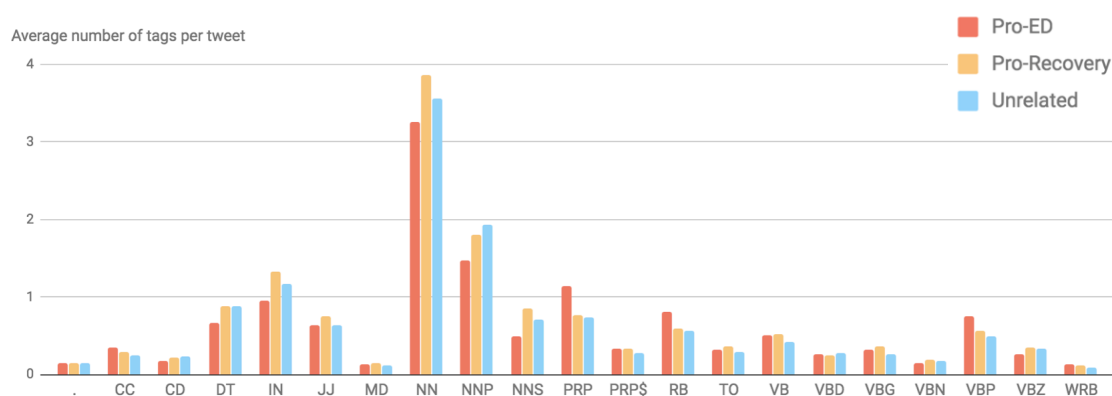
## 4. Data

### 4.7.5 Part-of-Speech

Part-of-Speech tag counts have previously been used as features for different NLP classification tasks, such as sentiment analysis (Gimpel *et al.* 2011). In order to evaluate whether this could serve as distinguishing features for the classification problem in this work, the NLTK package was used to perform POS-tagging on all tweets in the training data set.

Figure 4.10 visualises the average number of each tag per tweet for all three classes. The results presented only display tags with an average minimum average presence of 0.1 per tweet. The complete histogram, including a description of each tag, is presented in appendix B.

**Figure 4.10:** Average distribution of POS-tags in tweets<sup>17</sup>



According to the histogram, an average pro-ED tweet tends to include fewer nouns and more personal pronouns than tweets from the other two classes. However, the differences overall are small and POS-tags do not appear to contain much differentiating information.

### 4.7.6 References to Eating Disorders and Related Topics

Motivated by Arseniev-Koehler et al.'s (2016) study of references to eating disorders in pro-ana users' tweets and profile information, similar analyses were conducted to evaluate the presence of such in the training data set. However, Arseniev-Koehler et al.'s work only considered users, and their followers, from the pro-ana community, and they

<sup>17</sup> Adjusted for average tweet length per class

did not include any control set of users. The researchers assumed that words such as «exercise» or «dinner» would imply a reference to disordered behaviour, given that the tweet was written by a pro-ana user. However, for the data set in this work, which included users annotated as pro-recovery and unrelated, this assumption would not hold. Instead, a new codebook was defined. The new codebook was also divided into different domains in order to count references to each in the training data set. The complete codebook is presented in table 4.7. Explanations for different pro-ED terms and phrases are found in appendix A.

For the following analyses it was assumed that a word from the codebook implied a reference to the domain to which it belongs. It is worth mentioning that terms such as «kg» and «weight» not necessarily have to refer to *body* weight, and similarly «vomit» and «throwup» do not always reference a compensatory action, especially not when it is used by an unrelated user. «Ana» and «Mia» could also refer to a person's name, rather than the pro-ED communities. Because of this, it is acknowledged that the results presented in the following sub-sections are likely to include some false references. However, it is not expected that these will overshadow the interesting finds.

**Table 4.7:** Codebook<sup>18</sup>

<b>Domain</b>	<b>Words</b>
<b>Eating Disorders</b>	anorexia, anorexic, bulimia, bulimic, ednos, eating disorders, eating disorder, eatingdisorder, eatingdisorders, ed, eds
<b>Pro-ED Communities</b>	proana, pro-ana, promia, pro-mia, proed, pro-ED, ana, mia, thinspo, thinspiration, bonespo, bonespiration
<b>Body Image</b>	overweight, obese, fat, fatty, chubby, skinny, thin, skinnier, emaciated, bony, bone, bones, hipbones, thighgap, collarbone, collarbones, spine, backbone, bikinibridge, waist
<b>Body Weight</b>	weight, scale, kg, lbs, gw, lw, ugw, bmi, pound, cw
<b>Food &amp; Diet</b>	calorie, calories, abcdiet, Russian gymnast, binge, starve, fasted, fasting, starving, starvation, diet, projectthin, skinny4xmas, skinny4christmas, calorieapril, skinnyforsummer, weightloss
<b>Compensatory Behaviour</b>	laxies, laxatives, vomit, vomited, purge, puke, purging, throwup

<sup>18</sup> For explanations of pro-ED terms and hashtags, see appendix A.

#### 4. Data

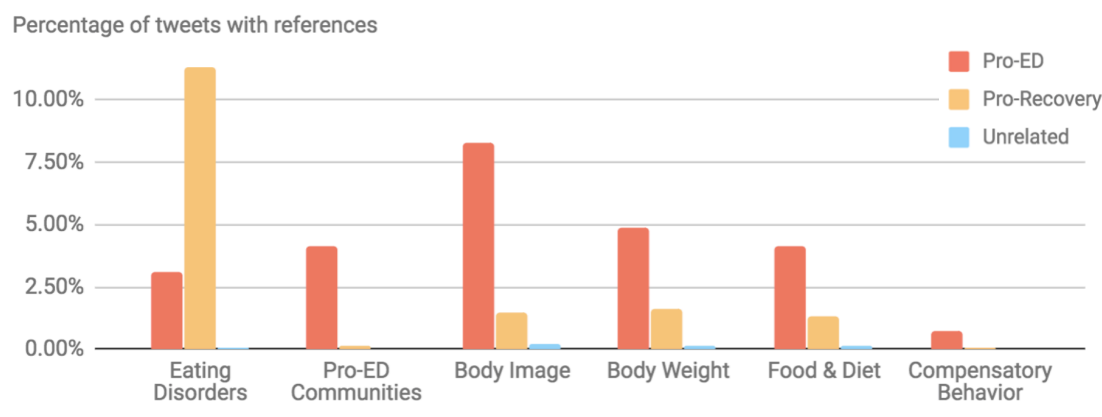
##### Tweets

Given the method used to collect and annotate data in this work, each pro-ED user does, strictly speaking, only have to include *one* positive display of a pro-ED attitude in either a tweet or profile information. The remaining tweets could, hypothetically, concern any other topic. In order to get a sense of how common references to the different domains in the codebook were among the tweets written by pro-ED users in our data set, and to be able to compare the results to the other two classes, all tweets in the training data set were checked for presence of each reference in the codebook.

Each tweet reference was considered valid if the exact word/phrase was found preceded and followed by either a space, punctuation or nothing. For instance, «Hogwarts» or «I'm eating a **banana**» would not be considered valid references, but «I'm getting closer to my **gw**» or «**Ana** is all that matters» would<sup>19</sup>.

Overall, 21.59% of the pro-ED profiles' tweets contained a reference from our codebook. In comparison, 14.98% and 0.55% of the tweets from pro-recovery and unrelated users included similar references. The percentages of tweets including references from each domain, as presented in figure 4.11, showed that 11.4% of the tweets from pro-recovery users' and 3.1% of the tweets from pro-ED users' included a reference from the «eating disorder» domain. For the remaining domains, the pro-ED tweets had a much larger percentage of references than the other two classes. For the unrelated class the percentages of tweets including references were less than 0.20% for all domains.

**Figure 4.11:** References in tweets.



<sup>19</sup> **GW:** Acronym for «Goal Weight»

**Ana:** Reference to the pro-anorexia community, often seen personified

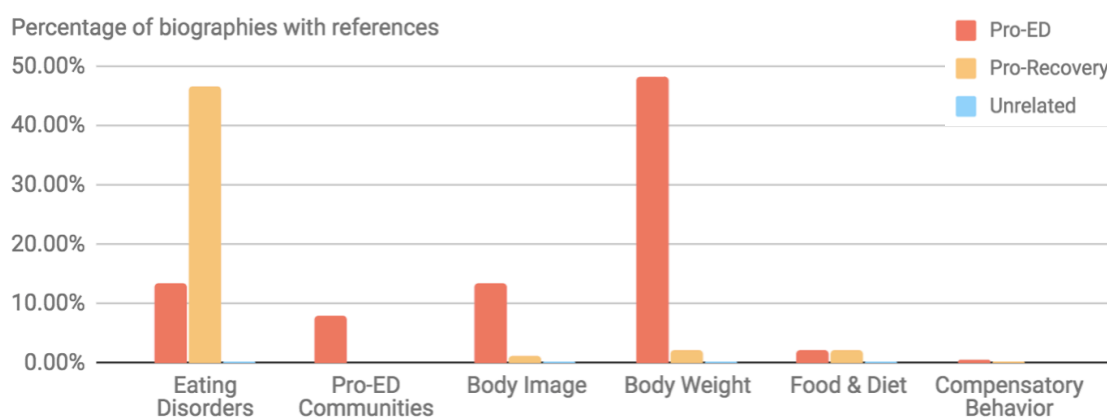
### Biographies

A similar study was conducted on the profiles' biographies, and the results are visualised in figure 4.12. As with tweets, the presence of references to eating disorders were largest among the pro-recovery users.

What seems to be a common trend in the pro-ED community is to inform other users about one's current weight (CW), goal weight (GW), ultimate goal weight (UGW) and lowest achieved weight (LW). This information is often seen presented in pro-ED user's biographies. Some users also include similar information such as height, body mass index (BMI) or clothing size. As this is not something one would assume either recovery users or unrelated users would typically chose to include in their biographies, it bore the potential to serve as a differentiating feature.

As seen in figure 4.12, the reference analysis showed that 48.3% of the pro-ED users in the data set included a weight reference in their biography, compared to 2.1% and 0.3% for the recovery oriented and the unrelated users.

**Figure 4.12:** References in biographies.



### Username and Display names

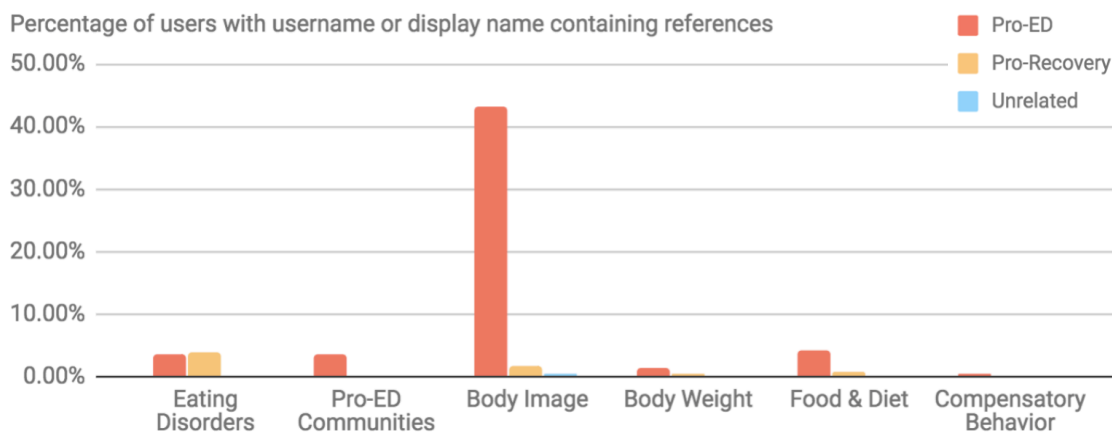
One would normally assume that usernames and display names on Twitter would reflect the name of the person or company behind the user profile. However, as many pro-ED profiles are anonymous, some users tend to choose names reflecting their affiliation to the communities. Usernames and display names were therefore also checked for references from the codebook.

#### 4. Data

In the previous experiments only spaced references were considered valid. However, usernames are required to not contain any spaces or non-alphanumeric characters<sup>20</sup>, thus following the same procedure would be pointless. Character sequences matching those in the codebook were because of this considered valid. In order to reduce the problem of false references, no two-letter words in the codebook were included. Additionally, «ana» and «mia», often referring to real names or parts of other words, e.g., «banana», were also not counted as references.

Figure 4.13 shows that more than 40% of the pro-ED users had a username or display name referencing body image. A more detailed study showed that 16.7% of the pro-ED users' names contained the word «thin» and 15.4% contained «skinny». References from the other domains in the codebook were not used to the same extent.

**Figure 4.13:** References in usernames or display names.



<sup>20</sup> With the exception of underscores.



## 5. Architecture

In order to achieve automatic identification of pro-eating disorder users on Twitter, a handful of different models and feature sets were explored with the interest of finding the best performing classifier. As previously described in the review of related literature, the general approach to building a text classification system can be divided into four steps; collection of data, pre-processing, feature extraction and model building. Whereas the proceeding chapter covered the first and parts of the second step, this chapter continues where the prior left off. First, section 5.1 presents the further pre-processing steps that were applied to the data. Section 5.2 then describes the process of extracting five sets of features that were used in later experiments, and finally, four supervised machine learners that were employed in this work are presented in section 5.3.

The goal of this study was to detect pro-eating disorder members on Twitter. Thus, for the remaining part of this thesis, the focus will switch from the three-class partitioning (pro-ED, pro-recovery and unrelated) to the binary classification problem; pro-ED vs. Non pro-ED (pro-recovery + unrelated).

The machine learning library Scikit-learn (Pedregosa *et al.* 2011) was employed during the pre-processing, feature extractions and experiments described in both this and the following chapter. For further information about the referenced tools and methods, the reader is referred to Scikit-Learn's online documentation (Scikit-Learn Developers, 2017).

### 5.1 Pre-Processing Continuation

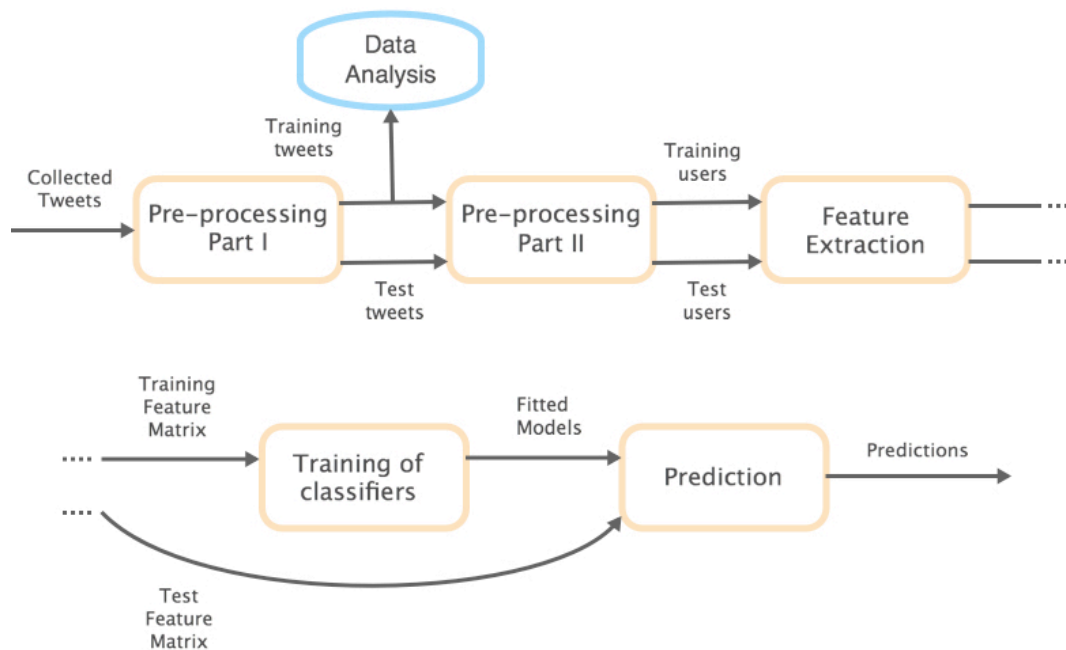
Section 4.5 presented the high-level filtering and pre-processing steps that were applied to the data set prior to the analysis of characteristics, including the removal of non-English tweets and users without enough source data, and replacement of URLs, user mentions, retweets and emojis with placeholder tags. Figure 5.1 illustrates how all tweets

## 5. Architecture

went through these initial pre-processing steps before the training section of the data set was analysed, and how the two parallel tracks of the test and training data were treated equally up to the training of classifiers.

Due to limited time, the effects of utilizing different pre-processing methods have not been explored in this work. Instead, some basic pre-processing steps, that have been seen used in related studies, were applied. All classification systems presented in this thesis went through the same pre-processing steps.

**Figure 5.1:** Classification system



### 5.1.1 Tweet Aggregation

The labelling of data in this study was conducted at user level, annotating each user as either 'pro-ED' or 'non pro-ED' (pro-recovery or unrelated). After the data collection phase, the complete data set consisted of millions of documents, each corresponding to a single tweet, along with the user information and class label. However, a tweet written by a pro-ED user does not necessarily reflect any pro-ED attitude, and since this work focuses on the classification of users, as opposed to single tweets, it made the most sense to aggregate all tweets per user together. This way, each user corresponded to one document with a single string field consisting of all consecutive tweets, similar to the

approaches presented in the related work of Preoțiuc-Pietro *et al.* (2015) and Balasuriya *et al.* (2016).

### 5.1.2 Pre-Processing in Scikit-Learn

During the extraction of features, which is described further in section 5.2, the Scikit-learn library was used. In the process of making feature vectors, some pre-processing steps were incorporated in the implementations and are thus presented in this subsection. The incorporated pre-processing steps included lowercasing and removal (i.e., ignoring) of non-alphanumeric characters and stopwords.

Scikit-learn's components for feature vector creation automatically treated all punctuation as token separators, which for instance caused hashtags to be considered normal words and terms like «pro-ana» to be regarded as two separate words: «pro» and «ana». All characters were also converted to lower case before tokenisation by default. Since hashtags are case insensitive, some users choose to write abbreviations and first letters in capital letters, while others write all hashtags in lowercase. Also first words of tweets tend to begin with upper case. By lowercasing, and thus treating all letters the same way, it is ensured that equal hashtags and words get treated as the same token. Moreover, it decreases the term space. The process of lowercasing will also incorrectly merge different terms that happens to have the same spelling. However, it is reasonable to assume that this would only affect a small portion of the data.

As described in section 2.2.4, stop words are commonly used words that are removed from text in natural language processing. The high-level vectorising components in Scikit-learn let the encoder specify a stop word list in order for the vectoriser to overlook these terms during the feature extraction. For the purpose of this work, words from the NLTK English stop word list and words that were used to find users during the data collection phase were excluded for the sake of preventing any bias created by this collection method. The complete list of stop words used in this thesis is to be found in appendix C.1.

## 5.2 Feature Extraction

Successful automatic classification relies on having vector representations of features for the data set objects, and the performance depends on choosing features that are informative and differentiating. In order to do so, domain knowledge is valuable. Based

## 5. Architecture

on the data analysis presented in the previous chapter, and the features used by the related work presented in section 3.3.3, five groups of features were extracted from the data: Unigrams and bigrams from tweets, emoji features from tweets, unigrams and bigrams from biographies and character  $n$ -grams from usernames and display names.

### 5.2.1 Feature extraction in Scikit-learn

The Scikit-learn module `feature_extraction.text` contains multiple tools for conversion of raw textual document collections to numerical feature matrices. This includes the `TfidfVectorizer` class, which implements tokenisation, occurrence counting and Term frequency-Inverse document frequency weighting, as well as the pre-processing steps aforementioned. To elaborate, the `TfidfVectorizer` first tokenises the documents, learns the vocabulary of the complete collection and calculates the inverse document frequency weights. Each document in the collection is then encoded as an array with the weighted frequencies of each term occurrence for the purpose of generating a clean and numeric input for the machine learning algorithms.

Scikit-learn also lets the encoder specify a lower and upper boundary for the vocabulary size, a custom stop word list, whether the extracted features should be made out of words or characters, and what range of  $n$  to consider during  $n$ -gram extractions.

In this work, the `TfidfVectorizer`, along with the stop word list described in section 5.1.2, was employed to extract all features. Unless otherwise specified, the default parameters were used.

### 5.2.2 Feature groups

With the interest of comparing the performance of models trained on different types of features from the Twitter users, five feature groups were established:

- **Unigram features from tweets:** This group included, as the name suggests, single words (i.e., unigrams) from the combined tweet text, including the internet/Twitter functionality placeholders, i.e., MENTION, URL and RT. Emojis were not included in this group.

- **Bigram features from tweets:** Similar to the unigram features, but with bigrams. This was achieved by changing the *ngram\_range* parameter of the `TfidfVectorizer` to (2,2).
- **Emoji features from tweets:** The emoji features consisted of the unigram emoji placeholders from tweets. The extraction of emoji features was achieved by specifying a fixed vocabulary, consisting of the emoji placeholder tags, to the `TfidfVectorizer`.
- **Biography features:** This group included unigrams and bigrams from the biographies of the Twitter users. Achieved by adjusting the *ngram\_range* parameter to (1,2). The bigram features included both MENTION, URL, and all of the emoji placeholders.
- **Name features:** By adjusting the *analyser* parameter of the `TfidfVectorizer` to 'char' instead of 'word', and the *ngram\_range* parameter to (3, 15), character *n*-grams of length 3 and up to 15 characters (maximum length of username) were extracted from the users' display names and usernames.

The decision of choosing and assembling the features in this fashion was motivated by the results of the data analyses, and the desire for comparing the feature types against each other. The results from the study of characteristics, presented in section 4.7, indicated large differences in the usage of specific words related to eating disorders and related topics. Unigram and bigram features from the collection of tweets and biographies were extracted in order to capture such textual information. Emojis were kept in a separate group as it was considered interesting to measure their differentiating abilities on their own. The data analyses also showed a great difference in the amount of users including ED references in their usernames or display names, particularly references to body image. As usernames are required not to contain spaces, character *n*-grams, rather than word *n*-grams, were extracted.

### 5.3 Classifiers

This work explored using four different classification models; a Naïve Bayes model, a linear Support Vector Machine, a Random Forest and a Logistic regression model. The models were chosen on the basis of their reputation of performing well over textual

## 5. Architecture

features, and their popularity among the studies presented as related work (see section 3.3.4).

These models also produce interpretable results, which give the opportunity to draw lines between the data analyses, prior content analyses of pro-ED content from the related work, and the features considered most informative by the classification models.

The experiments presented in the next chapter employed the following implementations:

- **NB**: Naïve Bayes (`sklearn.naive_bayes.MultinomialNB`)
- **SVM**: Linear kernel Support Vector Classification: (`sklearn.svm.LinearSVC`)
- **RF**: Random Forest (`sklearn.ensemble.RandomForestClassifier`)
- **LR**: Logistic Regression (`sklearn.linear_model.LogisticRegression`)

All models used default parameters unless otherwise stated.

Scikit-Learn also offers a collective voting classifier (`sklearn.ensemble.VotingClassifier`), **VC**, which was used towards the end of the experiments to wrap the best models and average their predictions.

## 6. Experiments & Results

This chapter presents the experiments conducted to find an efficient approach to classification of Twitter users with respect to pro-ED membership.

In the process of designing a text classification system, a variety of different combinations of pre-processing methods, feature types, and machine learning approaches could be explored, as well as extensive experimentation with parameters. However, due to limited time, this work decided to employ well-known pre-processing steps and machine learning algorithms, and focus primarily on exploring the different feature groups' predictiveness and contribution to solving the classification problem at hand.

### 6.1 Experimental Setups

Chapter 5 described the process of extracting features and presented four machine learning algorithms. The following sections present the procedure of experimenting with aforementioned features groups and learning algorithms on the training data set. First, for the sake of evaluating the feature groups' informativeness, each group was used separately to train four different classification models: a Naïve Bayes model (NB), a Support Vector Machine (SVM), a Random Forest (RF) and a Logistic Regression model (LR). Vocabularies of different size were examined for most of the feature groups. Predictive features within each group were investigated by studying the weight coefficients of the Support Vector Machine.

As this chapter later will show, all feature groups managed to classify users with good performance scores and were thus considered relevant in the continuation of the experiments. The different learning algorithms were then trained on a combination of all feature groups with equal weighting. Each feature group's contribution was assessed by conducting a feature ablation study, and its result was used to decide on a new weighting

## 6. Experiments & Results

of the features for the final training of models. The final models, along with a collective voting model, were finally tested on the unseen test data.

### 6.2 Interpretation of Results

To evaluate classifiers there are multiple metrics that could be used, some of which were introduced in section 2.1.5. For the following experiments with machine learning methods and informativeness of feature groups, the results were evaluated by considering the classifiers' precision, recall and  $F_1$ -scores. In this thesis, the three metrics are reported for the 'pro-ED' and 'non pro-ED' class separately because of the class imbalance in this work's data set.

The distribution of users in the data set, as discussed in section 4.4, did not accurately reflect the user population on Twitter. In the data sets used to train and test the classifiers, the proportion of pro-ED users was approximately one-third. However, in reality the pro-ED communities are small and their members make up a tiny fraction of the millions of users on Twitter. Consequently, the results needed to be examined with this caveat in mind, acknowledging that in particular the falsely classifications of pro-ED users were expected to be larger for the actual population compared to this work's data set. Moreover, if the purpose of the classification were restriction policies, a high false-positive rate on the pro-ED class would penalise outsiders. Although the magnitude is unknown, and not possible to estimate without further research, low false positive rates on the pro-ED classification was emphasised.

Overall, the ideal classifier was considered to be one scoring great on precision, while still obtaining a high recall and  $F_1$ -score on the pro-ED class. The scores of the 'non pro-ED' classification were also taken into account; however, it was reasonable to expect that this classification problem would overall yield better performance scores.

### 6.3 Feature study

Features are meant to represent aspects of the data objects that are relevant to the classification problem. Based on the observed differences from the data analyses and related work, section 5.2.2 presented groups of observable features expected to inform



the classifiers about the unobservable class label. However, the chosen feature groups would not all necessarily help the models to make better predictions. The following experiments first address each feature group's informativeness by training the models on each feature group on its own, and then its contribution when used in combination with other feature groups. In all of the experiments each of the four machine learning algorithms was used to train a model.

In addition to reporting the findings from each of the four models' predictions, the following sub-sections also present the top most informative features within each feature group according to the weight coefficients for the linear Support Vector Machine. Unlike some of the implementations available in the Scikit-learn library that act as black boxes, LinearSVC lets the encoder access the coefficients of the weight vector. As explained in the background chapter, a support vector machine creates a hyperplane for the sake of separating the data points belonging to each of the two classes. By studying the coordinates of the weight vector orthogonal to the optimal hyperplane, information regarding class (direction) and impact (relative size) is accessible (Guyon, 2003).

Due to time constraints and the fact that support vector machines often scored best on performance in the related work, this study decided to only focus on the SVM feature weights in order to measure the single features' informativeness. Nevertheless, studying the top features for one of the models makes the magic behind the predictions more transparent and gives the ability to compare the top features to the previously conducted analyses and findings of related work.

### 6.3.1 Unigrams from tweet text

Each of the four machine learning algorithms described in section 5.3 was used to build a model trained on unigram features from the user's tweets. In order to find a suitable vocabulary size, i.e., number of unigrams to consider, four different maximum limits for the vocabulary size were explored; 2000, 10,000, 20,000 and no limit (i.e., the total number of unique terms). The results, presented in full in appendix C, showed that using a vocabulary size of 20,000 yielded the best performance score for most of the models.

Table 6.1 presents the average scores of the four models, each trained on a tf-idf weighted vocabulary of 20,000 unigrams from tweets, evaluated under a 5-fold cross validation scheme.

## 6. Experiments & Results

**Table 6.1:** Classification results based on 5-fold cross-validation

Features	Model	Pro-ED			Non pro-ED		
		Precision	Recall	F1-score	Precision	Recall	F1-score
20,000 unigram features from tweets	NB	0.923	<b>0.994</b>	0.957	0.997	0.958	0.977
	SVM	0.982	0.974	<b>0.978</b>	0.987	0.991	0.989
	RF	<b>0.986</b>	0.966	0.976	0.983	0.993	0.988
	LR	0.981	0.974	0.977	0.987	0.990	0.989

The table shows that all models yielded strikingly strong performance scores. With this level of certainty, it was necessary to investigate what features the predictions were based on to make sure the classifiers did not have access to information they were not supposed to have. Tables 6.2 and 6.3 present the words with the highest and lowest weight coefficients for the SVM model, and provide useful insight into what unigrams this classifier used to make its class predictions. Positive weights contribute to a positive pro-ED classification and vice versa.

Most of the top features seemed very reasonable and many of them have already been addressed in previous parts of this thesis, including «skinny», «ana» and «collarbones». «Skinnynewyear» is referencing a pro-ED event, and the «edprobs» (Eating Disorder problems) unigram is a popular hashtag for tweets addressing relatable situations among people with eating disorder and has been shown popular in a prior content analysis of pro-ED communities (Bert *et al.* 2016). More surprising is perhaps the more common «ll»(will), «till»(until), «someday» and «bad». As the first three are related to future, one could hypothesise that their differentiation ability is related to the articulation of the transformation to a future thinner self, presented as a relevant feature of pro-ana profiles in Bates' study of self-descriptions in the pro-ana community (Bates, 2015).

Besides the word «recovery», most of the unigrams with negative weighting do not explicitly reference eating disorders or related topics.

**Table 6.2:** Unigrams with highest weights

Weight coefficient	Feature
0.1404	skinny
0.1134	ana
0.0996	skinnynewyear
0.0981	edprobs
0.0980	ll
0.0966	collarbones
0.0963	till
0.0838	thread
0.0832	someday
0.0824	bad

**Table 6.3:** Unigrams with lowest weights

Weight coefficient	Feature
-0.1445	thankyou
-0.0706	yall
-0.0683	god
-0.0676	doesnt
-0.0651	newprofilepic
-0.0650	things
-0.0640	hours
-0.0620	photography
-0.0583	world
-0.0583	recovery

### 6.3.2 Bigrams from tweet text

Following the same procedure as described for the unigram feature group, the bigram features from tweets were explored. As presented in appendix C, using a maximum limit of 10,000 bigrams in the tf-idf weighted vocabulary gave the best performance scores on a whole. Table 6.4 presents the classification results on the training data set with 5-fold cross-validation. As with unigrams, the classifiers scored overall remarkably well, although slightly worse than the classifiers trained on unigrams.

**Table 6.4:** Classification results based on 5-fold cross-validation

Features	Model	Pro-ED			Non pro-ED		
		Precision	Recall	F1-score	Precision	Recall	F1-score
10,000 bigram features from tweets	NB	0.959	<b>0.971</b>	0.965	0.985	0.979	0.982
	SVM	<b>0.985</b>	0.963	<b>0.974</b>	0.981	0.993	0.987
	RF	0.980	0.953	0.966	0.976	0.990	0.983
	LR	0.984	0.918	0.950	0.960	0.992	0.976

Table 6.5 and 6.6 display the ranked bigram features according to the weight coefficient for the SVM model. Similar to the unigram features, most of top discriminating features

## 6. Experiments & Results

for the pro-ED class were related to weight loss, body shape and eating disorders, while the features predicting the non pro-ED class were more general. Note that stopwords have been removed, such that «cup of coffee» has been converted to «cup coffee», etc.

**Table 6.5:** Bigrams with highest weights

Weight coefficient	Feature
2.7765	lose weight
2.2288	im fat
2.2035	ed twitter
2.2010	fat RT
2.0403	fat fat
1.8845	MENTION want
1.8238	hourly URL
1.8146	weight RT
1.7854	skinny URL
1.6890	MENTION im

**Table 6.6:** Bigrams with lowest weights

Weight coefficient	Feature
-1.0537	happy birthday
-0.8813	recovery URL
-0.8745	mention thankyou
-0.8496	mention happy
-0.6629	mention new
-0.6471	youre kind
-0.6396	URL happy
-0.6242	social media
-0.6131	newprofilepic URL
-0.5537	cup coffee











### 6.3.3 Emojis

For the experiments with emoji features, different vocabulary sizes were not explored and all possible emojis ( $n = 1248$ ) were taken into account during the training of the classifiers. The prediction results, measured with 5-fold cross validation, are shown in table 6.7. The classifiers trained solely on emoji unigram placeholders from tweets achieved lower scores than the classifiers trained on unigram or bigram features.











**Table 6.7:** Classification results based on 5-fold cross-validation

Features	Model	Pro-ED			Non pro-ED		
		Precision	Recall	F1-score	Precision	Recall	F1-score
1248 emoji features from tweets	NB	<b>0.911</b>	0.734	0.813	0.877	0.963	0.918
	SVM	0.871	<b>0.806</b>	<b>0.837</b>	0.905	0.939	0.922
	RF	0.877	0.694	0.775	0.859	0.951	0.903
	LR	0.882	0.771	0.823	0.890	0.948	0.918

**Table 6.8:** Emoji features with highest weight coefficients.<sup>1</sup>

Weight coefficient	Feature	Emoji
2.7230	pistol	
2.3033	pensiveface	
2.0306	pigface	
1.8308	spoutingwhale	
1.7339	happypersonraisingonehand	
1.7188	confoundedface	
1.7172	ribbon	
1.7127	pig	
1.6863	bikini	
1.5797	whale	

**Table 6.9:** Emoji features with lowest weight coefficients.

Weight coefficient	Feature	Emoji
-2.2376	rolf	
-2.1901	smilingfacewithsunglasses	
-2.0771	downpointingbackhandindex	
-2.0422	basketballandhoop	
-2.0325	dogface	
-2.0242	facepalm	
-2.0195	books	
-2.0194	rightpointingbackhandindex	
-1.8760	snake	
-1.7724	birthdaycake	

<sup>1</sup> Twemoji Graphics are licensed under CC-BY 4.0: (See image references)  
<https://creativecommons.org/licenses/by/4.0/>

## 6. Experiments & Results

In the ranking of the most informative single features for the SVM model, presented in tables 6.8 and 6.9<sup>2</sup>, most of the features have already been addressed in the discussion of emojis in the analyses of data characteristics in chapter 4 (see section 4.7.2).

### 6.3.4 Biographies

The study of biography features followed the same procedure as the above mentioned; however, both unigrams and bigrams were considered possible features, and instead of using the tweets as data source, the features were extracted from the users' biographies. The study of different vocabulary sizes are once again presented in appendix C. As vocabularies of size 2000 and 10,000 provided comparable results, the former was chosen in order to limit the number of dimensions. Table 6.10 displays the results of the classifiers trained on biography features on the training data set with 5-fold cross-validation.

Compared to the models trained on unigram and bigram features from tweet text, these models scored lower in terms of performance. However, it is worth noticing the large difference in amounts of data these models predictions are based on. Each user in the data set has one biography, consisting of somewhere between 0(NULL) to 160 characters. On the other hand, the average user in the data set has 1500 tweets, with each potentially reaching up to 280 characters.

**Table 6.10:** Classification results based on 5-fold cross-validation

Features	Model	Pro-ED			Non pro-ED		
		Precision	Recall	F1-score	Precision	Recall	F1-score
2000 unigram & bigram features from bios	NB	<b>0.934</b>	0.715	<b>0.810</b>	0.870	0.974	0.919
	SVM	0.880	<b>0.745</b>	0.807	0.880	0.948	0.913
	RF	0.846	0.718	0.777	0.866	0.933	0.899
	LR	0.928	0.660	0.772	0.849	0.974	0.907

Most of the highest ranked features in terms of informativeness, see tables 6.11, were familiar and already touched upon in previous parts of the thesis. As presented in section 4.7.6, references to body weight in biographies are a differentiating characteristic of pro-

<sup>2</sup> The EMOJI prefix is removed from the feature names in order to make them easier to read.

ED profiles on Twitter. This is supported by the weight coefficients to the support vector machine model as the highest ranked features includes both «cw» (current weight) and «gw» (goal weight). References to body shape, such as «skinny» and «fat», are also found. The «tw» («trigger warning») unigram was often used to warn other users against one's own content.

Interestingly, table 6.12 shows that the placeholders for URLs and user mentions are ranked highest on informativeness for the negative class, i.e., 'non pro-ED', in the SVM model. As users often tend to include other Twitter accounts they are associated with in their biography, e.g., «Journalist at @BBCWorld», or similarly, links to other social media accounts or home pages, it could appear that these placeholders' informativeness is related to the fact that many pro-ED users are anonymous. Another reason could be that the 'non pro-ED' class includes more organisations, brands and companies that maybe are more likely to include links to home-pages, or related Twitter profiles, in their biographies.

The only bigram feature in the top informative features was «recovering anorexia» for the non pro-ED class.

**Table 6.11:** Bio' features with highest weights

Weight coefficient	Feature
2.7913	cw
2.5922	avi
2.5431	skinny
2.4791	ana
2.3647	gw
2.2301	fat
1.8059	pro
1.7248	anymore
1.7143	tw
1.6404	ed

**Table 6.12:** Bio' features with lowest weights

Weight coefficient	Feature
-1.5465	URL
-1.3793	MENTION
-1.2647	recovering anorexia
-1.2164	make
-1.2083	blessed
-1.1776	fresh
-1.0818	positivity
-1.0114	lover
-0.9726	news
-0.959	kids

## 6. Experiments & Results

### 6.3.5 Names

As explained in section 5.2, the name feature group consisted of character  $n$ -grams of length 3 up to 15, derived from the users' usernames and display names. A vocabulary size of 10,000 character  $n$ -grams yielded the best result after cross-validation, as presented in table 6.13. As with the biography and emoji features, the results are overall lower than the results from the models trained on unigram and bigram features.

**Table 6.13:** Classification results based on 5-fold cross-validation

Features	Model	Pro-ED			Non pro-ED		
		Precision	Recall	F1-score	Precision	Recall	F1-score
10,000 char n-grams features from names	NB	0.936	0.623	0.748	0.836	0.978	0.902
	SVM	0.910	<b>0.741</b>	<b>0.817</b>	0.879	0.963	0.919
	RF	0.881	0.705	0.783	0.863	0.951	0.905
	LR	<b>0.958</b>	0.557	0.704	0.814	0.988	0.892

**Table 6.14:** Highest weighted name features

Weight coefficient	Feature
3.8810	ana
3.1775	thin
2.7817	hin
2.7535	fat
2.4240	thi
2.3434	skin
2.1880	nny
2.0870	ski
2.0617	bone
1.9254	lbs

**Table 6.15:** Lowest weighted name features

Weight coefficient	Feature
1.2256	ian
-1.1541	mmy
-1.1305	ede
-1.0976	bri
-1.0508	hing
-1.0024	nut
-1.0019	fav
-0.9891	jen
-0.9840	thing
-0.9542	cha

Similar to the experiments presented above, the features with the highest and lowest weight coefficients for the SVM are presented in the tables 6.14 and 6.15. The data analyses in chapter 4 showed large differences in the amount of body image references in



user and display names, and these are reflected in the list of informative features for the classification. Besides the character sequences; «ana», «bone», «fat», «thin» and «lbs», which are all clearly related to body image and eating disorders, the remaining  $n$ -grams could all be considered part of such references, as «thi», «hin» make up «thin», and «skin», «ski», «nny» similarly constitute the word «skinny».

## 6.4 Combining feature groups

The findings from the last section’s experiments demonstrated that all of the feature groups scored well on performance. As a result, it was decided to include all of them in a combination approach where each of the four machine learning algorithms was used to train a model on a combination of all the feature groups with equal weights. The results on the training data set with 5-fold cross-validation are presented in table 6.16.

**Table 6.16:** Classification results based on 5-fold cross-validation

Features	Model	Pro-ED			Non pro-ED		
		Precision	Recall	F1-score	Precision	Recall	F1-score
Unigram + Bigram + Emoji + Bio + Name	NB	0.921	<b>0.995</b>	0.956	0.997	0.957	0.976
	SVM	0.982	0.978	<b>0.980</b>	0.989	0.991	0.990
	RF	<b>0.988</b>	0.966	0.977	0.983	0.994	0.988
	LR	0.985	0.976	<b>0.980</b>	0.988	0.992	0.990

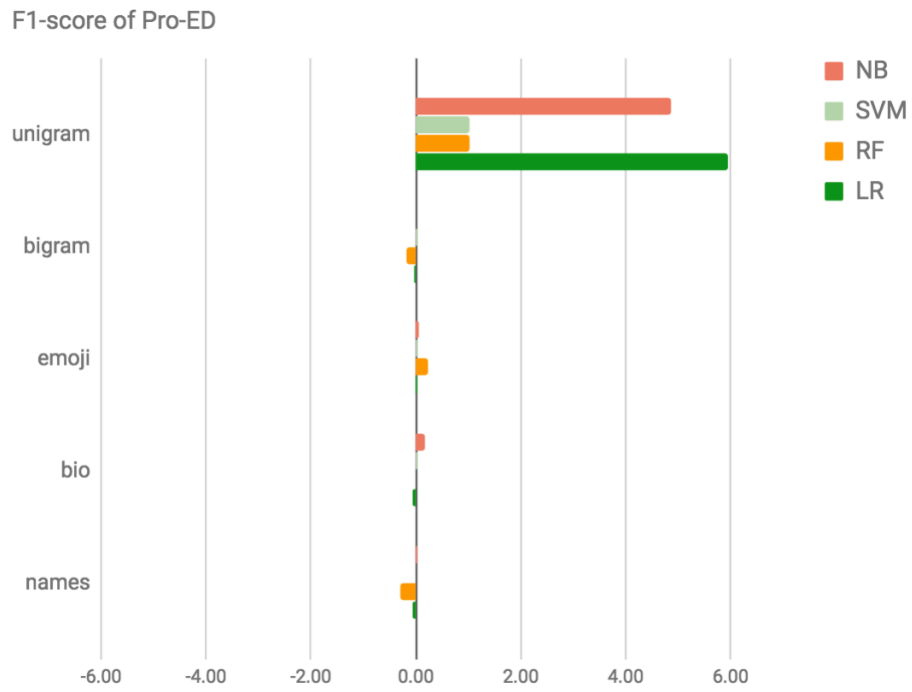
The performance scores for these models were overall higher than any of the models trained on single feature groups. The differences were not big; however, this is to be expected with this high level of accuracy.

In hope of improving the scores even further, the contribution from each feature group was calculated in order to weight each feature group according to their influence on the performance. The contribution from each feature group was assessed using a feature ablation study, where each group was removed from the total feature collection, one at a time, with equal weighting among the remaining feature groups.

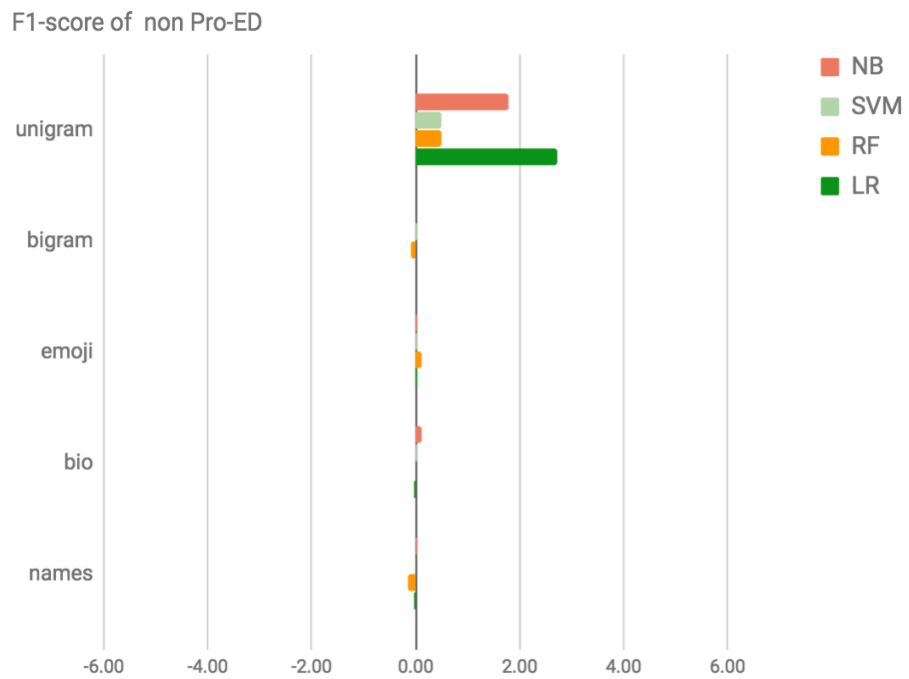
Figure 6.1 and 6.2 show the contribution from each feature group for the classification of pro-ED and non pro-ED users, respectively.

## 6. Experiments & Results

**Figure 6.1:** Contribution to the  $F_1$ -score of each of the four models for the pro-ED class



**Figure 6.2:** Contribution to the  $F_1$ -score of each of the four models for the non pro-ED class



The contribution is measured as the negative change in average  $F_1$ -score when the feature group is left out. E.g., the average  $F_1$ -score of the logistic regression model trained on all features except the ‘unigrams from tweets’ feature group dropped approximately 0.06 compared to the  $F_1$ -score of the same model trained on all feature groups. The unigram feature group’s contribution is therefore illustrated as positive 6 for the LR model in the diagram for the pro-ED class (figure 6.1).

Overall, the changes in average  $F_1$ -score were clearly largest for the unigram feature group; however, the changes were moderate and no feature groups could be said to be crucial in the identification of pro-ED users, as all models trained on all feature groups except unigrams still achieved an  $F_1$ -score of at least 0.90.

Each of the four learning algorithms was then used to train models on the collection of feature groups again, this time weighted according to their found contribution for the given algorithm. In cases where the inclusion of features contributed negative to the  $F_1$ -score, the feature group was simply ignored. Table 6.17 displays the results of classification after employing the found weighting scheme. The results did increase slightly for the NB, RF and LR approaches.

**Table 6.17:** Classification results based on 5-fold cross-validation

Features	Model	Pro-ED			Non pro-ED		
		Precision	Recall	$F_1$ -score	Precision	Recall	$F_1$ -score
Weighted Unigram + Bigram + Emoji + Bio + Name	NB	0.924	<b>0.994</b>	0.958	0.997	0.958	0.977
	SVM	0.984	0.974	0.979	0.987	0.992	0.989
	RF	<b>0.988</b>	0.971	0.980	0.985	0.994	0.990
	LR	0.985	0.976	<b>0.981</b>	0.988	0.993	0.990

These four models were finally combined using a voting classifier, **VC**, in order to wrap the models and average the predictions of the sub-models. As presented in table 6.18, the approach improved the performance score even further.

## 6. Experiments & Results

**Table 6.18:** Classification results based on 5-fold cross-validation

Features	Model	Pro-ED			Non pro-ED		
		Precision	Recall	$F_1$ -score	Precision	Recall	$F_1$ -score
Weighted feature groups	<b>VC</b>	0.989	0.977	<b>0.983</b>	0.988	0.994	0.991

This combination model achieved an extremely high precision, and had only 20 false-positives (Figure 6.3). Out of the false-positives, 12 were originally annotated as unrelated and 8 were annotated as pro-recovery. Considering the fact that there were almost 6 times more unrelated users than pro-recovery users in the training dataset, this suggests that the classifier had more difficulties in differentiating the pro-recovery users from pro-ED, than from the unrelated ones.

**Figure 6.3:** Confusion matrix for the voting classifier based on 5-fold cross-validation

		Predicted	
		Positive	Negative
True	Positive	1802	43
	Negative	20	3601

### 6.5 Test set results

In order to evaluate the classifiers' ability to detect pro-ED users in a set of unseen users, the four models trained on weighted combinations of features group, and the voting classifier (combination of these four) were tested on a set of 1376 unseen users (3.10% of the test users were removed in the high-level filtering process). The test data went through the same pre-processing steps and used the same text representation.

Table 6.19 presents the precision, recall and  $F_1$ -score for the five systems. Overall, compared to the results from cross-validation on the training set, the classifiers maintained their remarkably good performance, with a slightly higher precision and

lower  $F_1$ -score. For the combination model, only four users were falsely identified as pro-ED, out of which two were originally annotated as pro-recovery.

**Table 6.19:** Classification results on unseen test data

Feature	Model	Pro-ED			Non pro-ED		
		Precision	Recall	$F_1$ -score	Precision	Recall	$F_1$ -score
Weighted feature groups	NB	0.924	<b>0.973</b>	0.948	0.987	0.961	0.974
	SVM	<b>0.991</b>	0.969	<b>0.980</b>	0.985	0.996	0.990
	RF	<b>0.991</b>	0.935	0.962	0.970	0.996	0.982
	LR	0.984	0.962	0.973	0.982	0.992	0.987
	VC	<b>0.991</b>	0.969	<b>0.980</b>	0.985	0.996	0.990

**Figure 6.4:** Confusion matrix for the voting classifier on the test set

		Predicted	
		Positive	Negative
True	Positive	433	14
	Negative	4	925

6. Experiments & Results

## 7. Discussion & Conclusion

This chapter evaluates to what degree the goal of this study has been accomplished, and addresses the research questions formulated in the first chapter. It also discusses ethical consideration related to this study's line of work, limitations and recommendations for future work.

### 7.1 Evaluation of results

The preceding chapter presented the results from training different machine learners on single, and combinations of, feature groups for the purpose of detecting pro-ED users. The final evaluations on unseen test data supported the remarkably high performance scores observed using cross-validations on the training set in the prior experiments. The highest precision and  $F_1$ -scores were achieved using either a support vector machine, or a combination approach, trained on weighted feature groups with emphasis on unigrams from tweets.

The weight coefficients for the support vector machines demonstrated that the top most discriminating features within each feature group were usually related to body image or weight. The word «skinny», or fragments of it, were among the top ten most predictable features for all groups, except emoji. As shown in section 4.7.6, there were large differences in the amount of references to eating disorders, or related topics, in tweets written by users belonging to each of the two classes. 21.59% of the tweets from pro-ED users included a word from this study's defined reference codebook<sup>1</sup>. By aggregating and adjusting for the class imbalance, the same static for non pro-ED users were 3.04%. Moreover, references to community events and body concepts, such as «thigh gap», were hardly ever mentioned in tweets from users outside the pro-ED community. As each user in the data set included on average about 1500 tweets, these differences were likely

---

<sup>1</sup> Not all of these words were considered features since they were used in the data collection.

## 7. Discussion & Conclusion

constituting an essential part of the classifiers' ability to predict pro-ED affiliation with such high accuracies.

Although much attention was given to the approaches combining learning algorithms and weighted feature groups, it is worth pointing out that some of the single models trained on single feature groups, such as a linear SVM trained on tweet unigram features, produced good and comparable results.

### 7.2 Evaluation of Research Questions

The goal of this thesis was to construct a system for detection of pro-eating disorder users on the microblogging site Twitter. In order to reach the stated goal, three research questions were formulated. This section will address each questions and present the main findings.

**RQ1:** *How is the Twitter platform used by members of pro-ED communities, and what criteria should be used in annotation of such users?*

As the research related to pro-eating disorder communities on Twitter turned out to be limited, literature regarding the pro-ED phenomenon on social media in general was reviewed. Most studies, across social media platforms, reported similar observations of how pro-ED members employ specific community hashtags, such as #proana, and clear lines were drawn between thinspiration and the online pro-eating disorder communities. Some of these hashtags and terms were later used in searches during the collection of the data set. Arseniev-Koehler *et al.* (2016) presented criteria to evaluate whether tweets displayed a positive pro-ED attitude, and these were used to formulate the annotation guidelines used in this study.

**RQ2:** *What does previous research establish as useful methods and features for classification of user generated, textual data with respect to mental health or online subcultures?*

With regard to the second research question, most related studies reported the best achievement in performance by training classifiers on unigram features from tweets, which this work's experimental results support.



Support vector machines were the most employed machine learning algorithms used in the collection of related work, and did indeed perform best on the test data set in this study; however, the improvement over the alternative models was, for the most part, quite small.

**RQ3:** *What characterises the tweets and profile information of users taking part in pro-ED communities on Twitter?*

Both the study of characteristics in chapter 4 and the SVM weight coefficients reveal differentiating traits in both tweets and profile information of the users labelled as pro-ED. This work found that tweets written by pro-ED users included URLs and user mentions to a considerably lesser degree than the tweets written by users annotated as either unrelated or pro-recovery. Differences in usage of certain emojis were also confirmed. By designing a codebook of words related to eating disorders and similar topics, inspired by the approach of Arseniev-Koehler *et al.* (2016), tweets and profile information were checked for use of codebook words. 21.59% of the tweets from pro-ED users included a word from this codebook, compared to 3.04% of tweets written by a non pro-ED user. The results also demonstrated that tweets written by pro-recovery users included explicit references to eating disorders, such as «eating disorder» or «anorexia», more than three times more often than the pro-ED tweets.

In the data found at user-level, the biggest differences between pro-ED and non pro-ED users were shown in the amount of weight references in the profile biographies and references to body images in names. 48.27% of the pro-ED bios included a weight reference, as opposed to 0.58% of the non pro-ED biographies in the dataset. For body image references in display- and usernames the percentages were respectively 43.17% against 0.08%.

**RQ4:** *What methods and features are useful for the classification of pro-eating disorder users on Twitter.*

All four machine learning algorithms that were used in the experiments achieved high performance scores; however, the Naïve Bayes approach had a tendency to score less accurately than the other three, especially for the approaches where the features groups

## *7. Discussion & Conclusion*

were combined. The voting classifier scored best on the experiment evaluated with cross-validation on training data; however, it lost its superiority when applied to the test data set. Out of the five models, the SVM was the only model to maintain its performance scores on the unseen test data.

The study of feature groups found that unigram and bigram features extracted from tweets caused the highest performance scores. In combinations with the other feature groups, unigrams had the highest contribution according to the ablation study. For classifiers trained on single features groups, the features extracted from profile information resulted in less precise classifiers than those trained on unigram and bigram features extracted from tweets.

### **7.3 Ethical Considerations**

Digitalisation has changed the way information can be analysed, and along with the availability of social media data, it brings a variety of research opportunities, but also some ethical concerns. As Conway & O'Connor (2016) state: «...simply because social media is public, and in some cases freely available, it does not follow that it is always ethically appropriate to use it for any research purpose, particularly in relation to sensitive domains such as mental health». Inference on potentially stigmatising labels, such as mental illnesses, drug abuse, etc., at user-level should always be conducted with caution.

In particular, the moral guidelines regarding exploitation of user data and consent, in the age of social media, are blurry. Most of the pro-ED user data employed in this study came from users self-identifying as pro-eating disorder. However, some users were annotated as pro-ED despite stating otherwise. The fact that these users might oppose to being classified as pro-ED raises the question of whether it is ethically flawed to utilise their data as examples of something they do not self-identify as, without them knowing. This work considered it fair to interpret the act of sharing thinspiration, pro-ED events or desires for extreme weight loss as a confirmation of affiliation with some sort of pro-ED community.

As touched upon in the introductory chapter, moderation of pro-eating disorder content is a debated topic. Some claim the communities contribute to eating disorders, and argue that the social media platforms should censor them, or have a legal obligation to do so. On the other hand, people argue that moderation threatens the freedom of speech. The full impact of pro-ED content, and whether the communities represent a serious threat to vulnerable people on social media, is not known. Still, Twitter's young demographic and the nature of the content gives many people enough reasons to worry.

As opposed to other unwanted content in social media, such as pornography or spam, the pro-ED content does not originate from money-making industries, but often from adolescents suffering disordered eating. Besides moderation, detection of pro-ED users could provide opportunities to reach and advice individuals with eating disorders, or those heading down a dangerous path, to seek help. It is outside the scope of this project to draw conclusions on how to best handle the pro-ED related issues in today's social media, but the author hopes this line of work could open up possibilities for designing good solutions for both the pro-ED users and the general public.

### **7.4 Future work and Limitations of the study**

This study has some limitations with regard to the data set. First, the pro-ED users were found mostly through searches on well-known pro-ED hashtags, such as #proana. The returned tweets and users were then assessed against the inclusion criteria of having a positive pro-ED attitude. Thus, most of the users labeled as pro-ED had at some point employed an explicit pro-ED hashtag. This methods fail to capture users who meet the inclusion criteria, without ever employing these very pro-ED-specific tokens. Although the search words were overlooked during the training of the classifiers, it is likely that these users are more dedicated to the pro-ED community, which might amplify their differences to the non pro-ED users.

While the control set of unrelated users was diverse and included organisations, companies and people of different age, background and gender, the set of pro-ED users was more likely to consist of teenage girls as they tend to represent the main clientele of pro-eating disorder communities (Boniel-Nissim & Latzer, 2016, p.161; Giles, 2006). A demographically matched control set, as seen used in the CLPsych shared task

## 7. Discussion & Conclusion

(Coppersmith *et al.* 2015), could have given a more precise picture of the classifiers' ability to capture the users' affinities to pro-ED communities.

These concerns could be solved in further work by employing a different data collection scheme, and estimate the age and gender of the pro-ED users in order to construct a demographically matched control set. Future studies could also work towards detection of tweets containing pro-ED content, as opposed to users. Although this will cause substantially less data to base the predictions on, the found differences between the classes presented in this thesis give reason to believe an efficient classification at tweet level is achievable.

Another interesting approach would be to consider the users who are following the «explicit» pro-ED users on Twitter, as perhaps these users are less outspoken about the pro-ED communities, while still sharing many similarities. As eating disorders carry a social stigma, further research should honour privacy and evaluate the ethical challenges that arise as the classification target differentiates further from the user's own perception of self.

## *7. Discussion & Conclusion*

## *7. Discussion & Conclusion*

## References:

- Arseniev-Koehler, A. Lee, H. McCormick, T. & Moreno, M.A. (2016) #Proana: Pro-Eating Disorder Socialization on Twitter. *Journal of Adolescent Health*, 58 (6), Pages: 659 - 664.
- Balasuriya, L, Wijeratne, S. Doran, D. & Sheth, A. (2016) Finding Street Gang Members on Twitter, *International Conference on Advances in Social Networks Analysis and Mining*, 8, Pages: 685–692.
- Bardone-Cone, A. & Cass, K.M (2006) Investigating the impact of pro-anorexia websites: A pilot Study. *European Eating Disorders Review*, 14(4), Pages: 256 - 262.
- Bardone-Cone, A. & Cass, K.M. (2007) What does viewing a pro-anorexia website do? An experimental examination of website exposure and moderating effects. *International Journal of Eating Disorders*, 40(4), Pages: 537 - 548.
- Bates, C.F (2015) "I Am a Waste of Breath, of Space, of Time": Metaphors of Self in a Pro-Anorexia Group, *Qualitative Health Research*, 25(2), Pages: 189-204.
- Bellows, B.K. LaFleur, J. Kamauu, A.W. Ginter, T. Forbush, T.B., Agbor, S. Supina, D. Hodgkins, P. & DuVall, S.L. (2014) Automated identification of patients with a diagnosis of binge eating disorder from narrative electronic health records. *Journal of the American Medical Informatics Association*, 21(1) Pages: 163–168.
- Bert, F. Gualano, M.R Camussi, E. & Siliquini, R. (2016) Risks and Threats of Social Media Websites: Twitter and the Proana Movement. *Cyberpsychology, Behavior, and Social Networking*, 19(4), Pages: 233-238.
- Betton, V. Borschmann, R. Docherty, M. Coleman, S. Brown, M. & Henderson, C. (2015) The role of social media in reducing stigma and discrimination. *British Journal of Psychiatry*, 206(6), Pages: 443-444.

## References

- Boepple, L. & Thompson, J.K. (2015) A Content Analytic Comparison of Fitspiration and Thinspiration Websites. *International Journal of Eating Disorders*, 49 (1), Pages: 98-101.
- Boniell-Nissim, M, Latzer, Y. (2016) The Characteristics of Pro-Ana Community, *Bio-Psycho-Social Contributions to Understanding Eating Disorders*, Springer, Pages: 155-167.
- Borzekowski, D. Schenk, S. Wilson, J. & Peebles, R. (2010) E-Ana and E-Mia: A Content Analysis of Pro-Eating Disorder Web Sites. *American Journal of Public Health*, 100(8), Pages: 1526- 1534.
- Burnap, P. Colombo, W. & Scourfield, J. (2015) Machine Classification and Analysis of Suicide-Related Communication on Twitter. *The 26th ACM Conference, Conference Paper*, Pages: 75-84.
- Carr, C. (2017) Social Media and Intergroup Communication. *Oxfords Research Encyclopedias* [Online], August, Available at: < <http://communication.oxfordre.com/view/10.1093/acrefore/9780190228613.001.0001/acrefore-9780190228613-e-460?result=25&rskey=L6sLIM#acrefore-9780190228613-e-460-bibItem-0027> > [Accessed: 18. May 2018]
- Casilli, A. Pailler, F. & Tubaro, P. (2013) Online networks of eating-disorder websites: why censoring pro-ana might be a bad idea. *Perspectives in public health*, 133(2), Pages: 1-2.
- Chancellor, S. Pater, J.A. Clear, T.A. Gilbert, E. & De Choudhury, M. (2016) #thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, Pages: 1199-1211.
- Chancellor, S. Kalantidis, Y. Pater, J.A. De Choudhury, M. Shamma, D.A (2017) Multimodal Classification of Moderate Online Pro-Eating Disorder Content. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, Pages: 3213-3226.



- Cohen, J. (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), Pages. 37–46.
- Conway, M. & O'Connor, D. (2016) Social media, big data, and mental health: Current advances and ethical implications, *Current Opinion in Psychology*, 9, Pages: 77-82.
- Coppersmith, G. Dredze, M. Harman, C. Hollingshead, K. & Mitchell, M. (2015) CLPsych 2015 Shared Task: Depression and PTSD on Twitter, *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, Pages: 31-39.
- Fleiss, J. L. (1971) Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), Pages: 378–382.
- Ghaznavi, J. & Taylor, L.D. (2015) Bones, body parts, and sex appeal: An analysis of #thinspiration images on popular social media. *Body Image*, 14, Pages: 54-61.
- Giles, D. (2006) Constructing identities in cyberspace: The case of eating disorders. *The British journal of social psychology* 45(3) Pages: 463-477.
- Gimpel, K. Schneider, N. O'Connor, B. Das, D. Mills, D. Eisenstein, J. Heilman, M. Yogatama, D. Flanigan, J. & Smith, N.A. (2011) Part-of-speech tagging for twitter: Annotation, features, and experiments. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Pages: 42–47.
- Guyon, I. (2003) An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, (3), Pages: 1157-1182
- Hasan, H. (2012) Instagram Bans Thinspo Content. *Time* [Online], 26. April, Available at: <<http://newsfeed.time.com/2012/04/26/instagram-bans-thinspo-content>> [Accessed 10. May 2018]
- Holahan, C. (2001) Yahoo removes pro-eating disorder internet sites. *Boston Globe*, 4 August, Page: A2.

## References

- Homan, C.M. Johar, R. Tong, L. Lytle, M. Silenzio, V. & Cecilia, O.A. (2014) Toward Macro-Insights for Suicide Prevention: Analyzing Fine-Grained Distress at Scale. *Workshop on Computational Linguistics and Clinical Psychology Conference Paper*, Pages: 107-117.
- Jackson, R.G. Patel, R. Jayatilleke, N. Kolliakou, A. Ball, M. Gorrell, G. Roberts, A. Dobson, R.J. & Stewart R. (2017) Natural language processing to extract symptoms of severe mental illness from clinical text: the Clinical Record Interactive Search Comprehensive Data Extraction. *BMJ Open*, 7.
- Johnson, H.A. (2015) I Will Not Eat! A Review of the Online Pro-Ana Movement.
- Mowery, D. Smith, H.A. Cheney, T. Bryan, C. and Conway, M. (2015) Identifying Depression-Related Tweets from Twitter for Public Health Monitoring, *ISDS 2015 Conference*
- Mowery, D. Bryan, C. & Conway, M. (2017) Feature Studies to Inform the Classification of Depressive Symptoms from Twitter Data for Population Health.
- National Health Service (2015) How many calories do teenagers need. *NHS* [Online] 16. June, Available at: <<https://www.nhs.uk/chq/Pages/how-many-calories-do-teenagers-need.aspx>> [Accessed 26. May 2018]
- National Institute of Mental Health (2016) Health and Education: Eating Disorders, *NIMH*[Online] February, Available at: <<https://www.nimh.nih.gov/health/topics/eating-disorders/index.shtml>>[Accessed at: 10. May 2018]
- Novak, P.K, Smailović, J. Sluban, B. & Mozetič, I. (2015) Sentiment of Emojis. *PloS one*, 10(12)
- Pails, C. (2012) 'Pinterest Terms of Service Get Updated', *Huffington Post* [Online], 26.March, Available at: <[https://www.huffingtonpost.com/2012/03/26/pinterest-terms-of-service-update\\_n\\_1379486.html?1332780457](https://www.huffingtonpost.com/2012/03/26/pinterest-terms-of-service-update_n_1379486.html?1332780457)> [Accessed 10. May 2018]

- Pennacchiotti, M. & Popescu, A.-M. (2011) A machine learning approach to twitter user classification. *Proceedings of the Fifth International Conference on Weblogs and Social Media*, 11.
- Preoțiu-Pietro, D. Sap, M. Schwartz, H. A. & Ungar, L. H. (2015) Mental Illness Detection at the World Well-Being Project for the CLPsych 2015 Shared Task. *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, Pages: 40-45.
- Pedersen, T. (2015) Screening Twitter users for depression and PTSD with lexical decision lists. *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, Pages: 46-53.
- Pedregosa, F. Varoquaux, G. Gramfort, A. Michel, V. & Thirion, B. (2011) 'Scikit-learn: Machine Learning in Python.' , *Journal of Machine Learning Research*. 12, Pages: 2825-2830
- Raileanu, L.E & Stoffel, K. (2004) Theoretical comparison between the Gini Index and Information Gain criteria. *Annals of Mathematics and Artificial Intelligence*, 41, Pages: 77-93
- Ransom, D.C. La Guardia, J.G. Woody, E.Z. & Boyd, J.L. (2009) 'Interpersonal Interactions on Online Forums Addressing Eating Concerns.', *The International Journal of Eating Disorders*. 43(2), Pages: 161-170.
- Rao, D. Yarowsky, D. Shreevats, A. & Gupta, M. (2010) Classifying Latent User Attributes in Twitter. *International Conference on Information and Knowledge Management, Proceedings*, Pages: 37-44.
- Reel, J.J. (2013) '*Eating Disorders: An Encyclopedia of Causes, Treatment and Prevention*' Greenwood.
- Resnik, P. Armstrong, W. Claudino, L. Nguyen, T. Nguyen, V. & BoydGraber, J. (2015) The University of Maryland CLPsych 2015 shared task system. *Proceedings of the*

## References

- Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, NAACL.
- Rosen A. (Product Manager, Twitter) (2017) 'Tweeting Made Easier', *Twitter* [Online] 7.November, Available at: < [https://blog.twitter.com/official/en\\_us/topics/product/2017/tweetingmadeeasier.html](https://blog.twitter.com/official/en_us/topics/product/2017/tweetingmadeeasier.html) >[Accessed 11. December 2017]
- Russell, S. & Norvig P. (2010) *Artificial Intelligence: A Modern Approach 3rd edition*, Pearson Education.
- Saul, H. (2015) People running pro-anorexiaand 'thinspiration' websites to face prison in France, *The Independent* [Online], 3.April, Available at: <<https://www.independent.co.uk/life-style/health-and-families/health-news/people-running-thinspiration-or-pro-anorexia-websites-in-france-will-now-face-a-prison-sentence-and-10153909.html>> [Accessed 23. March 2018]
- Scikit Learn Developers (2017) Documentation of scikit-learn 0.19.1 [Online], Available at: <<http://scikit-learn.org/stable/documentation.html>> [Accessed 31. May 2018]
- Stewart I. Chancellor S. De Choudhury M. & Eisenstein J. (2017) #anorexia, #anarexia, #anarexyia: Characterizing Online Community Practices with Orthographic Variation. *IEEE International Conference on Big Data*, Pages: 4353-4361.
- Talbot T.S. (2010) The effects of viewing pro-eating disorder websites: a systematic review. *West Indian Medical Journal*, 59(6), Pages: 686-697.
- Talbot, C.A. Gavin, J. van Steen, T. & Morey, Y. (2017) A content analysis of thinspiration, fitspiration, and bonespiration imagery on social media, *Journal of Eating Disorders*, 5 (1), Pages: 5 - 40.
- Taylor, A. Marcus, M. Santorini, B. (2003) The Penn Treebank: An Overview, *Treebanks*, TLTB 20, Springer, Pages: 5-22.

- Tran, T. & Kavuluru, R. (2017) Predicting Mental Conditions Based on «History of Present Illness» Psychiatric Notes with Deep Neural Networks. *Journal of Biomedical Informatics*, 75, Pages: 138 - 148
- Tumblr (2012), 'A new policy against self-harm blogs', *Tumblr Staff* [Online], 23. February, Available at: <<https://staff.tumblr.com/post/18132624829/self-harm-blogs>> [Accessed 23. March 2018]
- Twitter (2018) 'Q1 2018 Letter to Shareholders', *Twitter Investor Relations* [Online], 25. April, Available at: <[http://files.shareholder.com/downloads/AMDA-2F526X/6220513289x0x978181/2FD6D58F-A930-4EB2-90B0-9C3A120648DE/Q1\\_2018\\_Shareholder\\_Letter.pdf](http://files.shareholder.com/downloads/AMDA-2F526X/6220513289x0x978181/2FD6D58F-A930-4EB2-90B0-9C3A120648DE/Q1_2018_Shareholder_Letter.pdf)> [Accessed 20. May 2018]
- Wilson, J.L. Peebles, R. Hardy, K.K. & Litt, I.F. (2016) «Surfing for thinness: a pilot study of pro- eating disorder Web site usage in adolescents with eating disorders», *Pediatrics*, 118(6), Pages: 1635-1643.
- World Health Organization (2018) BMI classification, *WHO* [Online] June, Available at: <[http://apps.who.int/bmi/index.jsp?introPage=intro\\_3.html](http://apps.who.int/bmi/index.jsp?introPage=intro_3.html)> [Accessed at: 7. June 2018]
- Yazdavar, A. Al-Olimat, H. Ebrahimi, M. Bajaj, G. Banerjee, T. Thirunarayan, K. Pathak, J. & Sheth, A. (2017) Semi-Supervised Approach to Monitoring Clinical Depressive Symptoms in Social Media. *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, Pages: 1191-1198.
- Yeshua-Katz D. & Martins N. (2013) Communication Stigma: The Pro-Ana Paradox. *Health Communication*, 28(5), Pages: 499-508.
- Yom-Tov, E. Fernandes-Luque, L. Weber, I. & Crain, S.P. (2012) Pro-Anorexia and Pro-Recovery Photo Sharing: A Tale of Two Warring Tribes. *Journal of Medical Internet Research* 14(6), Pages: 39-50

## *References*

## Image references:

'em dem', Flickr user (2016) *Stockings* [Online photography licensed under CC BY 2.0, Heavily modified from original]. Available at: < <https://www.flickr.com/photos/emden09/26528766003/in/photolist-GqfRTi> > [Accessed 13. April 2018].

Instagram (2018) Content Advisory [Online Search Results ©2018 Instagram]. Available at: < <https://www.instagram.com/explore/tags/anorexia/> > [Accessed 6. March 2018]. Screenshot by author.

Jlo, L. (2010) Everybody Hurts [Online photography licensed under CC BY 2.0, Desaturated from original]. Available at: < <https://www.flickr.com/photos/37167854@N08/4287444249/> > [Accessed 13. April 2018].

Macsurak, C. (2014) *Blanca Padilla walking the runway for Porsche Design* [Online photography licensed under CC BY 2.0, Desaturated from original]. Available at: <[https://en.wikipedia.org/wiki/File:Blanca\\_Padilla\\_Porsche\\_Design\\_2.jpg](https://en.wikipedia.org/wiki/File:Blanca_Padilla_Porsche_Design_2.jpg)> [Accessed 13. April 2018].

Pinterest (2018) Search Results for «Anorexia» [Online Search Results ©2018 Pinterest]. Available at: < [https://pinterest.com/search/pins/?q=anorexia&rs=rs&eq=&etslf=1965&term\\_meta\[\]=anorexia%7Crecentsearch%7Cundefined](https://pinterest.com/search/pins/?q=anorexia&rs=rs&eq=&etslf=1965&term_meta[]=anorexia%7Crecentsearch%7Cundefined) > [Accessed 6. March 2018]. Screenshot by author.

Tumblr (2018) Public Service Announcement [Online Search Results ©2018 Tumblr]. Available at: < <https://www.tumblr.com/psa/search/anorexia> > [Accessed 6. March 2018]. Screenshot by author.

Twitter (2018) *Twemoji* [graphics licensed under CC-BY 4.0] Available at: < <https://github.com/twitter/twemoji> > [Accessed 13. April 2018].

## *References*



## Appendix A: Pro-Eating Disorder

This chapter includes explanations of pro-ED related hashtags and terms referred to in this thesis. Table A.1 contains words related to body weight and eating disorders in general, while table A.2 presents hashtags and terms used by the pro-ED communities.

**Table A.1:** Medical terms and Diagnoses

<b>Term/Phrase:</b>	<b>Explanation:</b>
Anorexia Nervosa (Anorexia)	Anorexia nervosa is an eating disorder characterised by weight loss, and for many, a distorted body image (National Eating Disorders Association, 2018).
Bulimia Nervosa (Bulimia)	Bulimia nervosa is an eating disorder characterised by cycles of bingeing and compensatory behaviours (National Eating Disorders Association, 2018).
BMI	Body Mass Index (BMI) is an index of weight-for-height, commonly used to classify underweight, overweight and obesity in adults (World Health Organization, 2018).
ED	Short for Eating Disorders.
EDNOS	Eating Disorders Not Otherwise Specified is the former name of OSFED (National Eating Disorders Association, 2018).
OSFED	Other Specified Feeding or Eating Disorder is a category of eating disorders that do not meet the criteria for any other specific eating disorder diagnosis (National Eating Disorders Association, 2018).

**Table A.2:** Explanations of Pro-ED terms and phrases

<b>Term/Phrase:</b>	<b>Explanation:</b>
ABCDiet	Extreme diet.
Ana	Ana can reference both the disorder Anorexia nervosa and the pro-ana community. Often seen personified.
Ana buddy/Ana coach	Members of the pro-ana community encouraging each other to lose weight.

## Appendices

BikiniBridge	Having a space between bikini bottom and the lower abdomen, when bikini bottoms are suspended between the two protruding hip bones.
Binge	Episode of uncontrollable eating.
Bonespo / Bonespiration	Sub-type of thinspiration. Promoting the desirability of a skeletal appearance (Johnson, 2015).
CalorieApril	Pro-ED event.
CW	Current weight.
GW	Goal weight.
Laxies	Short for laxatives.
LW	Lowest (achieved) weight.
Mia	Mia can reference both the eating disorder, Bulimia nervosa, and the pro-mia community. Often seen personified.
Meanspo	Mean or strict content intended to promote weight loss.
Pro-ana	Pro-Anorexia (pro-eating disorder community).
Pro-mia	Pro-Bulimia (pro-eating disorder community).
Pro-Eating Disorder / Pro-ED	Pro-eating disorder refers to online communities endorsing engagement with disordered eating.
ProjectThin	Extreme diet.
Purge	To rid of whatever is impure or undesirable, often used in the pro-ed communities to references a compensatory behaviour such as self-induced vomiting, extreme exercise or misuse of laxatives.
Reversethinspo	Used to describe photos or content that does not fulfil the criteria of «thinspiration».
RG/Russian Gymnast	Extreme diet.
Skinnyforsummer	Pro-ED event.
Skinny4Xmas/ Skinny4Christmas	Pro-ED event.
Skinny4NewYear	Pro-ED event.
Thigh gap	Having a space between the inner thighs while standing upright.
Thinspo / Thinspiration	Content (words and images) intended to promote weight loss (Johnson, 2015).
UGW	Ultimate Goal Weight.
Wannarexic	Wannarexic is used in Pro-ED communities to describe individuals faking engagement in ED behaviour (Arseniev-Koehler <i>et al.</i> 2016).

## Appendix B: Data

### B.1 Data Collection

In the process of gathering data, the Tweepy<sup>1</sup> library was used. Tweepy is an open-source wrapper which provides easy access to the Twitter API (section 2.5.1) for the programming language Python, and is licensed under the MIT license<sup>2</sup>.

The user-scraper code was based on the following two Github codes:

- <https://gist.github.com/yanofsky/5436496>
- <https://gist.github.com/macloo/5c69cdf5294fa97eb41d6ad950233cee>

### B.2 Keywords for Data Collection

In the process of collecting data for the purpose of this study the following sampling tag words were used in searches on Twitter.

**Table B.1:** Sampling tag words

Domain:	keywords
Pro-ED	#anabuddy, #anacoach, #bonespiration, #bonespo, #calorieapril #meanspo, #reversethinspo, #proana, #proed, #promia, #skinny4xmas, #skinny4-christmas, #thinspiration, #thinspo, #ugw
Pro-Recovery	#BeatED, #EatingDisorderRecovery, Eating Disorder Recovery, #EDRecovery, #RecoveryWarriors, #effyourbeautystandards, #bodypositivity
Unrelated	#AcademyAwards, #Arsenal, #Audi, #Avengers, #Baseball, #Bernie, #Biology, #Brexit, Business Insider, Casey Neighstat #ball, #christmas, #college, #coke, DIY, #Eid, #Eurovision, #fact, #fashion, #fitness, #funny, #FreePalestine, Hadoop, #health, Jimmy Fallon, John Lewis, Legend of Zelda, #life, #lol, #love, Mona Lisa, Museum of Modern Art, National Geographic, #namaste , NFL, #nutrition, pegida, PyeongChang, Snapchat, Stand up to Cancer, Taylor Swift, The Breakfast Club, #today, Trump, video, #vlog, weed, #wedding, youtube

<sup>1</sup> Tweepy Documentation:  
<http://docs.tweepy.org/en/v3.5.0/index.html>

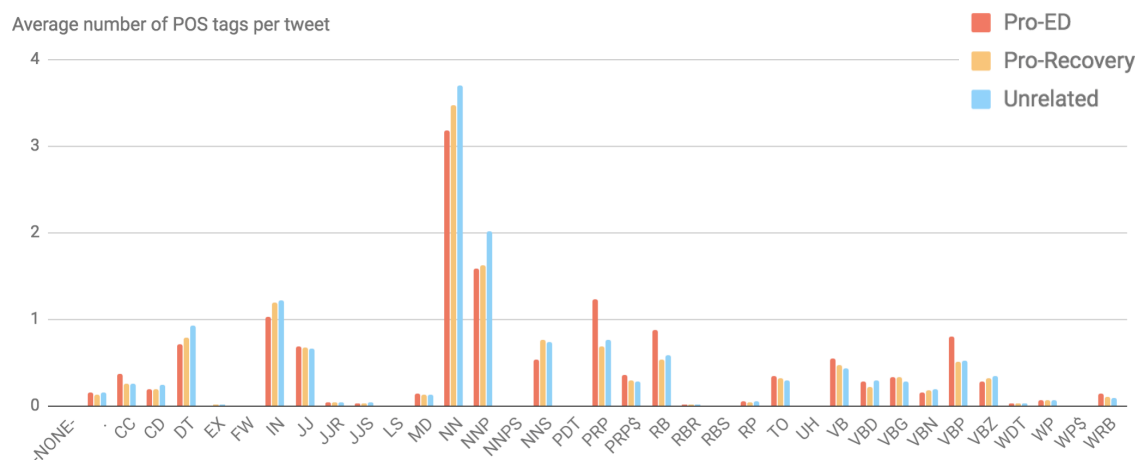
<sup>2</sup> MIT License:  
<https://opensource.org/licenses/MIT>

### B.3 Part-of-Speech

The following list contains all tags with corresponding part of speech, in alphabetical order (Taylor, Marcus & Santorini, 2003, page: 8). Figure B.1 displays the complete histogram over average number of each part-of-speech-tag in all tweets in the training data set.

CC	Coordinating conjunction	PP\$	Possessive pronoun
CD	Cardinal number	RB	Adverb
DT	Determiner	RBR	Adverb, comparative
EX	Existential <i>there</i>	RBS	Adverb, superlative
FW	Foreign word	RP	Particle
IN	Preposition	SYM	Symbol
JJ	Adjective	TO	infinitival to
JJR	Adjective, comparative	UH	Interjection
JJS	Adjective, superlative	VB	Verb, base form
LS	List item marker	VBD	Verb, past tense
MD	Modal	VBG	Verb, gerund/present participle
NN	Noun, singular or mass	VBN	Verb, past participle
NNS	Noun, plural	VBP	Verb, non-3rd person singular present
NNP	Proper noun, singular	VBZ	Verb, 3rd person singular present
NNPS	Proper noun, plural	WDT	Wh-determiner
PDT	Predeterminer	WP	Wh-pronoun
POS	Possessive ending	WP\$	Possessive wh-pronoun
PRP	Personal pronoun	WRB	Wh-adverb

**Figure B.1:** Complete histogram over POS-tags



## B.4 Most Popular Locations:

In the study of locations, presented in section 4.8.3, the location field of all the users in the training data set was examined. Table B.2 presents the top 15 locations used by the highest percentages of users.

**Table B.2** : Popular locations

Location	Total	Pro-ED	Pro-Recovery	Unrelated
<b>NULL</b>	35.4 %	52.5 %	18.2 %	28.4 %
USA	3,42 %	1,73 %	3,86 %	<b>4,37 %</b>
England	3,20 %	1,41 %	<b>6,43 %</b>	3,64 %
CA	2,98 %	0,54 %	7,72 %	3,50 %
London	2,38 %	1,03 %	2,09 %	3,27 %
NY	1,99 %	0,27 %	4,34 %	2,57 %
Los Angeles	1,94 %	0,16 %	2,57 %	2,90 %
UK	1,61 %	0,43 %	2,25 %	2,20 %
New York	1,56 %	0,38 %	3,70 %	1,83 %
United Stated	1,45 %	2,06 %	1,29 %	1,10 %
United Kingdom	1,43 %	1,46 %	2,09 %	1,27 %
Canada	1,26 %	1,03 %	2,25 %	1,20 %
India	0,80 %	0,05 %	0,80 %	1,27 %
California	0,77 %	0,22 %	0,96 %	1,07 %
Hell	0,79 %	<b>2,22 %</b>	0,00 %	0,07 %
Australia	0,66 %	0,43 %	1,93 %	0,53 %

## *Appendices*

## Appendix C: Architecture

### C.1 Stop Words

Table C.1 contains all stopwords used during the pre-processing step.

**Table C.1:** Stop words

Words from NLTK's English stop words	<p>all, six, less, being, indeed, over, move, anyway, fifty, four, not, own, through, yourselves, go, where, mill, only, find, before, one, whose, system, how, somewhere, with, thick, show, had, enough, should, to, must, whom, seeming, under, ours, has, might, thereafter, latterly, do, them, his, around, than, get, very, de, none, cannot, every, whether, they, front, during, thus, now, him, nor, name, several, hereafter, always, who, cry, whither, this, someone, either, each, become, thereupon, sometime, side, two, therein, twelve, because, often, ten, our, eg, some, back, up, namely, towards, are, further, beyond, ourselves, yet, out, even, will, what, still, for, bottom, mine, since, please, forty, per, its, everything, behind, un, above, between, it, neither, seemed, ever, across, she, somehow, be, we, full, never, sixty, however, here, otherwise, were, whereupon, nowhere, although, found, alone, re, along, fifteen, by, both, about, last, would, anything, via, many, could, thence, put, against, keep, etc, amount, became, ltd, hence, onto, or, con, among, already, co, afterwards, formerly, within, seems, into, others, while, whatever, except, down, hers, everyone, done, least, another, whoever, moreover, couldnt, throughout, anyhow, yourself, three, from, her, few, together, top, there, due, been, next, anyone, eleven, much, call, therefore, interest, then, thru, themselves, hundred, was, sincere, empty, more, himself, elsewhere, mostly, on, fire, am, becoming, hereby, amongst, else, part, everywhere, too, herself, former, those, he, me, myself, made, twenty, these, bill, cant, us, until, besides, nevertheless, below, anywhere, nine, can, of, your, toward, my, something, and, whereafter, whenever, give, almost, wherever, is, describe, beforehand, herein, an, as, itself, at, have, in, seem, whence, ie, any, fill, again, hasnt, inc, thereby, thin, no, perhaps, latter, meanwhile, when, detail, same, wherein, beside, also, that, other, take, which, becomes, you, if, nobody, see, though, may, after, upon, most, hereupon, eight, but, serious, nothing, such, why, a, off, whereby, third, i, whole, noone, sometimes, well, amoungst, yours, their, rather, without, so, five, the, first, whereas, once</p>
Words used in searches for users	<p>academyawards, arsenal, audi, avengers, baseball, bernie, biology, brexit, business insider, casey neighstat ball, christmas, college, coke, diy, eid, eurovision, fact, fashion, fitness, funny, freepalestine, hadoop, health, jimmy fallon, john lewis, legend of zelda, life, lol, love, mona lisa, museum of modern art, national geographic, namaste , nfl, nutrition, pegida, pyeongchang, snapchat, stand up to cancer, taylor swift, the breakfast club, today, trump, video, vlog, weed, wedding, youtube, anabuddy, anacoach, bonespiration, calorieapril, bonespo, meanspo, reversethinspo, proana, proed, promia, skinny4xmas, skinny4christmas, thinspiration, thinspo, ugw, beatED, eatingdisorderrecovery, eating disorder recovery, edrecovery, recoverywarriors, effyourbeautystandards, bodypositivity</p>

## C.2 Vocabulary size

Scikit-Learn lets the encoder define a maximum limit of the size of the tfidf-weighted vocabulary in its `TfidfVectorizer` class by adjusting the `max_features` parameter. When the parameter is changed from `None` (default) to  $m$ , the `TfidfVectorizer` builds a vocabulary that only considers the top  $m$  features, according to the term frequency across the document collection.

For the sake of finding an appropriate number of features to extract from each feature group, this study examined four different sizes of vocabularies per feature group<sup>3</sup>; 2000, 10,000, 20,000 and unlimited (`max_features = None`). For each size of vocabulary, the average precision, recall and  $F_1$ -score for all machine learners trained on the given type of feature were examined, with emphasis on the  $F_1$ -score. The following sections present the results from each feature group.

In cases where different vocabulary sizes yielded similar results, the smallest size of vocabulary was always preferred in order to reduce the number of dimensions to consider.

### C.2.1 Unigram Features from Tweets:

Table C.2 presents the average scores of the four models, each trained on a tf-idf weighted vocabulary of unigrams from tweets, evaluated under a 5-fold cross validation scheme.

The results demonstrated that a maximum limit of 20,000 and no limit (`None`) generated the best results. Considering the large difference in term space between these two options, this work concluded that the vocabulary of 20,000 was the better choice.

---

<sup>3</sup> With the exception of emoji features.



**Table C.2:** Unigram vocabulary size

Feature	Model	Pro-ED			Not Pro-ED		
		Precision	Recall	$F_1$ -score	Precision	Recall	$F_1$ -score
2000 Unigrams	NB	0.923	0.985	0.953	0.992	0.958	0.975
	SVM	0.973	0.972	0.973	0.986	0.986	0.986
	RF	0.973	0.972	0.973	0.986	0.986	0.986
	LR	0.977	0.976	0.976	0.988	0.988	0.988
10,000 Unigrams	NB	0.919	0.994	0.955	0.997	0.956	0.976
	SVM	0.978	0.972	0.975	0.986	0.989	0.987
	RF	0.986	0.963	0.974	0.981	0.993	0.987
	LR	0.977	0.975	0.976	0.987	0.988	0.988
20,000 Unigrams	NB	0.923	0.994	0.957	0.997	0.958	0.977
	SVM	0.982	0.974	<b>0.978</b>	0.987	0.991	0.989
	RF	0.986	0.966	<b>0.976</b>	0.983	0.993	0.988
	LR	0.981	0.974	0.977	0.987	0.990	0.989
1,087,181 Unigrams (no limit)	NB	0.931	0.994	<b>0.962</b>	0.997	0.963	0.979
	SVM	0.983	0.974	<b>0.978</b>	0.987	0.991	0.989
	RF	0.976	0.935	0.955	0.968	0.988	0.978
	LR	0.982	0.976	<b>0.979</b>	0.988	0.991	0.989

## Appendices

### C.2.2 Bigram Features from Tweets:

Table C.3 presents the average scores of the four models, each trained on a tf-idf weighted vocabulary of bigrams from tweets, evaluated under a 5-fold cross validation scheme.

The results demonstrated that a maximum limit of 10,000 and 20,000 generated comparable results. This work concluded that the vocabulary of 10,000 was the better choice, as it reduced the number of dimensions.

**Table C.3:** Bigram vocabulary size

Feature	Model	Pro-ED			Not Pro-ED		
		Precision	Recall	$F_1$ -score	Precision	Recall	$F_1$ -score
2000 Bigrams	NB	0.959	0.952	0.955	0.976	0.979	0.977
	SVM	0.985	0.949	0.967	0.975	0.993	0.984
	RF	0.981	0.938	0.959	0.969	0.991	0.980
	LR	0.969	0.895	0.931	0.949	0.985	0.967
10,000 Bigrams	NB	0.959	0.971	0.965	0.985	0.979	0.982
	SVM	0.985	0.963	<b>0.974</b>	0.981	0.993	0.987
	RF	0.980	0.953	<b>0.966</b>	0.976	0.990	0.983
	LR	0.984	0.918	<b>0.950</b>	0.960	0.992	0.976
20,000 Bigrams	NB	0.962	0.972	0.967	0.986	0.980	0.983
	SVM	0.984	0.965	<b>0.974</b>	0.982	0.992	0.987
	RF	0.979	0.946	0.962	0.973	0.990	0.981
	LR	0.984	0.918	<b>0.950</b>	0.959	0.993	0.976
16,091,600 Bigrams (no limit)	NB	0.994	0.946	<b>0.969</b>	0.973	0.997	0.985
	SVM	0.991	0.948	0.969	0.974	0.996	0.985
	RF	0.973	0.888	0.929	0.946	0.987	0.966
	LR	0.983	0.887	0.932	0.945	0.992	0.968

### C.2.3 Unigram and Bigram Features from Biographies:

Table C.4 presents the average scores of the four models, each trained on a tf-idf weighted vocabulary of unigrams and bigrams extracted from biographies, evaluated under a 5-fold cross validation scheme.

**Table C.4:** Biography features vocabulary size

Feature	Model	Pro-ED			Not Pro-ED		
		Precision	Recall	$F_1$ -score	Precision	Recall	$F_1$ -score
2000 Unigrams and Bigrams from Biographies	NB	0.914	0.747	0.823	0.882	0.964	0.921
	SVM	0.882	0.755	0.813	0.884	0.948	0.915
	RF	0.875	0.713	<b>0.786</b>	0.866	0.948	0.905
	LR	0.932	0.686	<b>0.791</b>	0.859	0.975	0.913
10,000 Unigram and Bigrams from Biographies	NB	0.941	0.733	<b>0.824</b>	0.878	0.977	0.924
	SVM	0.895	0.754	<b>0.819</b>	0.884	0.955	0.918
	RF	0.879	0.709	0.785	0.865	0.950	0.906
	LR	0.934	0.674	0.783	0.855	0.976	0.911
20,000 Unigram and Bigrams from Biographies	NB	0.940	0.701	0.803	0.865	0.977	0.918
	SVM	0.882	0.751	0.811	0.882	0.949	0.914
	RF	0.839	0.715	0.772	0.865	0.930	0.896
	LR	0.928	0.654	0.767	0.847	0.974	0.906
44,502 Unigram and Bigrams from Biographies (no limit)	NB	0.951	0.675	0.790	0.856	0.982	0.915
	SVM	0.885	0.754	0.814	0.883	0.950	0.916
	RF	0.800	0.744	0.771	0.874	0.905	0.889
	LR	0.924	0.649	0.762	0.845	0.973	0.904

**C.2.4 Character  $n$ -gram Features from Names:**

Table C.5 presents the average scores of the four models, each trained on a tf-idf weighted vocabulary of character  $n$ -grams of length 3 to 15, extracted from usernames and display names, evaluated under a 5-fold cross validation scheme.

**Table C.5:** Name features vocabulary size

Feature	Model	Pro-ED			Not Pro-ED		
		Precision	Recall	$F_1$ -score	Precision	Recall	$F_1$ -score
2000 Character $n$ -grams from Names	NB	0.915	0.582	0.711	0.820	0.972	0.890
	SVM	0.865	0.714	0.782	0.866	0.943	0.903
	RF	0.850	0.688	0.760	0.855	0.938	0.895
	LR	0.927	0.602	<b>0.730</b>	0.828	0.976	0.896
10,000 Character $n$ -grams from Names	NB	0.928	0.647	<b>0.762</b>	0.844	0.974	0.905
	SVM	0.902	0.743	0.815	0.880	0.959	0.918
	RF	0.876	0.714	<b>0.787</b>	0.867	0.949	0.906
	LR	0.958	0.580	0.723	0.822	0.987	0.897
20,000 Character $n$ -grams from Names	NB	0.936	0.623	0.748	0.836	0.978	0.902
	SVM	0.910	0.741	<b>0.817</b>	0.879	0.963	0.919
	RF	0.881	0.705	0.783	0.863	0.951	0.905
	LR	0.958	0.557	0.704	0.814	0.988	0.892
211,380 Character $n$ -grams from Names (no limit)	NB	0.956	0.499	0.656	0.795	0.988	0.881
	SVM	0.907	0.731	0.809	0.875	0.962	0.916
	RF	0.845	0.682	0.755	0.852	0.936	0.892
	LR	0.951	0.455	0.615	0.781	0.988	0.872

