USER BEHAVIOR IN SOCIAL MEDIA: ENGAGEMENT, INCIVILITY AND DEPRESSION

by

Farig Yousuf Sadeque

A Dissertation Submitted to the Faculty of the

SCHOOL OF INFORMATION

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2019

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by Farig Yousuf Sadeque, titled User Behavior in Social Media: Engagement, Incivility and Depression and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

Steven Bethard — Date: 18 March 2019

Mihai Surdeanu — Date: 18 March 2019

Yotam Shmargad — Date: 18 March 2019

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

Date: 18 March 2019

Dissertation Director: Steven Bethard
Assistant Professor
School of Information

ACKNOWLEDGEMENTS

Thank you Steve, for being the best mentor anyone can ever dream of. You have been such a magnificent guide to me throughout the years- this would not have been possible without you pointing me towards the right direction. Thanks a lot to the my co-advisors for their wonderful assistance, and thanks to the all professors who have equipped me with the knowledge that made this dissertation possible. Special thanks to my marvelous teammates: Dongfang, Vikas and Egoitz- I have learned so much from all of you. My friends- who stuck with me through thick and thin- I cannot thank you enough. Thanks to all my family members- without whom I would not be here. But the biggest thanks goes to my parents- who have instilled in me the value of education, and who always believed that I can do it. And finally, thanks to my little sister, who, despite being three years younger than me, will always be my inspiration.

This has been a wonderful journey, and thanks to everyone who shared it with me.

DEDICATION

To my parents.

Ammu and Baba, this is your achievement as much as it is mine.

TABLE OF CONTENTS

TABLE OF CONTENTS – *Continued*

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

User behavior in online social media has been a much researched topic in various fields– and although some aspects of user behavior like political orientation and online harassment have received much of the limelight, some other aspects have remained mostly obscured. In this research we are exploring three specific behavioral aspects: engagement, incivility and mental health; and our ability to predict these aspects. Predicting future engagement of users can be a behavioral research topic, where user-generated contents and activity frequencies can provide valuable insights. These attributes can be used to analyze and predict how civilly users behave in these social platforms, and can also be used to analyze mental health of a user. All three of these behavioral aspects contribute to the health of a community, and have profound influence on the social capital and the sustainability of the social media platforms.

We have built prediction models for engagement in multiple social media, and analyzed the features that we have used over a certain period of time and in a cross-platform environment. We built models for identifying incivilities from user-generated contents and used it in social media as uncivil behavior has the potential to effect user engagement in a platform. We built depression detection models from user texts, and introduced a new performance metric that can measure the quality of a prediction model based on

its observational latency- and we argue that it is a more expressive metric of an early prediction model than the current state-of-the-art. We believe we have had significant contributions in the fields we have worked on, and have published our works in various conferences and workshops.

CHAPTER 1

INTRODUCTION

One of the most prevalent uses of the Internet now-a-days is the online social networks as they have established themselves as an integral part of human interaction and communication over the last decade. The term "Online Social Network" covers a wide range of services that provide online interaction- from micro-blogging sites such as Twitter to support group based health forums such as DailyStrength. As these networks grew, a lot of research has been done on them over the years. Various aspects of user behavior have been central in these works, yet some aspects have remained quite obscure. In this research we focus on three of these behaviors which are intertwined: engagement, incivility and mental health. Our goal for this research is to observe, analyze and predict these behaviors in various forms of social media, and in this dissertation, we will present the work that has been done for this purpose.

As in any other research, we started with a set of questions that we were planning to answer in the course of this research.

Engagement

User engagement has been a much researched topic in a handful of other fields like telecommunications- but in social media it has never garnered much attraction. Engagement

can be an indicator of the health of a social media platform and thus can be used for predicting the sustainability of the said platform. When we first started this research, we asked ourselves a plethora of questions- can we predict user-level engagement in a social media platform? What features are the most important? Does the importance change over time? Are all features similarly useful in a multi-community environment? To answer these questions, we experimented with state-of-the-art machine learning models. We explored a huge number of features that can be used in these machine learning models- we looked into user-generated contents- language structure, word usage, sentiment etc.; we explored the activities and their frequencies over time, and we analyzed user and community level engagement networks as a potential feature set towards a better prediction. We analyzed the effects of these attributes in multiple social media platforms (and also in different communities in the same platform) to identify the contributions of these attributes over a user's future engagement.

Online Harassment: Incivility

While exploring the phenomenon of engagement (whether a user will participate in a community or not), we were convinced that it is important to look into how these users engage- and the question of civility comes in. We would like to identify users who do not participate in a community in a civilized manner, thus harming the community health. We asked ourselves, can we build a model that can identify user generated contents in a social media platform that display some form of incivility? We identified attributes of

user-generated text contents that is unique to uncivilized conversations for this purpose. We built prediction models and trained them to identify uncivil conversations. Our goal is to use these models as a filter that can be used along with human moderators, and will help reducing the human effort that currently goes in detecting incivility.

Mental Health: Depression

While exploring the phenomenon of engagement in social media, and how people behave within it, we figured out that there are repercussions of these activities on an individual's mental health. Although it is difficult to identify the scale of these repercussions, it is possible to identify users who are going through a difficult time. As the spectrum of mental health is vast, we focused our efforts on the most prevalent mental health problem- depression. We built prediction models that used user generated contents in a social media to predict the state of depression of the users. We used state of the art natural language processing and machine learning techniques to identify risk users with high recall. The model we created was fast- that is, it will identify a user with risk with lowest possible observations. This speed of observation (or observational latency) was an important part of our research, as we could not let a prediction model observe years of user activity before it identifies risk users- as the more time it takes, the riskier it gets for users who need help. We introduced an evaluation metric that measured the performance of said prediction model not only based on its accuracy, but also the number of observations it required to properly predict risk users.

1.1 Organization of the Thesis

This dissertation is organized as follows:

- In Chapter 2, we will present our research domain, techniques that we used for building prediction models and processing natural language, and performance metrics that we used to measure how a model performed in a specific task.

- In Chapter 3, we will dive deeper into engagement prediction in social media. We will discuss our motivations, will survey related works and present our effort to analyze and predict continued participation in multiple social media platforms.

- Chapter 4 will contain our motivations behind working towards building prediction models to detect incivility in public discourse. We will talk about our data collection and cleaning process, and will discuss the prediction models we built, and how we used these models in a cross-platform environment.

- In Chapter 5, we will describe our works in depression detection- why we were interested in it, what were the challenges, and how we faced those challenges. We will show our effort in forms of multiple depression detection models, and will introduce a novel performance metric that will combine traditional performance measures with observational latency to rank depression detection models.

- In the last chapter, we will summarize the works that have been presented in this thesis, and will conclude the thesis.

Welcome to the last five years of my life. Let's begin.

CHAPTER 2

BACKGROUND

As I continued working on the topics that are presented in this thesis, I had learned a lot of things that I did not know before. I had to learn machine learning and natural language processing techniques, and during the course of my PhD, I have seen the both these fields moving swiftly from non-sequential learning approaches to sequential and deep learning approaches. I have used these knowledge I gathered along the way to write up this thesis.

In this chapter, we will learn about some of the basic concepts that will be useful for understanding the work we have done during the course of this thesis and will establish the research domain of this thesis. Background and motivations for each major part of the thesis will be found in their individual chapters. We do not have a dataset that we have used for multiple tasks, so all the data description (sources, collection and cleanup process) are also in their individual sections.

2.1 Research Domain

Online social networks now dominate the Internet, having established themselves as an integral part of human interaction and communication over the last decade. The term "Online Social Network" covers a wide range of services that provide online interaction – from micro-blogging websites such as Twitter to support group based health forums such

as DailyStrength. Popular social networks include Facebook, instagram, Tencent Weibo etc. There is a handful of other media platforms that are commonly considered as social media despite their focus not being a networking platform, e.g. Usenet newsgroups, Reddit, Yahoo! answers etc. Even the platforms that do not have a clear display of a structured interaction among the users may fall into the category of online social network because of their strong inherent social networking properties– for example, massively open online courses (Sinha et al., 2014) and massively multiplayer online role playing games (Kawale, Pal, and Srivastava, 2009; Milosevic, Zivic, and Andjelkovic, 2017). We define online social network as:

An online, open-for-all platform where people can register to access content and perform similar activities among themselves which can be represented in a structured user-user, user-community or community-community communication networks

These *activities* include a wide range of actions performed by users in a platform- e.g. personal messages (chatrooms), user generated or external content sharing (photos, news, stories), creating or commenting in a thread (newsgroups), asking questions or answering them, participating in the same activity together (playing an online game in a common server or taking the same online course) etc.

Although we explored and surveyed a lot of these social networks, to contain our research in a more focused direction, we picked only thread-and-comment based social networks– which are the most common social network out there. We specifically selected

two support group based social networks: Dailystrength[1] and HealthBoards[2], the largest online social news platform: Reddit[3], discourse forum of one newspaper: Arizona Daily Star[4] and one of the most popular social media platforms around: Twitter[5].

2.2 Techniques and Tools

As most of this thesis is on predicting some kind of user behavior in social media, machine learning has a big part in it. We have used both non-sequential and sequential supervised machine learning models for different tasks, and have found varied degrees of success. Our most commonly used machine learning techniques were:

Logistic Regression

Logistic regression is one of the most common non-sequential machine learning techniques. It is a binary classification technique that uses the logistic function to predict the probability of an example belonging to a class. The logistic function looks like this:

$$y = \frac{e^{\hat{y}}}{1 + e^{\hat{y}}}$$

Where $\hat{y}$ is the linear combination of the model's bias, input examples and their respective weights.

[1] www.dailystrength.org

[2] www.healthboards.com

[3] www.reddit.com

[4] www.tucson.com

[5] www.twitter.com

We have used logistic regression mostly as baseline models for our tasks as it is really simple to implement and understand. We have used various implementation of this algorithm- Weka, Liblinear and Scikit-Learn. Although this is popular as a binary classification technique, we have used it for multinomial classification too, using one-vs-rest classification technique.

Support Vector Machines

A Support Vector Machine, popularly known by the acronym SVM, is a machine learning technique that is used for binary classification tasks. In this technique, input vectors are non-linearly mapped to a high dimensional feature space, where a linear decision surface is constructed to classify the inputs. It is one of the most popular non-sequential machine learning techniques because of its high generalization capability. For the non-linear mapping of the feature space to a higher dimension, this algorithm uses a technique called kernel tricks. In our research, we have experimented with various kernels (polynomial kernels, radial basis function etc.) and have used Weka, LibSVM and Scikit-learn's implementations of this algorithm. Like logistic regression, we used this technique for multinomial classification too.

Recurrent Neural Networks

Recurrent neural networks, or RNNs are a form of artificial neural network where neurons are connected to form a directed cycle, allowing the network to exhibit temporal behavior,

and thus be used as a sequential learning model. General structure of an RNN can be seen in figure 2.1.

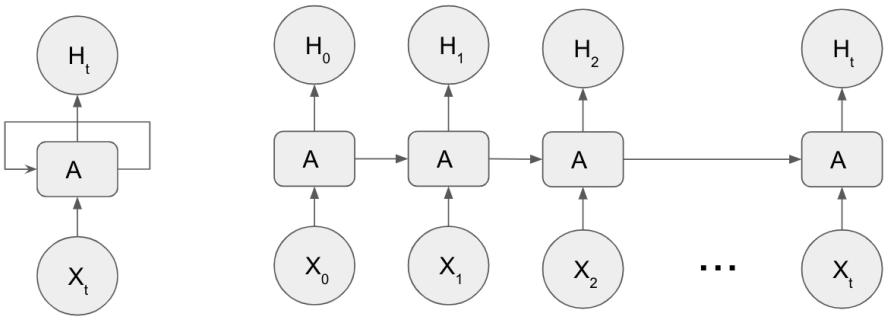


Figure 2.1: Recurrent neural network, Left: rolled, Right: unrolled

We have used RNNs to create prediction models that uses user-generated texts as inputs. As texts are sequential (hence exhibits temporal behavior), RNN proved to be an extremely powerful tool for some of our tasks.

The recurrent layer of an RNN can be made of different recurrent units- two of the most common of them are Long Short-Term Memory (LSTM) units and Gated Recurrent Units (GRU). LSTMs are recurrent units that are capable of learning long term dependencies, and were introduced by Hochreiter and Schmidhuber, 1997b. Gated Recurrent Units are a variation of LSTMs, and was introduced by Cho et al., 2014. The reason of selecting GRUs as units for the recurrent layer is that they can outperform LSTMs in terms of parameter updates and CPU time convergences with the same number of parameters. Internal structures of both these units can be seen in figure 2.2. A simple yet detailed description of the inner workings of these units can be found in Olah, 2015.

RNNs are deep neural networks, and have other layers apart from recurrent layers:

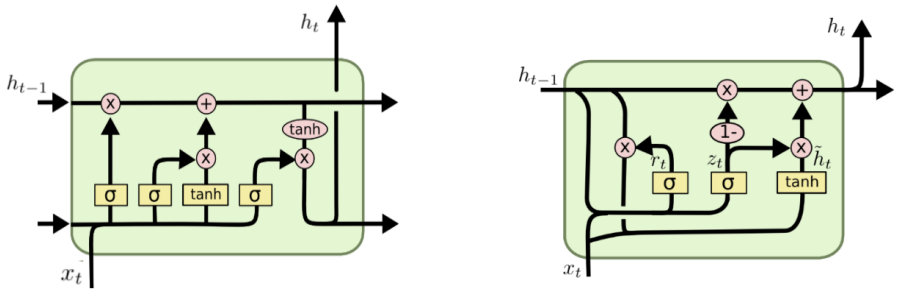


Figure 2.2: Internal structure of a recurrent unit, Left: LSTM, Right: GRU. Image taken from Olah, 2015

Dropout Layer This layer is used to avoid overfitting during training an RNN. It works by randomly deactivating a certain percentage of neurons and forces a layer to learn from a different set of neurons.

Pooling layer This layer is used to reduce spatial dimension in a neural network. **Max pooling layer** outputs the maximum value from the vector it has been applied, and **average pooling layer** outputs the average.

As for the output of the RNNs in our research, we used two forms of dense layers: sigmoid and softmax. Sigmoid is used for binary classification tasks, as it outputs the probability of a class using the sigmoid function: $f(x) = \frac{1}{1+e^{-x}}$. We have used a collection of sigmoids in a single output layer for multilabel classification. Softmax is used for multiclass classification as it generalizes the sigmoid function for multiple classes.

Another important concept that we used in this research is Embedding. We have used word and character embeddings for various tasks. Embeddings are vector representations

of a particular lexical entity, and can capture the context of that entity, along with syntactic and semantic characteristics associated with it. GloVe is the most popular word embedding (Pennington, Socher, and Manning, 2014), and we have used these pretrained embeddings for multiple tasks. We have also used pretrained FastText embeddings (Joulin et al., 2016) for our incivility detection task.

2.2.1 Linguistic Resources

We have taken advantage of some linguistic resources that are freely available out there: we have used Stanford CoreNLP (Manning et al., 2014) for a variety of natural language processing tasks e.g. tokenization, parts-of-speech tagging, sentiment analysis etc. We have also used Linguistic Inquiry and Word Count[6] for emotional word usage analysis in user generated contents.

2.2.2 Information Retrieval Techniques

We have used a handful of information retrieval techniques that are popular in natural language processing. The most popular of them is TF-IDF. TF-IDF (short form for *term frequency-inverse document frequency*) shows how important a term is to a document in a corpus. It is calculated by multiplying two measures:

Term frequency measures how frequently a term occurs in a document. Term frequency

[6] `http://www.liwc.net/`

of a term t in document d is defined as:

$$TF_d(t) = \frac{N_d(t)}{\sum_{t \in T} N_d(t)}$$

where $N_d(t)$ represents the number of times t has occurred in document d.

Inverse document frequency measures how important a term is in a corpus. This is important as it discounts the term frequency importance of terms that occur frequently in multiple documents like articles and conjunctions. Inverse document of a term t is defined as:

$$IDF(t) = \log \frac{|D|}{\sum_{d \in D} d_t}$$

Where $|D|$ represents the number of documents in the corpus and d_t represents documents that have the term t in it.

TF-IDF is the multiplication of these two measures.

Another technique that we used for information retrieval is Pointwise Mutual Information (PMI). It measures the association of two events given their joint probability and individual probability, and assumes independence between these. It can be represented like this:

$$PMI(x,y) = \log \frac{p(x,y)}{p(x)p(y)}$$

Where x and y are two independent events, $p(x,y)$ represents their joint probability, and $p(x)$ and $p(y)$ are their individual probabilities. We used this technique to find out the association of sets of words with certain classes, and used these association values as inputs for our machine learning models.

2.3 Performance Metrics

For all of our models, we have presented (at least) four performance metrics: accuracy, precision, recall and F-measure.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2.2}$$

$$Recall = \frac{TP}{TP + FN} \tag{2.3}$$

$$F\text{-}measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{2.4}$$

Where TP is the number of *true positives* (where model predicted true and the ground truth was also true), TN is the number of *true negatives* (where model predicted false and the ground truth was also false), FP is the number of *false positives* (where model predicted true but the ground truth was false) and FN is the number of *false negatives* (where model predicted false but the ground truth was also true).

We have also used other performance metrics for specific tasks (e.g. Early Risk Detection Error, or ERDE)- they are described in detail in their specific chapters.

CHAPTER 3

ENGAGEMENT

3.1 Background and Motivations

There is little research on computational assessments of the level of engagement in popular social networking sites– one of the most significant contributors of social capital in online forums and social networks. Engagement in large services like Twitter or Usenet newsgroups has been explored over the years, but smaller networks, specifically support group or community based forums have received little to no attention.

Engagement, or continued participation is a frequently researched topic across many industry sectors. A common way of talking about continued participation is in terms of *churn* – a portmanteau of *change* and *turn* – which is the rate of loss of customers from a company's customer base to another company. Research on churn has a simple motivation: loss of customers is loss of revenue, and retaining a customer is much cheaper than winning a new one (Hadden et al., 2007). Generally a company tries to identify a churning customer early in their lifecycle so that customer management departments can efficiently target these customers and provide incentives to prevent them from leaving the company. Among these industries, telecommunication sectors have contributed extensively in the research of churn among their customers (Kim and Yoon, 2004; Gerpott, Rams, and

Schindler, 2001; Keaveney, 1995; Mozer et al., 1999; Burez and Poel, 2009; Dasgupta et al., 2008).

Consequently, engagement is an important factor for social network services since they follow the same business model as the service providers in telecommunication sectors: you lose revenue when a customer leaves the network. However, in social networks, the threat is much more than monetary. As social networks thrive on the interactions among users, loss of users means loss of social capital within the service, which ultimately affects the sustainability of the service. The strict definition of churn also typically does not apply to social networks, as users may or may not join another service after leaving the current one. Instead, terms like *continued participation*, *engagement*, *attrition*, or *defection* are more commonly used. In our research, we adopt the term *engagement* since it encompasses the broadest range of phenomena.

Factors that influence engagement in social networks can vary from service to service. Graph based features can play a big role in predicting participation in those services which maintain an extensive architecture of relationships among the users like Facebook, whereas the frequency of activities plays a bigger role in the prediction task in services like forums and discussion boards. Demographic information, contents of texts, and timelines within user lifecycles can contribute significantly depending on the paradigm of the prediction task.

3.1.1 Challenges in Predicting Engagement in Social Media

The first challenge while exploring engagement in social networks comes from the motivation of a user to participate in a network. Social network users invest their time, sharing views or opinions or simply participating in a discourse, without expecting any immediate return from the network (Constant, Sproull, and Kiesler, 1996). In sociology this type of activity is known as the "Gift Economy" (Rheingold, 2000), which, in contrast to the the service or commodity economy, is not driven by exchanging service or commodities for monetary benefits, but rather is driven by the expectations of social contracts. Several motivations drive users to participate in this economy of gift transactions, for example, the expectations of future payback in terms of new information and social interaction, recognition as a source of valuable information from peers or idea diffusion among other users in the community; and when these expectations are not met, users tend to leave the community, thus hurting the social capital of the network in the process. Social networking services lose revenue when users leave their network, just like other industries; but this loss of social capital poses a greater threat to the services as this threatens the survival of the social networks in the long run. Identifying the phenomenon of not meeting expectations is difficult, and poses the biggest challenge to the analysis of future engagement.

Another challenge in predicting continued participation in online social networks is that there are no predefined "triggering events" (Gustafsson, Johnson, and Roos, 2005) in social networks as there are in telecom sectors. In telecommunication services, a subscriber is bound by a service contract or he buys credits before using the service. When the contract

expires, or the credit dries up, churn is triggered based on the other factors like service quality, tariffs or poor customer experience. In social networks, users are weakly tied by a non-binding social contract (Constant, Sproull, and Kiesler, 1996). A user can leave a social network any time without incurring any kind of explicit monetary penalty, and can again join the network any time as there is low-entry barrier to join most social networks. This absence of triggering events makes it more difficult to predict continued participation in social networks than to predict churn in industries like telecom.

One other challenge while predicting continued participation in social networks is the diversity and the growth of the social networks (Karnstedt et al., 2011). There are chatrooms, discussion boards, community forums, photo and video sharing websites, blogs, massively multiplayer online games, online courses and many others which accumulate two or more of these services into them. The inner structures of these services are highly diverse and complex. Discussion boards and blogs are mostly for sharing ideas and views by posts and replies in threads, and interpersonal communication among the users in these services are generally sparse, whereas chatrooms and online games depend mostly on the dense interpersonal communication among the users. Also, there are hierarchies of engagement in most of these services: a user can stop communicating with a single user or a set of users, or he can stop participating in a forum or a single thread, or he can leave the network entirely.

Another challenge that makes engagement prediction in social networks more difficult than predicting churn in the telecom sector is that in social networks, participation is a

continuous process. A user does not suddenly drop off of a social network, it happens over a significant period of time. There is no certain triggering event in social networks as there is in telecom services; a user may gradually decrease his or her participation in the community and eventually stop participating at all.

Due to these challenges, engagement prediction in social networks is still largely unexplored, and thus represents a major research opportunity in this field. There have been a few works on predicting future participation in popular paradigms like micro-blogging (e.g., Twitter; Mahmud, Chen, and Nichols, 2014; Chen and Pirolli, 2012) and massively multiplayer online role playing games (e.g., EverQuest II; Kawale, Pal, and Srivastava, 2009), but paradigms like health forums (e.g. DailyStrength) are still mostly unexplored. A number of these social networking services provide their data for nonprofit and research purposes, and there is a huge opportunity to apply data mining and natural language processing in these data to establish successful engagement prediction models for these social networking paradigms.

3.1.2 Related Works

As we have mentioned earlier, not a lot of work has gone into the research of engagement prediction in social networks. One of the most prominent works among those done in the field is from Elisabeth Joyce and Robert E. Kraut. They published a paper in 2006 which attempted to discover significance of various characteristics (e.g. numbers of replies received, length in words, being a question or testimonial, emotional tone etc.) of user's

first post in a Usenet newsgroup and the replies to that on future participation (Joyce and Kraut, 2006). They also hypothesized that characteristics of the initial post should also influence the continued participation as it explicitly influences whether it will get a reply and also the quality of the reply. Using a probit analysis the authors found out that the group in which the initial message was posted influenced the likelihood of getting a reply, and that longer initial posts and receiving a reply both have positive significance over the prediction value. Effects of the characteristics of the replies also varied from group to group, but eventually the only characteristic that had some consistent significance over the prediction value in all groups was whether the reply was a question or not. The effect of emotional tones in both the initial post and the replies vary over the different groups- and thus the authors conclude that, out of the six hypotheses, only the first one ("Receiving a response to an initial post will increase the likelihood that the poster will post again") is supported, the second one ("An initial post that receives a response that provides information rather than asks a question will increase the likelihood that the poster will post again") was disconfirmed and the other four are not supported.

Arguello et al. attempted to find the factors that influence the number of replies a post gets in Usenet newsgroups which in essence captures the success of an online community (Arguello et al., 2006). As Joyce and Kraut suggested that the responses a user gets from his or her first post play a crucial role in his or her continued participation, this paper also tries to predict whether a user returns or not based on these analyses. The factors that the authors explored (and also used as features for learning) are divided into certain

categories: Group-level factors (group identity, cross-posting and group size and volume), Individual level factor (Newcomer status) and message characteristics (Rhetoric, Topical coherence, linguistic complexity and word choice: Both linguistic complexity and word choice took advantage of the LIWC lexicon). The authors used the same probit analysis used by Joyce and Kraut to predict whether a message receives a reply, with four different sets of independent variables- where each set introduces a certain class of new features to the base model. The analysis showed that the group a user posts into has a significant influence on whether that user receives a reply, along with some characteristics of the post, i.e. being a testimonial increased the likelihood of getting replies by 10% and being a topical question increased the probability by 6%. Usage of longer and more complex sentences reduced the probability of getting a reply, whereas sentences containing more first person singular pronouns and third person pronouns increased the likelihood. The authors included a new independent variable *gotReply* which denoted whether a user has received a reply or not in his or her first post. Based on this variable, they reported that getting a reply increased posters' probability of posting again by about 6.2%. They also found out that receiving replies from newcomers or having complex replies hurts the probability, whereas receiving replies with more positive emotion words actually improves the probability. Their concerns about the model were that the dataset was not large enough and the usage of bag of words as a measure of topical coherence, as this ignores syntax and context and only considers the usage of words.

Twitter has received more attention than any other social network in this field, as

we have seen in the works of Mahmud, Chen, and Nichols, 2014 and Chen and Pirolli, 2012. Mahmud et al. worked on predicting social engagement behavior by means of response and retweet where they used various psycholinguistic categories obtained from Linguistic Inquiry and Word Count (LIWC) database. In their analysis, they found out some categories (anger, cognition, communication, anxiety, social process, positive feelings, positive emotions etc.) that have a noticeable statistical significance- both positive and negative- with two independent variables- reply rate and retweet rate. Their reported system could predict response and retweet rate with below 30% mean absolute error and could predict future engagement based on these rates with a 72-85% accuracy based on which LIWC categories were used. Chen and Pirolli focused on exploring factors influencing engagement of Twitter users in a real-life event (#OccupyWallStreet movement) also based on retweets and replies- but rather emphasizing on the contents of the tweets, they looked into the activities like number of tweets, number of followers, number of followees, number of retweets, number of posted mentions, number of retweets and mentions from followers, user demographics etc. The study found strong support for one of their hypotheses, that more interaction before the movement led to more engagement during the movement.

In their 2013 paper, Danescu-Niculescu-Mizil et al. focused entirely on the linguistic attributes of the activities performed by users in a community to predict lifecycle of the said user in two beer rating communities (Danescu-Niculescu-Mizil et al., 2013). Their target was to create models that could analyze a user's linguistic change over time based on his or her adoption of lexical innovations, similarity to group's linguistic trend, use of

certain classes of words etc. These models (called snapshot language models- essentially bigram language models with Katz back-off smoothing(Katz, 1987)) were then used to predict future engagement of a user in the said community with considerable performance improvement over a previously set baseline.

Hamilton et al. explored loyalty- which is a different take on continued participation- in online communities in Reddit[1] in their 2017 study (Hamilton et al., 2017). Two key aspects of loyalty, which they define as a combination of preference and commitment, are explored by the authors in this study: user loyalty, where a loyal user prefers a community over others, and community loyalty, where a loyal community retains its loyal users over time. User loyalty depends on individual user's linguistic and behavioral attributes. Upon analyzing the contents of the posts where loyal users post more than the vagrants (those who are not loyal to the community) do, the authors concluded that loyal users prefer posts with more esoteric contents- where the esotericity of a post is calculated by averaging the inverse document frequency of the noun phrases in the content. In their loyalty prediction task using only the first post of a user, the authors found out that these linguistic features are decent predictors of loyalty in 58% of the subreddits- which indicates that loyal users display affinities to certain stylistic elements really early in their user lifecycle.

Online multiplayer games have received considerable attention over the years. Milosevic et al. predicted churn in a popular mobile social game named *Top Eleven- Be a Football Manager* where they used user activities, virtual monetization and gameplay styles to

[1] `http://www.reddit.com`

identify churners (Milosevic, Zivic, and Andjelkovic, 2017). Kawale et al. approached the gaming domain from a different perspective- they created influence diffusion models from player activities in EverQuest II and used network driven features to predict player churn (Kawale, Pal, and Srivastava, 2009). Sinha et al. used clickstream and forum activities to form user-level activity graphs in the popular massively open online course website Coursera and used network features to predict user attrition.

3.2 Predicting Engagement in DailyStrength

We started our work on user engagement in social media on one of the largest support group based social media platforms, *DailyStrength*[2] (Sadeque et al., 2015). This website has more than 500 support groups to date, and is a thread-and-comment sort of platform. Users can create a thread, or comment to other threads. The commenting hierarchy was flat- a user could not comment on other user's comments. For our purpose, we selected 20 support groups, which focused on either physical or mental ailments of users, or in some case, both of them. These groups were: Acne, ADHD (Attention Deficit Hyperactivity Disorder), Alcoholism, Asthma, Back Pain, Bipolar Disorder, Bone Cancer, COPD (Chronic Obstructive Pulmonary Disease), Diets and Weight Maintenance, Fibromyalgia, Gastric and Bypass Surgery, Immigration Law, Infertility, Loneliness, Lung Cancer, Migraine, Miscarriage, Pregnancy, Rheumatoid Arthritis, and War in Iraq. Gastric and Bypass Surgery was the largest among these 20 with 21507 posts and 158020 replies,

[2] `www.dailystrength.org`

whereas Bone Cancer was the smallest with only 40 posts and 51 replies. Individual activity also varied greatly as there are people who posted or replied only once in their lifetime, and there were people who have more than 5000 posts or replies. Although these groups had varied characteristics, we did not consider each of them as different communities, rather, we considered all of them as parts of a single, larger community (Our next work considers each different type of support group as a different community- we will talk about that later in this chapter). The general statistics of this larger community are given in table 3.1.

Support groups	20
Posts	110316
Replies	788119
Users	39905

Table 3.1: Summary of the data collected from DailyStrength

For our task, we crawled all of the thread initiations and replies to existing threads for all of these support groups from the earliest available post until the end of September 2013 (we started working on this task back in 2014). The posts and replies were downloaded as HTML files, one per thread, where each thread contained an initial post and zero or more replies. We then parsed and filtered these files to extract pertinent information (user id, date, post and reply texts), part-of-speech tagged all texts using the Stanford part-of-speech tagger (Manning et al., 2014) and used Linguistic Inquiry and Word Count (LIWC) lexicon to tag emotion words. For the users' demographic information, we collected the user

profile pages of all the users we identified in the previous step. We filtered out the users with the most incomplete profiles, where they were missing both age and gender. These users do not appear in the train, development or test sets, but their replies they post on other users' posts who are not filtered out contribute to the participation prediction task of those users.

3.2.1 Definition of Engagement

To proceed with our work in engagement, we needed to establish a definition of what an engaged user meant in support group based social media, as there were no prior work on engagement done in this paradigm. We came up with the simplest of definitions: *A user is identified as engaged if a user has already participated in a community for a previously determined period of time (observation period), and then continues their participation beyond that period.* Any user who does not have any activities at any point beyond the said observation period has discontinued their participation. We introduced an engagement prediction model based on this definition:

$$m_{\Delta t}(u) = \begin{cases} 1 & \text{if } \exists a \in A: \quad a.u = u \wedge a.t > u.t + \Delta t \\ 0 & \text{otherwise} \end{cases}$$

where u is a user, Δt is an amount of time which we call the *observation period*, A is the set of all activities (from any user at any time) such as posting or replying to a post, $a.u$ is the user whose activity it was, $a.t$ is the time of the activity, and $u.t$ is the time at which the user account was created. Intuitively, m should predict 1 (engaged) iff Δt time has

elapsed since the user created their account and there is any new participation (posting or replying) any time in the future after that.

3.2.2 Features

For this prediction model (which we use as a supervised classification model), we explored a set of features:

Activity features

These features gather information of a user's activity on DailyStrength. In general, we would expect users who are more active during the observation period to also be more likely to continue to participate in the future.

PostCount The number of threads a user has initiated on the DailyStrength website over the observation period.

ReplyCount The number of replies a user has posted to other users' posts on the DailyStrength website over the observation period.

SelfReplyCount The number of replies a user has posted to their own posts over the observation period.

OtherReplyCount The number of replies a user has received to their posts from other users over the observation period.

Time features

These features provide a look into the timing of a user's participation on DailyStrength. In general, we would expect users who are participating frequently throughout the observation period to be more likely to participate in the future.

TimeGap1 The number of days between the point at which the user created their DailyStrength account and their first activity (post or reply). This is a measure of how long it took a user to start actively participating in the community.

TimeGap2 The number of days from the time of the last post or reply of a user to the end of the observation period. This is a measure of how long the user has been idle since their last activity.

AvgDays The average number of days between any two sequential activities (posts or replies) by the user during the observation period. This is a measure of how often a user is idle.

Personal features

These features are gathered from a user's account information page. Since providing age, gender, location and a profile photo are all optional during the DailyStrength account creation process, many users are missing one or more of these pieces of information. In general, we would expect users with more complete profiles to be more likely to continue to participate.

Age The user's age.

Gender The user's gender, either *male*, *female* or *unknown*.

HasLocation A binary feature representing whether or not the user has provided their location.

HasImage A feature representing whether or not the user has provided a profile photo.

Content features

These features examine the content of the text in the posts and replies of a user. In general, we would expect users with longer posts to be more likely to continue to participate that users with short posts.

PosUnigrams The total number of words over the observation period that were identified as positive emotions by the LIWC lexicon.

NegUnigrams The total number of words over the observation period that were identified as negative emotions by the LIWC lexicon.

TotalUnigrams The total number of words a user posted over the observation period. This includes all the words (including stop words), not only the emotion words.

Question The total number of questions the user has asked over the observation period in either posts or replies. Questions were identified by looking for sentences ending in question marks.

Period	Baseline	Accuracy	Error Reduction	Precision	Recall	F-measure
1	50.48	83.06	65.81	88.3	80.2	84.0
3	63.53	85.69	60.76	92.0	86.4	89.1
6	72.01	87.69	56.02	94.7	89.0	91.7
9	77.75	89.12	51.10	94.9	91.4	93.1
12	82.19	90.71	48.01	96.3	92.7	94.5
15	85.09	92.03	46.54	97.1	93.8	95.4
18	86.96	92.29	40.87	97.3	94.0	95.6
21	87.65	92.34	37.98	97.7	93.8	95.7
24	87.84	92.34	37.01	97.8	93.7	95.7

Table 3.2: Performance across different observation periods (months)

Url The total number of URLs a user has posted over the observation period.

3.2.3 Experiments and Analysis

For our prediction task, we used 60% of the users to train our prediction model, 20% of the users as the development set and the remaining 20% of the users to test the model. Users were partitioned into each of these sets randomly, and the sets were kept unchanged for the purpose of comparing models with different observation periods. As our learning algorithm, we used logistic regression implemented in Weka v.3.6.11 (Witten and Frank, 1999) as it outperformed other techniques for this task.

We had four major questions that we wanted to answer in this research:

- Can continued participation be predicted?
- How long must a user be observed?

- Which features are most important?
- Does feature importance change over time?

Can continued participation be predicted?

Our first research question is whether a user's continued participation on the forum can be predicted given the features we developed. To test this, we consider an observation period of 1 month and train and test the corresponding classifier. The first row of table 3.2 shows the results. Our model achieves 83.06% accuracy, compared to the 50.48% accuracy of the baseline model. For the task of identifying just those users that have stopped participating, we achieve 88.3% precision and 80.2% recall. These high performance numbers suggest that while our models are still imperfect, our features are capturing a large proportion of the information necessary to predict continued participation.

How long must a user be observed?

Our second research question aims to determine the optimal observation period for predicting continued participation. For this experiment, we created 9 observation periods: 1 month, 3 months, 6 months, 9 months, 12 months, 15 months, 18 months, 21 months and 24 months. We then evaluated models trained on these different evaluation periods to see how performance increased or decreased.

Table 3.2 shows the results. Model accuracy always rises as the observation period grows longer, ranging from 83.06% at 1 month to 92.34% at 24 months. However, the

Activity Features		Timeline Features		Personal Features		Content Features	
Feature	Weight	Feature	Weight	Feature	Weight	Feature	Weight
PostCount	0.526	TimeGap1	3.172	Age	-0.346	PosUnigrams	3.002
ReplyCount	-10.652	TimeGap2	5.727	HasLocation	-0.092	NegUnigrams	-3.069
SelfReplyCount	0.001	AVGDays	0.809	HasImage	-0.845	TotalUnigrams	4.772
OtherReplyCount	-0.051					Question	0.827
						URL	1.834

Table 3.3: Weights of the features for a 1-month observation period

biggest gains are in the shorter periods, with the model increasing accuracy by 7.65% between 1 and 12 months, but only by 1.63% between 12 and 24 months. The performance of the baseline model also increases with the size of the observation period, so that after 24 months 87.84% of all users will not return.

For the task of identifying just those users that have stopped participating, we observe that precision and recall also both rise as the observation period grows, with precision making moderate gains, from 88.32 at 1 month to 97.8 at 24 months, and recall making larger gains, from 80.20 at 1 month to 93.7 at 24 months. As with accuracy, the biggest gains are between 1 and 3 month observation periods.

Overall, these results suggest that observing a user for even 1 month gives reasonable performance, observing for 12 months gives noticeably better performance, and observing for longer than 12 months gives diminishing returns.

Which features are most important?

Our third research question aims to prioritize our features based on how useful they are to the task of predicting continued participation. To investigate this, we turn to the coefficients (weights) for the independent variables (features) in our logistic regression, which represent the importance of each variable in the classification model. The larger the absolute value of the coefficient, the bigger the impression of that variable on the output. The sign indicates positive or negative effect of that variable on the result, where a negative value means that the feature is associated with continued participation, while a positive value means that the feature is associated with stopping participation.

Table 3.3 shows the weights of the features obtained from the test data for a 1-month observation period. The most important features (the features with the highest absolute values) are the number of times the user has replied to other users (ReplyCount), the time since the user's last activity (TimeGap2), the time between creating a DailyStrength account and the user's first post (TimeGap1) and the content (Unigram) features. The least important features are mostly the ones aimed at measuring completeness of the profile (Age, Gender, etc.), suggesting that profile completeness is not a good predictor of continued participation. However, the presence of a profile photo (HasImage) did make a small contribution to the model.

The signs of the weights of the features reveal the direction of predictiveness. The TimeGap1 and TimeGap2 weights are positive, indicating that longer gaps between activities predict someone leaving the forum. PostCount is positive while ReplyCount is

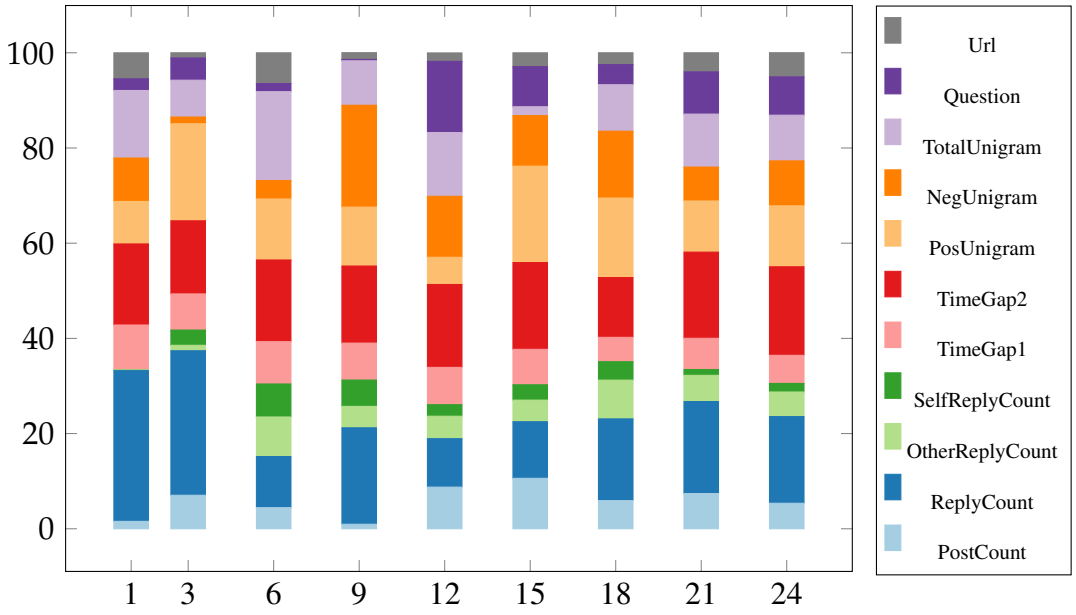


Figure 3.1: Relative importance of features over the different observation periods. The height of a bar segment represents the absolute value of the weight of the feature, scaled so that the sum of the feature weights is 100%.

negative, suggesting that people who only post will likely leave the forum, while people who reply to others will likely stay. Posting questions and URLs are associated with leaving the forum, along with higher usage of positive unigrams, while higher usage of negative unigram is associated with continued participation.

Does feature importance change over time?

Our fourth research question asks whether the importance of features is consistent across all observation periods, or whether some features become more or less important than others as the observation period grows.

Obs. Period (months)	1	3	6	9	12	15	18	21	24
Baseline	50.48	63.53	72.01	77.75	82.19	85.09	86.96	87.65	87.84
PostCount	4.91	3.05	1.53	0.81	0.13	0.43	0.23	0.08	0.08
ReplyCount	6.55	3.75	1.89	1.03	0.29	0.47	0.20	0.14	0.19
OtherReplyCount	2.97	2.43	1.40	0.74	0.11	0.37	0.18	0.00	0.01
SelfReplyCount	2.13	1.42	0.77	0.24	-0.13	0.27	0.05	0.00	-0.01
TimeGap1	0.00	0.00	0.00	0.00	-0.20	0.15	0.00	0.00	0.00
TimeGap2	31.7	20.7	14.9	10.8	7.76	6.16	4.92	4.43	4.42
AvgDays	4.72	0.00	0.00	0.00	-0.20	0.14	0.00	0.00	0.00
Age	0.36	0.14	0.00	0.00	-0.20	0.14	0.00	0.00	0.00
Gender	-0.20	0.00	0.00	0.00	-0.20	0.14	0.00	0.00	0.00
HasLocation	0.00	0.00	0.00	0.00	-0.20	0.14	0.00	0.00	0.00
HasImage	11.9	0.00	0.00	0.00	-0.20	0.14	0.00	0.00	0.00
PosUnigram	5.10	3.06	1.60	0.73	0.15	0.33	0.15	0.07	0.08
NegUnigram	4.32	2.65	1.43	0.72	0.09	0.25	0.10	0.07	0.07
TotalUnigram	4.98	2.73	1.50	0.82	0.13	0.35	0.09	0.09	0.09
Question	4.03	2.52	1.46	0.83	0.03	0.36	0.13	0.19	0.20
Url	0.10	0.30	0.11	0.07	-0.18	0.25	0.02	0.02	0.01

Table 3.4: Accuracy gain over Baseline over observation periods when a classifier is trained using only a single feature

Figure 3.1 shows the percentage importance of the eleven most significant features over the different observation periods. Features like TimeGap1 and TimeGap2 are fairly stable in importance over time, with TimeGap1 accounting for 5-9% of the weight and TimeGap2 accounting for 12-18%. ReplyCount is a very strong feature, accounting for as much as 30% in the 1 and 3 month observation periods, but it receives a lower weight for longer observation periods (as little as 10% in the 12 month period). SelfReplyCount and OtherReplyCount, which had almost no weight in the 1 month model, increase in importance for longer observation periods. The other features have less consistent patterns. For example, content features (TotalUnigram, NegUnigram, PosUnigram, Question, Url) account for around 40% of the model weights for most observation periods, but the

distribution of weight across these 5 features is erratic over time.

As another measure of feature importance over time, table 3.4 shows the increase in accuracy over the baseline majority class model for models trained using only a single feature. Note that the baseline model's accuracy increases for longer observation periods (because more users leave), so the absolute gains over the baseline always correspondingly decrease. TimeGap2 (TG2) always gives the largest increase in accuracy on its own, as much as 31.7% at a 1 month observation period, and is the only feature that continues (by itself) to give gains over the baseline all the way out to 24 months. ReplyCount (RC) is the next best feature by itself, achieving 6.55% improvement over the baseline at a 1 month observation period, but dropping to less than a 1% improvement by 12 months. The content features PosUnigram (Pos), NegUnigram (Neg), TotalUnigram (TUn) and Question (Que) each achieve a 4-5% improvement over the baseline for a 1 month observation period, but drop below a 1% improvement by 9 months. The personal features generally achieve very little on their own, except for HasImage (Img), which is very useful at 1 month (giving a 11.9% improvement), but giving no improvement for any other observation period.

3.2.4 Discussion

Our findings have several implications for social interaction in online health forums. This is the first study that attempts to predict continued participation of users in such support groups. Though the model is not perfect, it produces results with high accuracy, precision and recall. The high precision and recall has greater significance in this experiment, as they

represent our model's correctness in identifying the people who leave the group after a certain observation period. Identifying these people early in their lifecycle will help social health platforms identify users that are not being fully served, allowing the platforms to analyze the reason for the departure and create a more favorable environment for everyone.

This is also the first study that examines the effect of different lengths of observation period to determine the minimum amount of time required to accurately predict future participation. With a 12-month observation period, we can predict continued engagement with high accuracy, precision and recall, though even at a 1-month observation period, performance is good.

Our work has shown which features contribute the most to predict a user's continued participation. As we can see from the results, personal features covering demographics and profile completeness play little to no part in predicting user's engagement, whereas the other three categories have varied significance over time. The predictiveness of time based features, especially the time from account creation until a user's first activity and time since a user's last activity, are consistently predictive over all lengths of observation. The predictiveness of replies to other users' posts is very large for 1 and 3 month observation periods, but is a little less informative for larger observation periods. The predictiveness of content features (word count, negative/positive words, etc.) is generally good, though which of these features is most important varies somewhat over time.

3.3 Engagement Analysis in HealthBoards

Recent activity and change of policy in DailyStrength made us move to a similar support group based social media named *HealthBoards*[3]. Healthboards also features similar thread-and-comment based inter-user communication structure, and was one of the most popular support group communities on the Internet. We extended from our previous work and focused more on individual communities rather than considering all of them as one large community. We also focused more on the linguistic features of the user generated contents and the timeline information, as they were the most contributive features in our previous work. Also, during this research period we found out that depression related communities have some unique attributes compared to other communities that can help us predicting users' depression levels from an assortment of communities (Sadeque et al., 2016).

3.3.1 Data Collection and Analysis

We started this work with data collection- like the previous task, we crawled HTML pages from the website, and stored pertinent information in files compliant with JSON-based Activity Stream 2.0 specification from the World Wide Web Consortium (W3C, 2015). We had three major focus groups for this Task- depression, relationship health and brain/nervous system disorders. The last forum consisted of multiple subforums: Arachnoiditis, Alzheimer's Disease and Dementia, Amyotrophic Lateral Sclerosis (ALS), Aneurysm, Bell's Palsy, Brain and Head Injury, Brain and Nervous System Disorders,

[3] `www.healthboards.com`

Brain Tumors, Cerebral Palsy and Dizziness/Vertigo. The reason behind selecting these particular forums was simple: we tried to focus on the relationship of engagement and mental health at this point, and focused on one forum that may represent a set of social factors interacting heavily with mental health (relationship health) and one forum that represent neuropsychiatric disorders from a more physical perspective, while keeping our main focus on depression, which is a combination of social factors and neuropsychiatric disorder. Like the previous work, we part-of-speech tagged all the texts and extracted emotion words from LIWC lexicon. Table 3.5 shows the summary of the data we collected from Healthboards. As we can see, all three forums are roughly similar in number of users. However, users in the Depression forum are less engaged than users in Relationship Health, having a lower average number of replies per post and a lower average number of replies per user. While the Depression forum is similar to the Brain/Nervous System Disorder forum in terms of posts and replies per user, there are more users in the Depression forums that choose not to specify their gender.

Unlike the previous work, we did not jump into the prediction task right away– rather, we analyzed the useful features we identified in that work for engagement analysis in this community. Our first hypothesis was that a user's last post may contain some linguistic cues of their decreasing social interaction. To experiment on this, we considered all users from the three forums who were inactive for at least one year preceding the day of data collection. We used pointwise mutual information (PMI) between users' last posts and n-grams collected from these users' posts to identify phrases that have more association with

	Depression	Relationship	Brain/Nervous
Posts	19535	17810	13244
Replies	105427	199430	74974
Users	15340	12352	14072
Reply/Post	5.4 (0.1)	11.2 (0.1)	5.6 (0.1)
Post/User	1.3 (0.03)	1.4 (0.03)	0.9 (0.03)
Reply/User	6.9 (0.4)	16.1 (1.3)	5.3 (0.7)
Gender: male	20.77%	22.15%	22.99%
Gender: female	54.07%	57.52%	59.16%
Gender: unspec.	25.16%	20.33%	17.85%

Table 3.5: Summary of the data collected from HealthBoards. Numbers in parentheses are standard errors.

last posts than other random activities. A list of top 10 unigrams and bigrams according to PMI for each forum is given in table 3.6. These phrases suggest differences in reasons for leaving different types of forums. Depression has some especially revealing phrases: people appear to withdraw from the forum after starting treatment (*of Pristiq*, *depression medication*), but also after apparent calls for help (*'m suffering*, *cut myself*, *Any help*).

Our next hypothesis was that there may be observable changes over time in the language of users who are disengaging from the community. Using PMIs as above, we identified the top five LIWC psycholinguistic classes most associated with last posts: Social, Cognition, Affect, Positive Emotion, and Negative Emotion. Then we selected the top 100 most active users from two cohorts- one with the top 100 users who were inactive for at least one year preceding the day of data collection, which we call the non-returning (NR) cohort, and the

	Unigram	PMI	Bigram	PMI
Depression	iv	0.48	I+Feel	0.54
	Husband	0.45	of+Pristiq	0.53
	Ritalin	0.41	My+fiance	0.52
	pristiq	0.40	My+partner	0.50
	electric	0.38	depression+medicaiton	0.48
	cheated	0.37	in+middle	0.47
	adderall	0.37	'm+suffering	0.47
	Due	0.36	slept+with	0.46
	depression	0.36	cut+myself	0.46
	affair	0.36	Any+help	0.46
Relationship Health	wat	0.67	i+no	0.77
	introvert	0.63	this+disorder	0.74
	narcissist	0.62	a+narcissist	0.70
	iv	0.60	wife+said	0.67
	Bipolar	0.59	He+constantly	0.66
	thankyou	0.58	dad+does	0.65
	idk	0.57	confessed+that	0.65
	ADD	0.55	Just+recently	0.64
	schizophrenia	0.54	my+fiance	0.63
	episodes	0.52	she+continued	0.63
Brain and Nervous System	Bells	0.49	got+Bells	0.76
	tingly	0.45	centre+of	0.73
	hypochondriac	0.44	neural+canal	0.72
	ventricles	0.43	prominence+of	0.72
	temple	0.42	ears+from	0.68
	ms	0.41	bulge+with	0.68
	tumour	0.41	mild+posterior	0.67
	ADD	0.40	small+intestine	0.67
	temporal	0.40	your+biggest	0.67
	cyst	0.40	are+increasing	0.67

Table 3.6: Top 10 unigrams and bigrams from each forum based on their PMI with last posts.

other included the top 100 users with high activity but not marked as inactive yet, which we call the returning (R) cohort. these users were selected based on two features:

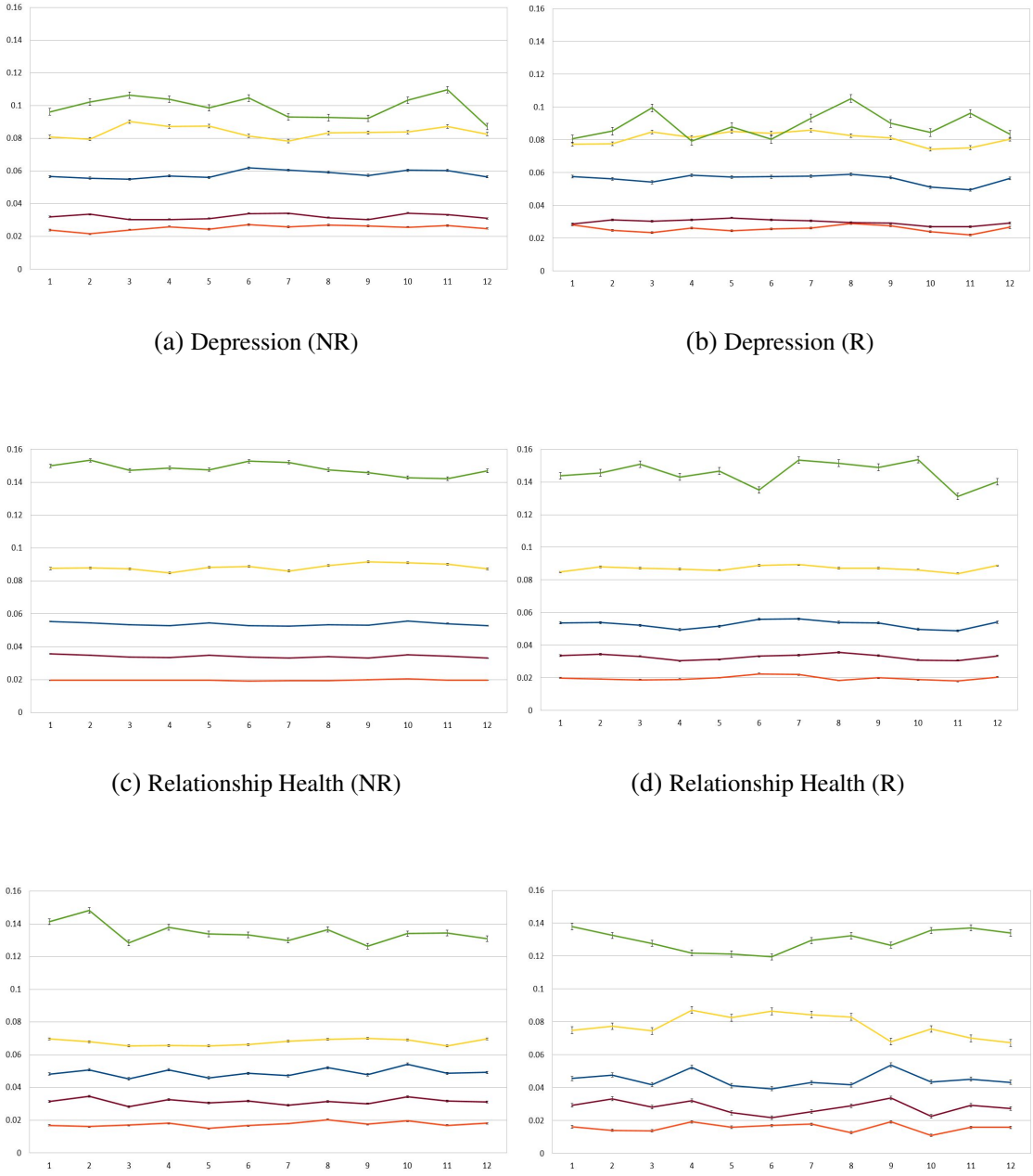
- posted in at least two different years
- made at least 100 posts or replies

Figure 3.2 shows the use of words from different psycholinguistic classes over the last 12 months of the selected users' timeline. For most word classes, usage is fairly constant over time and similar across the forums. However, use of social words in the Depression forum is about 40% lower than in Relationship Health or Brain/Nervous System Disorder. This reduced use of social words may indicate less social interaction and less energy, consistent with signs of recurring depressive episodes. Interestingly, both the returning (R) cohort and the non-returning (NR) cohort exhibit this behavior.

During the period we collected the data, we encountered many posts with negative sentiment after which the user stopped participating in the forum, for example:

> . . . I was really frightened of what was happening to me, my Mum took me straight back to the doctors, to a different one, they were useless, they put me straight on zoloft, I took the zoloft for about 3 days when everything got worse, I couldn't eat, I kept throwing up, I was having constant panic attacks I just wanted to sleep but lived in fear when I was alone. . . [4]

[4] http://www.healthboards.com/boards/2346283-post1.html



(a) Depression (NR)

(b) Depression (R)

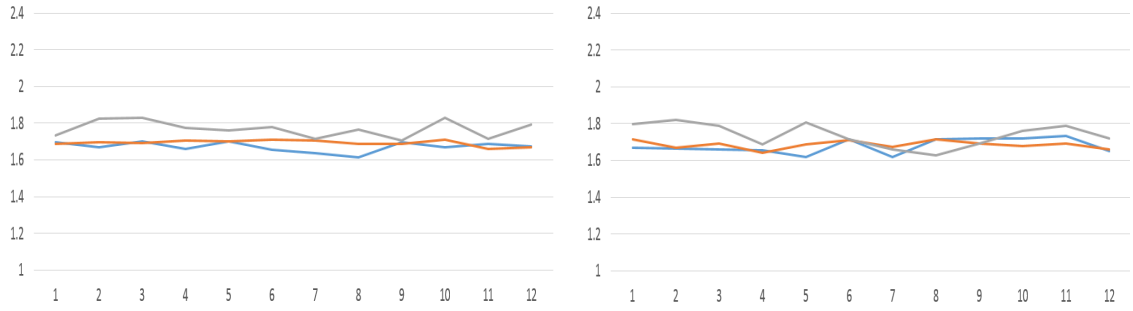
(c) Relationship Health (NR)

(d) Relationship Health (R)

(e) Brain/Nervous System (NR)

(f) Brain/Nervous System (R)

Figure 3.2: The final 12 months of psycholinguistic word use by category: Social (top; green), Cognition (2nd from top; yellow), Affect (middle; blue), Positive Emotion (crimson; 2nd from bottom), and Negative Emotion (orange; bottom)



Sentiment score- not returning

Sentiment score- returning

Figure 3.3: Sentiment score of activities over the final 12 months for three forums for Depression forums (blue), Relationship Health forums (orange), and Brain/Nervous System Disorder forums (grey)

To investigate this phenomenon, we took the same users from the language analysis and calculated sentiment for all of their posts and replies using the Stanford CoreNLP sentiment analyzer (Socher et al., 2013). The analyzer scores each sentence from 0 to 4, with 0 being extremely negative and 4 being extremely positive. We then average the sentence-level scores for an entire post to assign that post a sentiment score. We hypothesized that these scores may provide some insights towards a user's disengagement in the forum. Unfortunately, after graphing these sentiment scores averaged over all the users for each forums, we could not see any significant change over time, and the lines follow the average score for the respective forums (Depression: 1.68, Relationship Health: 1.70, Brain/Nervous System Disroder: 1.78) (Figure 3.3).

Our final hypothesis was that times users spend in these forums may have some indications of their future engagement– as we have seen that time related features had

	Depression		Relationship		Brain/Nervous	
	Return	Non-return	Return	Non-return	Return	Non-return
AvgInit	114.3	192.5	139.3	420.1	236.3	332.8
AvgMax	218.7	453.3	215.8	492.8	225.2	445.0
AvgMed	1.5	8.8	1.1	15.9	2.7	3.0

Table 3.7: Average initial (AvgInit), maximum (AvgMax), and median (AvgMed) idle time (in days) for users in the forums.

consistently high weight in our previous prediction task. We focused on the idle time of a user in a forum, which is the time passed between two sequential activities. For each forum, we identified all users who posted in at least two different years, and selected 50 random users who were active within the one year preceding the day of data collection, and 50 random users who were not. We then calculated the initial idle time (from account creation to first activity), maximum idle time, and median idle time.

Table 3.7 shows average initial, maximum, and median idle times across the forums. In general, non-returning users wait longer before their first activity, and have larger maximum and median idle times. Depression forum users have smaller initial idle times than Relationship Health or Brain/Nervous System Disorder users, both for returning and non-returning users.

3.3.2 Prediction Task

After finishing all the analyses, we started our prediction task. We used the definition for engagement from our previous task with one slight change- instead of identifying a user

Feature Set	Description
D	User profile demographics: gender and whether a location and/or an avatar image was provided
A	Activity information: number of thread initiations, number of replies posted, number of replies received from others, number of self-replies
T	Timeline information: initial, final, maximum and median idle times
U/B/G	Bag of unigrams/bigrams/1-skip-2-grams from the last post of the observation period
P	Counts of words for each LIWC psycholinguistic class in the last post of the observation period
S	Sentiment score of the last post of the observation period

Table 3.8: List of features used in the Healthboards prediction task

as disengaging if (s)he has not done any activity any time in future (after the observation period), we identified disengaged users if they have spent more time than their maximum idle time after their last post in a forum. Our model also changed a little based on this

definition:

$$m_{\Delta t}(u) = \begin{cases} 1 & \text{if } \exists a \in activities(u): \\ & start(u) + \Delta t < time(a) < start(u) + \Delta t + \max_{a \in activities(u)} time(a) \\ 0 & \text{otherwise} \end{cases}$$

where Δt is the observation period, u is a user, $start(u)$ is the time at which the user u created an account, $activities(u)$ is the set of all activities of user u, $time(a)$ is the time of the activity a. Intuitively, m should predict 0 iff Δt time has elapsed since the user created their account and the user will be inactive in the forum for longer than ever before.

We trained an L2 regularized logistic regression from LibLinear (Fan et al., 2008) using the data collected from the Depression forum and the features described in table 3.8. Throwaway accounts (Leavitt, 2015), defined as accounts with activity levels below the median (2 posts or replies), were excluded from training and testing, though their replies to other users were included for feature extraction. After removing such accounts, 8398 user accounts remained, of which we used 6000 for training our model, and 2398 for testing.

Table 3.9 shows the performance of this model on three different observation periods (1 month, 6 months, 12 months) and different combinations of the feature classes. We did not go beyond 12-month observation period as our previous research suggested that after 12 months we get diminishing returns on our performance measures. The table also shows the performance of a baseline model that predicts that all users will be inactive, the most common classification. We measure performance in terms of accuracy and F_1 (the harmonic mean of precision and recall) on identifying users who withdraw from the forum

Observation Period	1 month		6 months		12 months	
	ACC	F1	ACC	F1	ACC	F1
Baseline	64.7	78.6	80.7	87.3	87.8	92.5
D	65.2	78.7	71,3	82.9	76.8	84.1
A	57.9	66.8	63.0	75.5	66.3	78.8
T	72.0	81.9	82.5	90.2	88.2	93.7
DAT	75.7	83.4	84.4	91.1	89.0	94.0
DATP	75.4	82.9	83.8	90.7	89.0	94.0
DATU	70.4	78.4	84.3	91.0	88.9	94.1
DATB	73.4	81.2	84.4	91.1	88.9	94.0
DATG	71.3	79.3	84.4	91.1	89.0	94.0
DATS	75.6	83.4	84.5	91.2	89.0	94.1

Table 3.9: Accuracy and F-1 scores predicting which users will stop participating in the Depression forum, for different observation periods and different feature sets. Baseline is a classifier that predicts all users as disengaging.

by the end of the observation period. The most predictive features are the timeline (T) features, resulting in F_1 of 93.7 for a 12 month observation period. Though demographic (D) and activity (A) features underperform the baseline alone, adding them to the timeline features (DAT column) yields a better accuracy and a 6% error reduction: 94.0 F_1. The improvement is larger for 1 and 6 month observation periods: 8% and 10% error reductions, respectively.

Adding the language-based features (the DATP, DATU, DATB, DATG columns) does not increase performance despite our findings in language analysis that some phrases were associated with final posts in the forum. Adding sentiment features did not improve the

model (DATS column), and it is consistent with our analysis. This failure of linguistic features may be due to the relatively modest associations; for example, *cut myself* had a PMI of 0.46, and is thus only 38% more likely to show up in a last post than expected by chance. It may also be due to the simplicity of our linguistic features. Consider *Im getting to that rock bottom phase again and im scared*. By PMI, *rock bottom* is not highly associated with last posts, since people often talk about recovering from *rock bottom*. Only present tense *rock bottom* is concerning, but none of our features capture this kind of temporal phenomenon.

3.4 Engagement Network Analysis in Reddit

As we continue our experiments on engagement in thread-and-comment based social media, and gradually focusing more on the relationship of engagement and depression, we decided to apply our knowledge on engagement in the largest of this type of platform– Reddit[5]. We already know timeline features work decently well in engagement prediction, and there are some linguistic cues that can be discovered during observation– and although we can improve upon these features, we wanted to explore another type of features that is not prevalent in Reddit, or any other social media of its kind. These features are generated from interpersonal relationships of users, and are pretty useful in platforms where these are well-structured. For example, in online role-playing games (RPGs), user-level engagement is defined using the number of hours a player has logged in the same game with another player

[5] `www.reddit.com`

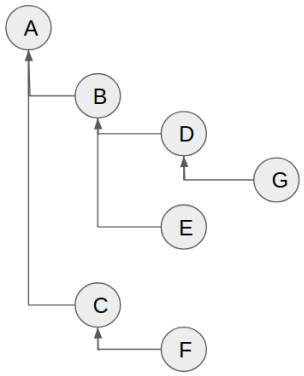


Figure 3.4: User relationship in Reddit

(Kawale, Pal, and Srivastava, 2009). For Twitter, user-level engagement has been identified using replies and retweets (Chen and Pirolli, 2012). Although forms of relationship are not defined by the platform itself, we believe, if we can formalize a structure out of the activities in Reddit, we can create interpersonal relationship graphs for all users in specific subreddits, and will be able to use these graphs to improve engagement prediction.

As our first step towards building interpersonal relationship graphs, we defined one particular relationship among users who contribute to a particular community:

Being-posted-on: A user is said to be *Being-posted-on* by another user if the second user is anywhere below in the same comment chain. This is a directed relationship among users, where the direction of the relationship goes from the poster (user in a bottom level of a comment chain) to the being-posted (user in a higher level).

Figure 3.4 explains this relationship. In this, two users connected using the black line have being-posted-on relationship if they are at different levels in the same chain (arrows indicate directions from poster to posted-on). For example, A is being posted on

by all of the other users, whereas B is being posted on by D, E and G. B is not posted on C as they are in the same level, and B is also not being posted on by F as they are in different comment chain. E, F and G are not being posted on by anyone as they are the last commenters in their particular comment chain.

For our data, we used the reddit dump collected by the Redditor Stuck_in_the_Matrix[6]. (S)he collected the entire comment history of reddit from 2005 (when Reddit was created) till 2015. The history contained 1.7 billion comments and is around 250 Gigabytes compressed (uncompressed, a year itself may go up to a Terabyte) in a single torrent file. The comments are chronologically ordered in JSON files divided in months. The JSON files follow the official Reddit structure, but does not maintain the comment chain hierarchy that we can see when we go to Reddit. This phenomenon made our work particularly hard as we had to recreate this hierarchy. Fortunately, Reddit JSON format preserves the parent id of a comment, which we utilized to recreate the hierarchy. Our main focal point, the *Depression* subreddit, was created back in 2009, and we used all the comments from the first 3 years. Initially, users in this subreddit were extremely irregular (only two posts can be found for the month of February in 2009 where the thread initiator is not being deleted)- but as time went, the subreddit picked up, and is now a community of with more than 350 thousand redditors, with thousands being active at any given time.

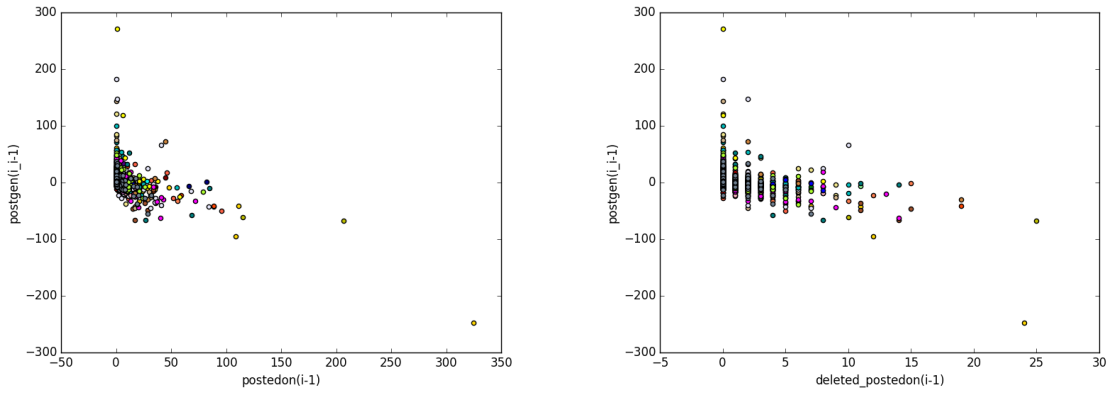
We were fortunate that this data was not completely unstructured- but it was not in a

[6] `https://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment/`

structure that we wanted, as we mentioned before. Creating our preferred structure had one particular difficulty- many users have been removed from this subreddit over time. We did not know why someone was deleted, or whether they deleted their accounts themselves, or were deleted by the administrators of the subreddit. This phenomenon created random holes in our comment-chain-based structure. We filled out these holes by imagining a generic [deleted] user- which is just a placeholder as no information of that user can be retrieved.

We tried to analyze the level of activities (frequency of comments) of a user based on the Being-posted-on relationship of that user in the previous month. We focused on two types of Being-posted-on reciprocation a user has received from other users based on the status of the poster- either an active user, or a deleted one. Our hypothesis was that getting more comments from a user who has been deleted may discourage someone to participate in the platform in the future as we consider deleted users as not useful participants in a healthy community- whether or not they were deleted by an administrator or deleted the account themselves. We were hoping to see an overall general decline in activities from previous month for the deleted-posted-on variable. We specifically looked into the relationships of these variables:

- number of a user being-posted-on in month $i-1$ vs number of comments generated by the user in month i
- number of a user being-posted-on in month $i-1$ vs change in number of comments (actual change and percentage change) generated by the user



(A) All users (B) Deleted users

Figure 3.5: Relationships between being-posted-on in month i-1 (y-axis) and difference of comments generated in months i and i-1 (x-axis)

We observed all these similar relationships for Being-posted-on from deleted users. We plotted these numbers in graphs, and from these plots, we could see that there is a general decline (correlation coefficient: -0.44) in difference in number of activities in month i and number of activities in month i-1 with increase in the number of being-posted-on by other users (Figure 3.5(A)). This is interesting as this is not what we expected- our hypothesis was that there should be a positive relationship between these two variables. We have seen similar decline (correlation coefficient: -0.37) when we observe the relationship of these two variables but using only those being-posted-on by deleted users (Figure 3.5(B)). This is expected as we hypothesized that those who have received reciprocation from users who have not been a useful member of the community may discourage the posted-on user involve in further discussions. We were not able to observe any clear pattern for other

relationships.

3.5 Discussion and Future Works

An interesting future research direction regarding this network-oriented user engagement analysis is to use the features we can obtain from these networks in a prediction model to observe whether we can improve our engagement models further. There are are research works done in the similar field- Ngonmang, Viennet, and Tchuente, 2012 has done some comprehensive research on a French social networking site using user-level network analysis and its effect on user engagement. Massively multiplayer online roleplaying games have also received some attentions regarding inter-user relationship as a predictor for future engagement (Kawale, Pal, and Srivastava, 2009).

We started this research with some specific questions in mind- and we tried our best to answer these questions. We built prediction models, observed features and their contributions over time, we experimented with cross-platform environments- and in the process, we published multiple papers in reputable venues (Sadeque et al., 2015; Sadeque et al., 2016). We believe we have contributed to the development to this particular research area- and we will continue our effort.

CHAPTER 4

INCIVILITY

4.1 Background and Motivations

Online harassment, colloquially known as cyberbullying or cyber harassment has been rampant since the introduction of Internet to the general population. It has been a major cause of concern since the mid- and late-90's, and is a thoroughly researched topic in the fields of social science, behavioral science, network science and computer security. Cyberbullying is a form of harassment that is carried out using electronic modes of communication like computer, phone, and in almost all the cases in recent years, the Internet. Cyberbullying is defined as a "willful and repeated harm inflicted through the medium of electronic text" by (Patchin and Hinduja, 2006)- but this phenomenon goes far beyond the scope of just electronic text. A more comprehensive definition of cyberbullying can be found in one of their later works, where they defined cyberbullying as "a form of harassment using electronic mode of communication" (Hinduja and Patchin, 2008). Fauman, 2008 described cyberbullying as "bullying through the use of technology such as the Internet and cellular phones". Cyberbullying has multiple forms- from online trolling to cyberstalking, even death threats.

Cyberbullying has some distinct characteristics, for example:

Anonymity and/or Physical Distance

In a lot of scenarios cyberbullies are anonymous- they often hide behind a computer screen and take part in bullying. This provides the bullies with a sense of security that they can do whatever they want without any consequence. Even if anonymity is not achieved through electronic communication methods, the distance between the victim and perpetrator provides the bully with the aforementioned sense of security.

Lack of Inhibition

One aspect of anonymity through the veils of electronic communication is that people tend to become less inhibited than they are in a physical confrontation (Fauman, 2008). It is observed that even in cases where anonymity is not achieved, inhibition plays a role through the phenomenon of established physical distance because of the lack of immediate physical ramification towards the perpetrator.

Power Balance

Often times physical bullying requires the bullies being able to get an upper hand in physical confrontation against the victims. This is not true in cyberspace- anonymity and lack of inhibition compensates for the lack of physical power in cyberbullying.

Longer Shelf Life

There is a saying that "nothing is ever deleted from the Internet". This provides us with another aspect of cyber bullying, where a defamatory photo, or a life-destroying text can linger in the cyberspace for an alarming amount of time (Faucher, Jackson, and Cassidy, 2014; Hinduja and Patchin, 2008).

Non-repetitiveness

Physical bullying is marked as a repetitive behavior- a bully performs acts over and over again to intimidate the victim. This is slightly different in the case of cyberbullying because of the longer shelf lives of the acts performed in cyberspace. Fauman, 2008 said, "... aggressive behavior does not need to be repetitive to have desired effect. A single posting of derogatory information about a victim on a web site is sufficient to repeatedly injure that individual, because the information is widely disseminated".

Wide and instant Dissemination

From the quote in the previous subsection, we know that dissemination indeed plays a role in cyberbullying. In the age of social media, bullying is happening more and more in online platforms, and it is easier to disseminate defamatory information about someone in these platforms than in other types of electronic communications like instant messaging or emails. This dissemination is as fast as it is wide- for example, a tweet from someone with a million followers can reach literally tens of millions of Twitter users in minutes.

"No Safe Place"

In the case of physical bullying, victims can (at least, temporarily) escape the bullies in protected environments such as their home. This provides an opportunity (albeit not much) of bullying being reduced. This is not the case of cyberbullying- perpetrators can now virtually infiltrate the homes of victims through the electronic means. In fact, research has shown that victims are more likely to be cyberbullied in their own home (Hinduja and Patchin, 2008).

Delayed Response

Often times cyberbullying can reach a victim later than the action actually being perpetrated (not opening emails immediately, or being offline from a social media site)- which results in a delayed response from a victim. This reduces the chance of eliciting empathy from the bully (Fauman, 2008; Faucher, Jackson, and Cassidy, 2014).

The spectrum of online harassment is vast; hence, we focus on one segment of this phenomenon: online incivility. Incivility has been rampant in American societies for quite some time. Incivility is described as *features of discussion that convey an unnecessarily disrespectful tone toward the discussion forum, its participants, or its topics* (Coe, Kenski, and Rains, 2014). It is often said that incivility is "very much in the eye of the beholder" and what is civil to someone may be uncivil to another (Kenski, Coe, and Rains, 2017), some are is universal nevertheless. A study in 2018 has suggested that 69% of Americans believe that incivility in public discourse has become a rampant problem, and only 6%

do not identify it as a problem (Shandwick, 2018). The average number of incivility encounters per week has also risen drastically in both physical world and cyberspace. Social media encounters are especially alarming: a person who has encountered any form of incivility anywhere, has on average 5.4 uncivil encounters per week in online social media platforms in 2018, which is almost double the amount from late 2016. We can certainly deduce the importance of identifying incivility in online social media platforms from these numbers.

4.1.1 Related Works

The most prominent work on incivility is by Kenski, Coe, and Rains, 2017, where the authors have attempted to establish the difference of perception of incivility among different classes of people. They have looked into five different forms of incivility: name-calling, vulgarity, lying accusation, pejorative and aspersion. They used comments posted by regulars in a newspaper website with apparent incivility that has been annotated by the authors, and their research focused mostly on the demographics and other individual attributes of readers of these comments and how they perceived incivility in these comments. The perception of different forms of incivility among different types of readers was vastly different- name-calling and vulgarity were the most perceived ones, whereas aspersion received the lowest incivility evaluation. The authors hypothesized that name-calling would be the most perceived form of incivility which the experiments supported, but vulgarity being almost as uncivil as name-calling was a surprise. The authors concluded

that the frequency of certain incivility plays a role as less frequent incivility tend to be noticed more.

Another research that worked on incivility and focused more on the perpetrators rather than the readers is (Rains et al., 2017). In this research, the authors attempted to establish a relationship between the commenter's political orientation and his pattern of incivility. They researched a handful of news articles published in the Arizona Daily Star newspaper website and the comments posted on these articles, then manually annotated these comments and their posters for their incivility and political orientation. They used the same five forms of incivility from the previous research, and measured the political orientation of users within a spectrum ranging from liberal to conservative. The authors found out that conservatives were significantly less likely to be uncivil in these public discussions compared to liberals, and the likelihood of liberals being uncivil increased with the presence of conservatives in the same discussion. Liberals were also found to be more repercussive compared to the conservatives. The authors discussed that the reason behind this phenomenon may be its non-normativity of incivility in public discussions and thus commenter's desire for intergroup distinctiveness. Liberals are also found to be more reactive to existing incivility– that may be the reason of conservative users' more tolerance for incivility.

In recent times, there have been a handful of works that have focused on particular forms of incivility- especially on vulgarity and namecalling. Habernal et al., 2018 analyzed ad hominem attacks in *Change My View*- a "good faith" argumentation platform that was is

hosted on Reddit. They have used stacked bidirectional LSTMs and Convolutional Neural Networks to identify ad hominem attacks in that platform, and achieved 78% and 81% accuracy respectively. One of their most interesting finding was that in 48.6% of the cases, ad hominem attacks are in the last comment of the thread, which shows that personal attacks and namecallings can affect user participation in public discourses. Cachola et al., 2018 used vulgarity score for a better sentiment prediction from a collection of 6800 tweets. They found out that vulgarity interacts with key demographic variable like gender, age, religiosity etc. There are other research works that also identified demographic keys that are closely associated with vulgarity: Wang et al., 2014 presented a quantitative analysis on the frequency of curse word usage in Twitter and their variation with certain demographics, and Gauthier et al., 2015 has analyzed the usage of swear words based on Tweeter users' age and gender. None of these papers present any machine learning model that can be used for vulgarity detection though- and Holgate et al., 2018 claim their work to be the first in vulgarity prediction. They have classified functionality of vulgarity in five different cohorts: aggression, emotion expression, emphasis, auxiliary and signalling group identity- and used binary logistic regression classifiers to identify vulgar texts. They also showed the correlation among the demographic variables and the vulgarity functionality and found age, faith, and political ideology have significant correlation with vulgarity usage. They have showed that using these vulgarity features can contribute towards identifying hate speech in social media.

Reynolds, Kontostathis, and Edwards, 2011 developed machine learning models that

can detect cyberbullying by identifying curse and insult words in social media posts. They have collected a small set of posts from a website named *formspring.me* and used various non-sequential learning algorithms on this dataset to build a binary classifier for cyberbullying detection. Waseem and Hovy, 2016 has presented machine learning models that can be used to detect racism and sexism in social media. They have collected and annotated a set of almost 17000 tweets, and used them to build character based n-gram models for offensive tweet detection. They have provided an extensive list of criteria that identify a tweet as racially and sexually offensive, and showed that demographic information does not add much performance to a character-level model. Wulczyn, Thain, and Dixon, 2017 introduced a methodology to generate annotations for personal attacks. They have used crowdsourcing to identify a set of Wikipedia comments, and used a machine learning model to imitate this annotation on a much larger scale. Agrawal and Awekar, 2018 have developed deep neural models that can detect cyberbullying (Reynolds, Kontostathis, and Edwards, 2011), racism/sexism (Waseem and Hovy, 2016), and personal attacks (Wulczyn, Thain, and Dixon, 2017) in multiple social media platforms. They claim that theirs is the first work to systematically analyze cyberbullying in social media towards building deep prediction models. They have shown that hand-crafted features using lexicons is not a good idea as abusive word vocabularies vary a lot from one social media platform to another, and swear words are not always considered to be uncivil in social media. Their neural models outperform traditional non-sequential machine learning models for cyberbullying detection.

Works that closely resemble what we are trying to do have one major issue with the datasets that have been used- they are often annotated by mechanical turks Wulczyn, Thain, and Dixon, 2017; Reynolds, Kontostathis, and Edwards, 2011. Incivility is based on the perception of the person in the receiving end, and this perception varies wildly from person to person. Using turkers that we know almost nothing about is not ideal- as difference in perception may introduce unintended bias in the dataset. Hence, we need a dataset that is annotated by experts who have extensive knowledge on incivility detection. Coe, Kenski, and Rains (2014) presents one such dataset, and we plan to use this for our incivility detection task.

4.1.2 Incivility Classification and Definitions

For our work, we will use the incivility classification presented by Coe et al. in their 2014 paper (Coe, Kenski, and Rains, 2014). In their paper (and also in their followup papers) the authors identified five most rampant incivilities in online interactions:

Namecalling Ad hominem attacks. Although ad hominem attacks are often used to derail a conversation by using derogatory terms towards another person, the authors have included every instances of derogatory remarks, irrespective of target and intention. For example, *At least the morons in the state capital no longer have control of this process!* is identified as an uncivil comment as it has the word *moron* in it (Kenski, Coe, and Rains, 2017).

Vulgarity Contents that include any sort of curse words, including minor ones such as

damn (Kenski, Coe, and Rains, 2017). For example, *I hope the voters will kick that politician out on his pompous ass next election.* is marked as vulgar, as it contains the word *ass* in it.

Lying accusations Contents including charges against someone of being dishonest. For example, *Americans have been screaming at the top of their lungs that this government is wrong, is corrupt, is lying, is deceiving the people, and is violating our constitution.* is a comment that is marked for lying accusation as it contains remarks on American government being corrupt and deceptive.

Pejorative for Speech Contents that are used to mock someone else for their expression or opinion. For example, *Quit crying over spilled milk* indicates that the target of this comment was complaining about something that already happened, and the way it presents itself is considered a mockery of the said target's opinion.

Aspersion Contents that, instead of attacking a person, attacks an idea or a nonhuman entity with derogatory remarks. For example, *Our justice system is just as corrupt and lousy as any in the world* is marked as aspersion as it attacks the justice system instead of a person (which would have been marked as ad hominem, or namecalling).

All these examples are extracted from Kenski, Coe, and Rains, 2017.

4.1.3 Challenges in Identifying incivilities from User Contents

As we have mentioned before that incivility is in the eye of the beholder, it is sometimes challenging to identify what can be unequivocally considered as uncivil interaction. These challenges include:

Frequency

Although researchers have identified incivilities being rampant in public discourse (Shandwick, 2018), it is still minuscule compared to regular civil discourses in any social platform. As most of our identification and prediction techniques are data-driven, it is difficult to create a model that can identify incivilities from this small number of examples.

Linguistic Variations and Creativity

Oftentimes people refrain from using an exact version of uncivil phrase, and use an abbreviation or spelling variation of that said phrase instead. For example, in this sentence *All BS, just like the politicians—the same crap*, the term BS is clearly an abbreviation of the word bullshit- but it is abbreviated, and common in public discourse. Problem is, there are instances in our data where we observed BS is being used to abbreviate something else (a person's name), which clearly is not an example of uncivil comment. Also, people often like to write uncivil words in such spellings that are clearly a derivative form of the said uncivil phrase. For example, people often use *sh!t* instead of *shit*- which clearly are the same thing in a public discourse. These variations are not easily identifiable, as hundreds

of these variations may exist.

Another challenge in identifying incivilities is the never-ending human creativity. Some people can be really creative when they try to attack someone. This often happens when someone tries to indulge in ad hominem with plausible deniability- for example, we have observed people using the words "DemocRat" instead of "Democrat" to identify someone with a democratic political orientation. Although these two words looks similar, and sounds exactly the same, democRat indicates that the target democrat is also a "rat", a colloquial word for a spy, or a dishonest person. This variation comes in other forms too, e.g. democraps- which can also be considered as an ad hominem attack. This phenomenon is sometimes referred as *Obscenity Obfuscation*, and researchers have found that it is becoming increasingly common in user generated contents in all sorts of social media platforms (Rojas-Galeano, 2017).

Difficulty in Comprehension

It is sometimes really difficult to understand whether a word or a phase is used in an uncivil manner without understanding the context. For example, the word "lazy" can be used to describe the state of something that is actually slow or ineffective, or it can be used as an ad hominem attack to someone; e.g. *the lazy politicians have ruined this country*. As understanding the context of a content in a public discourse is really difficult, separating these aforementioned cases based on their contexts becomes really challenging.

Another difficulty in context understanding lies in the definitions of incivilities- as it is

sometimes hard to differentiate between two incivilities. Perfect example this is the case of namecalling vs. aspersion- one is targeted towards a person, the other is towards an idea. It is tough for humans to identify which is which in some cases- hence, it will also present difficulties for the prediction models.

All these challenges have made incivility annotation a difficult task to partake. As we need annotated data to build supervised machine learning models for incivility identification, this shortage of annotated examples is not ideal. We were fortunate enough to obtain an expert-annotated dataset, and will discuss about it in details in the following section.

4.2 Incivility Prediction

As we have seen in the previous section, there are works that defines and analyzes incivility in various platforms, and have presented several machine learning techniques to identify specific incivilities from user generated contents. Most of these works have taken advantage of the users' demographic information obtained from the social media platforms- which is not always available as a large portion of public discourse is anonymous. In this chapter, we are going to focus on our attempt to create a machine learning model that can be used as an incivility filter for moderators in social media platforms. Our model will exclusively use features obtained from the contents and reciprocations in the platform, while avoiding demographic information to facilitate content-based prediction in anonymous forums.

4.2.1 Data Collection and Cleaning

Data was collected from the comment section of the Arizona Daily Star newspaper by Coe, Kenski, and Rains, 2014 and was graciously shared with us for further analysis. The authors selected Arizona Daily Star as their for multiple reasons:

- Arizona Daily Star is the only print daily in the Tucson metropolitan area with well over two hundred thousands readers on a weekday back in 2013.
- It had the same interaction format (log-in requirement, unique screen names, comment rating) with 15 other mid-size regional newspapers the authors have analyzed.
- It provided a conservative amount of incivility present in a newspaper discussion following the January 2011 shooting in Tucson- which sparked the discussions regarding incivility in public discourse.

The authors collected the data between 17 October and 6 November, 2011. Articles and comments were collected from eight news sections- Business, Entertainment, Lifestyles, Local News, Nation and World, Opinion, Sports and State News. All data was downloaded and saved manually by one research assistant one day after the articles were posted to provide enough time for the article to garner comments, yet not long enough for the article to be deleted. At the end of the data collection period, a total of 706 articles and 6535 comments were collected, out of which 6444 were coded for further analysis.

Data Coding

Articles and comments were coded by three teams of 3-5 research assistants, who had extensive training on the coding procedures (Coe, Kenski, and Rains, 2014). The coding process took approximately six weeks, and chance-corrected intercoder reliability was established prior to the coding- which ranged between 0.61 to 1.0 Krippendorff's alpha score for different codes. The coders not only coded the incivilities present in the comments- they also coded a variety of other metadata- e.g. author's name, reactions received for other readers (thumbs up or thumbs down), word counts etc. All the results of the coding procedure were saved in a metadata file created using Microsoft Excel for further use.

Data Cleaning

Although a decent amount of effort has gone into the data processing and coding by Coe et al., it was not ready for computational analysis. Hence, we had to spend a decent amount of time to clean and format the data. The biggest challenge in this process was to retrieve the comment texts from the files- the files were saved as PDFs, and were then given comment IDs by writing numbers manually at the side of the comments in the PDF file in red ink. This number was not written using any common text annotation techniques offered by any common PDF editors, so it was not possible to extract these numbers from the PDF. Because of this, we could not align the comments with the metadata which was keyed by the title of the article and comment number. A screenshot of a page of a PDF file in the dataset can be seen in figure 4.1.

Figure 4.1: Screenshot of a page from a PDF containing comments. Blacked out boxes on the left side of the page includes the name and profile picture of the commenter.

We had to trust our extraction process (explained in detail in the following paragraphs) and its ability to extract the comments in a sequential manner, then we re-numbered the comments. We then manually cross-checked hundreds of these comments and their metadata to be sure that we are not messing anything up. The metadata file had problems of its own. As it was edited by multiple encoders, there were inconsistencies: dates were written in multiple formats in multiple columns, title of the articles were written in multiple ways, there were multiple metadata fields with same attributes and so on.

The comment files were stored in a format which was easy for humans to read, but was not so easy for a computer program to parse. Comments were saved in PDF files as we have mentioned earlier, and then stored in a location like this: *NCID week_of_the_article Numbered/date_of_the_article/section/title_of_the_article*. This naming convention posed a huge problem for us as a lot of them were severely inconsistent. The *title_of_the_article* started with the string CM (representing comments) and then followed by the date (day and month) of the article, the section from which the article was collected, and finally a shorthand form of the article title. The day and month part of the title were not written in a proper format like *ddmm*, so articles written on November 2nd received a date code 112 (11 for November and 2 for the day), whereas articles written on October 26th received a date code 1026 (10 for October and 26 for the day). Fortunately for us we only had articles that were written between October 17 and November 6, so no conflicts occured (a possible conflict could have been 112- is it November 2nd or January 12th?). Also, these dates were written in different formats by different annotators in the metadata file (mm/dd/yy,

mmddyy, mm-dd-yy and so on), which made our task of aligning comments with metadata even harder.

Another inconsistency in the naming convention was that the data collectors created multiple files for one article if that said article had more than 100 comments. For each 100 comments, they created separate files, and gave each file its respective number as a count token (1 for the first 100, 2 for the second and so on). They put this number right after the CM string in the filename- e.g. a filename starting CM11017 means that it contains the first 100 comments of an article that was written on October 17. Unfortunately, the data collectors were massively inconsistent with this convention. They gave the count token to those articles that had more than 100 comments, i.e. an article that had less than 100 comments did not have any count token. Hence, the numeric string followed by the CM string was inconsistent. It could have been five character long (like we have seen in CM 11017), or four characters long (first character for count token, other 3 as the date, or all four are for date), or three character long (only date). Confusions arose when we could not identify whether a length-4 numeric string contains only date or a token count and a date- and solving this problem was not trivial.

The next inconsistencies happened in the *section* part of the filename. In a lot of cases, these section names did not match the sections written in the metadata file- which made our task to match the comments with their metadata even harder. There were eight sections Coe, Kenski, and Rains, 2014 used as their data source, but the names of these sections were written in multiple ways for multiple dates (possibly they were collected by multiple

persons). For example, the section *Nation and World* was mistakenly written as *Nation and News* in multiple locations. Similar things happened with Lifestyles, Local News and State News sections too.

The final inconsistencies in the naming convention came from the shorthand forms of the actual titles of the articles. We believe the data collectors attempted to use the first two words of an article as the shorthand form, which they were not able to maintain in multiple cases. There were spelling errors, usage of more than two words, misrepresenting punctuation marks in the title and so on. This caused problems when we tried to align the comments from the PDF files to the metadata file, and after attempts to solve this problem with rules and regular expressions, we ended up manually renaming all the problem titles with a proper format.

Extracting text from the PDF was not easy. As we have already said, aligning the comments with the metadata was already a big problem- but other problems surfaced when we actually looked inside the texts we have retrieved. Screen name of a commenter and the time of the comment usually appeared before a comment, and we used these two markers as the beginning of a new comment. Unfortunately, some of the screen names spanned more than one line, and was often considered as a content of a comment. Also, the format of how dates were represented was changed during the time of data collection (it moved from "xx hours, yy minutes ago" to a more standard mm-dd-yyyy format)- so there were two ways the times of the comments were represented in the data. We had to resort to regular expression matching to solve these problems. There were other artifacts in the

PDF files (e.g. links to report, thumbs up or down counts, links to next or previous pages, community guidelines etc.) that we had to filter out using regular expressions.

Another problem occurred during the text collection process when we found out that readers can quote other readers' comments. This quotation can be of multiple levels (user 3 can quote user 2, who has already quoted user 1, thus creating a 2-level quote) and happens mostly when two or more persons are involved in a debate. These quotes are easily identifiable by a human eye as they are confined in a text box in the PDF, but is not readily distinguishable from the extracted text. We could identify the starting point of a quote as it always started with a sentence that included *screen_name wrote:* or the word *Quote*, but we could not figured out the end point of it. This was important, as most of these debates were pretty heated and contained a decent amount of incivilities-hence we risked tagging a comment that quoted an uncivil comment without using any incivility in itself as uncivil, or vice versa. We tried using regular expressions and rules to solve this, but the most effective technique was a brute force process, which went through all the previous comments in the comment chain to find a match with that quote, and if found, deleted it from the last comment's body. This process identified almost 100% of the comments with quotes.

All these problems, along with some encoding errors that occurred during the text collection process, forced us to discard some of the articles from the original dataset. After all the cleaning, we ended up with 6175 comments from the original set of 6444 comments, and we were satisfied with the accuracy of our collected data. We then took all the data we

cleaned up and stored them in JSON format, a much more useful format for computational analysis.

4.2.2 Prediction Task

Our main focus was to build a prediction model that can work as a filter for incivility in public discourses. We were also interested in how a model trained on a public discourse data work on a social media platform. As we have mentioned earlier, a lot of effort had gone into the data cleaning and organization process. We first divided our dataset into three smaller sets: train, development and test sets. Comments are randomly assigned sets, and we ended up with 3950 comments in the training set, 989 comments in the validation set and 1236 comments in the test set. We set the the test set aside for our final evaluation, and worked only on the training and validation dataset to find the best model that can fit the problem.

Once we had the data ready for training, we started on our prediction task. For our basic analysis, we used logistic regression on the TF-IDF vectors obtained from the comments. As our problem is not a binary classification problem, we created five one-vs-rest classification models for the five incivility types we had (namecalling, vulgarity, lying accusation, aspersion and pejorative for speech). We created TF-IDF vectors for each comment in our training set using Python's Scikit-learn's TFIDF vectorizer and used these vectors as an input to our logistic regression models. We used these same inputs for a support vector machine (SVM) model.

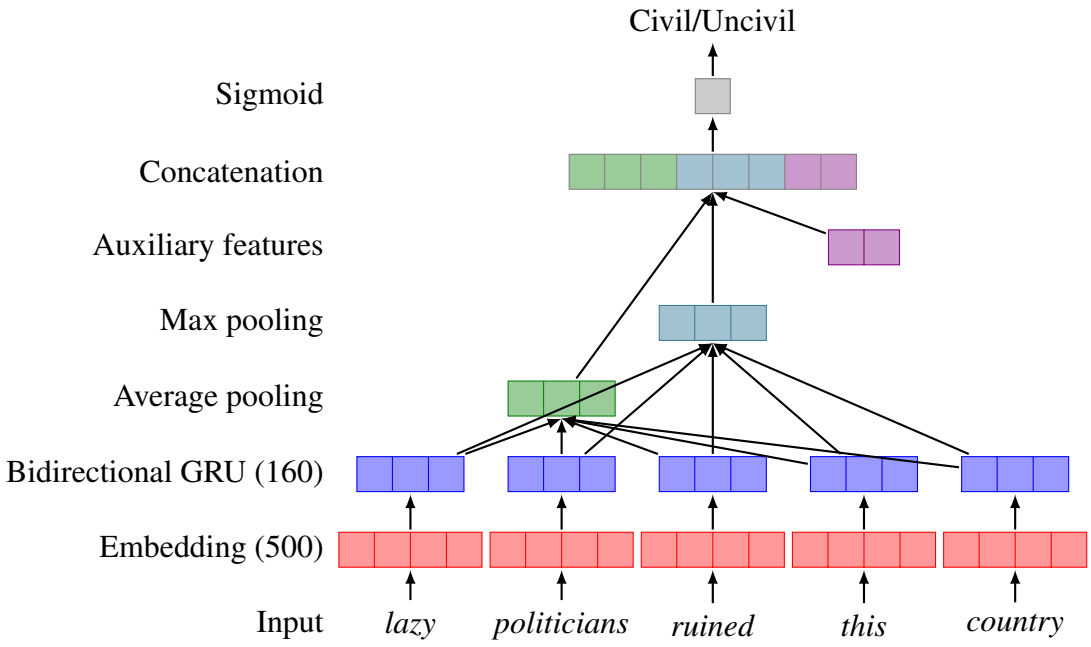


Figure 4.2: General structure of the RNN model. Auxiliary features are optional.

We found a similar task in Kaggle[1] (Srivastava, Khurana, and Tewari, 2018) that tries to identify toxicity of comments in the discourse section of Wikipedia. In that task, the best performing model was a recurrent neural network model with gated recurrent units (GRUs)- but we were skeptical about the performance as non-sequential models (logistic regressions and SVMs) also performed really well in that task- almost as well as the sequential model. Despite our skepticism, we started building sequential models that can fit our task.

For our sequential model, we used recurrent neural networks (RNNs), and gated recurrent units (Cho et al., 2014) for the recurrent layers. We used FastText embeddings (Joulin et al., 2016) to create input vectors for each of the comments in our training data,

[1] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

and fed these vectors to our recurrent layer- which was a bidirectional GRU layer. The outputs of this layer was then fed into a pooling layer, which was a concatenation of an average pooling layer and a max pooling layer (this format of pooling layer worked well for (Demidov, 2018), and also performed well in our preliminary analysis). The output of this layer is then fed through a sigmoid layer, that produced the outputs. To avoid overfitting, we used a dropout layer (Srivastava et al., 2014) with 0.2 probability in between the input and hidden layer. This model was trained with Adam optimizer (Kingma and Ba, 2015) on mini-batches of size 32, with other hyperparameters set to default apart from the maximum length of the input- which was set to 500 words for each comment, as this length garnered the best validation performance in our preliminary analysis. We ran each instance of this model for at most 500 epochs, with the option of early stopping if the validation accuracy did not improve for 10 consecutive epochs. A general structure of this model is shown in figure 4.2. Our first version of this sequential model was an non-class-weighted version of the said model, with 5 sigmoid units in the output layer for five incivility classes. At this point, we moved on from predicting all five incivilities to only the two most common ones- namecalling and vulgarity.

Our first attempt to improve the model was to introduce class weighting. As non-namecalling comments are 7 times more common than the namecalling ones, and non-vulgar comments are 35 times more common than vulgar ones, we introduced a weighting scheme of 1:7 for namecalling and 1:35 for vulgarity. To further improve our model, we wanted to incorporate any metadata that were available to use. From the 2014 paper of

Coe, Kenski, and Rains, 2014, we knew that the thumbs up and thumbs downs received by a comment, the section of the article and author of the article all had some significance regarding incivility in the forum. So we introduced these metadata as features in our model. We created a normalized feature vectors built on these attributes, and introduced them as auxiliary features right before the sigmoid layer, by concatenating them with the output of the pooling layer.

At this point, we wanted to explore external resources that we could use to improve our model. We chose to create a pretrained model on the Kaggle dataset we have mentioned earlier, as it had a large amount of annotated comments (over 160 thousand comments obtained from Wikipedia contributor's community). We used the same RNN model to train on the Kaggle data until it reached convergence, then retrained the model using our Arizona Daily Star data. We had to remove the output sigmoid layer after the pretraining was completed and reintroduced our original sigmoid layer, as the labels are different for the two datasets.

As a final step in our analysis, we wanted to compare our models' performance to a state-of-the-art out-of-the-box text classification model. We selected Flair's text classification model (Akbik, Blythe, and Vollgraf, 2018), which uses GloVe word embeddings (Pennington, Socher, and Manning, 2014) and a couple of character embeddings. We thought character embeddings would be helpful in our task as the linguistic variation and creativity challenges we mentioned earlier are much more likely to be captured by a character model rather than a word embedding model.

Validation								
	Namecalling				Vulgarity			
	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
Logistic Regression	86.33	56.13	11.05	18.46	-	-	-	-
SVM	86.39	54.10	14.89	23.35	-	-	-	-
Unweighted GRU	88.65	59.52	39.07	47.17	97.67	75.00	22.00	34.29
Weighted GRU	84.70	43.65	61.72	51.13	96.05	37.5	66.67	48.00
GRU with Aux features	84.60	44.38	59.85	50.96	96.05	37.5	66.67	48.00
GRU with Pretraining	88.45	69.44	19.53	29.79	97.26	50.00	11.11	18.03
Flair	87.34	52.17	28.12	36.55	96.86	25.00	7.41	11.43
Test								
	Namecalling				Vulgarity			
	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
Weighted GRU	85.81	45.76	50.63	48.07	97.24	48.72	57.57	52.77

Table 4.1: Performance of the sequential models in %. Acc: Accuracy, Prec: Precision, Rec: Recall, F1: F-measure

A snapshot of all our sequential models' performance can be seen in table table 4.1. As we can see, both non-sequential models performed really poorly on the development dataset: both of them failed to identify one single instance of pejorative of speech, aspersion, lying accusation or vulgarity. Logistic regression had a decent precision of 56.13, but a measly recall of 11.05, hence the F-measure is a poor 18.46. The performance of the SVM model was not much better either as precision went down to 54.1, and recall went up to 14.89, resulting in a slightly better F-measure of 23.35.

Our first GRU model performed much better compared to the the non-sequential

models, with F-measure of 47.17 on the validation dataset for namecalling, and 34.29 for vulgarity. Introducing class weighting improved the model further, as F-measure went up to 51.13 for namecalling and 48 for recall. Using auxiliary features had virtually zero effect, with slight improvement on the model's precision but a slight drop in recall for namecalling, and absolutely no change for vulgarity. Retraining our GRU model with a pretrained model trained from the Kaggle dataset performed poorly- almost as badly as the non-sequential models. Flair text classification was not up to the mark either, as it only achieved 36.55 and 11.43 F-measure for namecalling and vulgarity respectively.

At this point, we decided that we will continue working on predicting previously unseen data with our overall most balanced model: the weighted GRU model with FastText embedding. This model performed quite well on the previously unseen test data (48.07 F-measure for namecalling and 52.77 for vulgarity)- and encouraged us to use this on a cross-platform environment.

4.3 Incivility Prediction in Twitter

Our initial goal was to use the model we have created to be effective in a cross platform environment- and as Karan and Šnajder (2018) has showed that cross-domain adaptation for detecting abusive language is possible, we test our model on Twitter, specifically on troll accounts.

In June 2018, The United States House Intelligence Committee released a list of 3841 Twitter account names that were human operated troll accounts associated with Russia's

Internet Research Agency (IRA) (Linvill and Warren, 2018). This was a part of Russia investigation by special counsel Robert Mueller. Darren Linvill and Patrick Warren from Clemson University collected all the tweets published since June 2015 from these accounts, cleaned them, and published a set of almost 3 million of these tweets. These tweets are publicly available in FiveThirtyEight's Github page[2].

As prior research suggest that trolls are a big source of incivility in social media platforms (Fauman, 2008; Hinduja and Patchin, 2008), we took this opportunity to use our model to observe how our model performs on this dataset. We downloaded all the tweet texts and ran our weighted GRU model on these texts. Results of this experiment can be found in the author's GitHub repository[3].

4.3.1 Observations

As we have seen from the predictions our model generated for the tweets, 13% of all tweets are marked as namecalling and 1.7% are marked as vulgarity (compared to 14% and 2.8% respectively in our Arizona Daily Star training data). We do not have an expert annotator who can go through all 3 million of these tweets and tag them for namecalling and vulgarity- hence we could not calculate how our model performed in terms of precision, recall or F-measure. The annotation task is going to be equally costly and time consuming, hence, right now we opted for the analysis of the confidence level of the model on its

[2] https://github.com/fivethirtyeight/russian-troll-tweets

[3] https://github.com/farigys/incivility-in-the-wild/tree/master/outputs/russ_troll_data

prediction for randomly selected tweet texts.

The model did surprisingly well to detect namecalling and vulgarity in terms of confidence score. For example, if the model predicted over 90% on a tweet that it has some form of namecalling or vulgarity in it, we have almost always found it to be correct. We have gone through the top 250 namecalling tweets and top 250 vulgar tweets selected by the model, and we have found only 7 instances of mistakenly tagged namecalling and 5 instances of mistakenly tagged vulgarity. On the other spectrum, the model almost never makes a mistake when the prediction score is below 10%- we found only one instance of mistaken namecalling, and no instance of mistaken vulgarity in the bottom 250 tweets that we manually annotated.

Table 4.2 shows some of the tweets that have been classified by our model. As we have said, the model makes occasional mistakes. For example, the model is confident that there is a namecalling in the tweet that is in the third row of the table, but there is not. Our assumption of that happening is because the terms GOP and POTUS frequently appear with namecalling in our training data, and our model mistakes them as a signal for namecalling. There are some other mistakes that we could observe- e.g. in the fourth example, the model identifies the tweet as namecalling because of the presence of the word *pathetic*, but it is not a namecalling by definition. It is aspersion as it attacks an idea, not an individual. The ambiguity of a word that can be used as both vulgarity and non-vulgarity creates some problems too. For example, the word “hell” in the last example in the table has not been used as a vulgarity- but the word is associated so much with vulgarity that

Namecalling	
Tweet text	Score
RT Jason_toronto: immigrant4trump Delusional Waters, Head Clown Schumer, Joke Perez, Senile Pelosi, Sleazy Schiff	0.997
#IHateItWhen incompetent idiots try to teach us how to live	0.997
@dapsixer GOP POTUS GOPChairwoman Primary these GOP candidates	0.979
#alis Dobbs obliterates Mitch McConnell and his pathetic excuses	0.989
Vulgarity	
Tweet text	Score
Damn #BillCosby !! Damn damn damnnnn	0.996
I'm just going to say it. This is the stupidest tweet I've seen today. This BS bullying is not	0.973
"White Nationalism" WTH came up with this moniker? democrats?	0.985
Hell hath no fury like a bureaucrat scorned	0.969

Table 4.2: Examples from the Twitter vulgarity prediction

it is identified as such. The model can also handle abbreviations- it detects BS (short for *Bullshit*) and WTH (short for *Who the hell*) as vulgarity.

4.4 Discussion and Future Works

The work we have done here has decent significance towards keeping a civil environment in public discourse forums and social media platforms. We tried to build a filtering system that can work alongside human moderators to reduce their workload. This will be objective and independent of user reporting, and will also be capable to perform in a previously unseen environment. There are much work to do in this area- annotation of the troll tweets can show how well the model actually performed, self learning can be used to improve the performance of the model, and so on. We have used word n-grams for features in our baseline models, which can be improved by using features obtained from domain-specific lexicons. There are lexicons of abusive words (Wiegand et al., 2018)- which can be used to create non-sequential models with smaller feature sets. Whether these simpler models are better is yet to be proven - as Agrawal and Awekar (2018) has shown that vocabulary of words used for cyberbullying varies significantly from one social media platform to another. They have also showed that swear words are not necessary to be uncivil in online social media- hence these types of detection techniques should not rely on such hand-crafted features.

A big research question that follows this work is to observe whether incivility affects user engagement in social media. We have seen receiving replies can have effects in user's engagement (Joyce and Kraut, 2006; Sadeque et al., 2015), and the language of these replies can also have some consequences (Arguello et al., 2006). Habernal et al., 2018 has showed that 48% of comments that included ad hominem attacks ends the argument-

which is indicative of lower engagement by the entire community. Hence, we believe that incivility has significant influence on user engagement, and in turn may contribute to a community's sustainability. This is yet to be proven, and more work needs to be performed to prove or disprove this hypothesis.

Incivility detection can be used as a part of user content moderation, and this raises the issue of oppressing freedom of speech. Balance between content moderation and freedom of speech is delicate and is often overlooked. Commercial content moderation still heavily relies on human workers, and the bias introduced by the workers directly contradicts the myth of the Internet being a *site for free, unmediated expression* (Roberts, 2016). The human element in the moderation process is what allows sexist, racist, homophobic contents to persist in public discourse platforms, even though the platform itself disallows them (Roberts, 2014). Introducing an algorithmic model alongside this human element in the content moderation pipeline may seem to resolve the issue to some extent, but concerns are raised due to the bias being induced into the model itself. Supervised machine learning models are often likely to inherit bias from annotators, and ensuring that this bias not being transferred into the model is a challenge (Binns et al., 2017). Conception of offense and perception of incivility is a major contributor towards the bias inheritance (West, 2018), and before we introduce incivility detection models as a component of content moderation pipeline, a lot more research needs to be done in this field- so that moderation can never be used to oppress freedom of expression.

CHAPTER 5

DEPRESSION

5.1 Background and Motivations

In their Global Burden of Disease 2000 study, the World Health Organization estimated that depression is responsible for more than four percent of the Disability-Adjusted Life Years (DALYs) lost and will be the second leading cause of DALYs lost, behind ischaemic heart disease, by 2020 if the trend continues (WHO, 2003). Depression also accounts for 11.9% of all years Lived with Disabilities (YLDs) - the highest among all the mental and neurological conditions - with nearly 350 million people suffering from it worldwide (WHO, 2001). In 2000, depression imposed an annual economic burden of 83 billion dollars in the US - most of which was attributed to reduced productivity and increased medical expenses (PE et al., 2003). Depression is also a major cause of suicide: according to a study by Goodwin and Jamison, 1990, 15-20% of all major depressive disorder patients take their lives. This outcome is largely avoidable if there are proper interventions, and early detection of depression is the first step towards these interventions. Most studies of early detection of depression rely on diagnoses based on patients' self-reported experiences and surveys (Halfin, 2007). The cost of these diagnoses is extremely high, and as of 2009, 30% of world governments who provide primary health care services do not have these

programs (Detels, 2009).

The ubiquity of social media among the world population can provide a solution to this problem. Studies have shown associations between usage of social media and depression (Lin et al., 2016; Primack et al., 2017). Activities in social media can be used as predictors for well-being (Paul and Dredze, 2011) and social participation (Sadeque et al., 2016). Different social media provide different sets of activities that can be leveraged for detection: for example, Moreno et al. found that Facebook status updates of college students could show symptoms of depressive episodes (Moreno et al., 2011), and De Choudhury et al. attempted to predict depression in Twitter using attributes like demographics, language use, engagement, diurnal activity and aggregated behavior over an one-year observation period (De Choudhury et al., 2013). Hu et al. used similar attributes for detecting depression in a popular Chinese microblogging website, Sina Weibo, but with several observation windows ranging from 15 days to 3 months. Leveraging anything other than the contents posted by users in a social media can be challenging as data collected from various social media needs to be properly anonymized to mitigate the risk of reidentification, but the contents themselves can be useful sources of predictors of mental well-being (Coppersmith, Dredze, and Harman, 2014; Schwartz et al., 2016; Wang et al., 2017). Whereas Wang et al. used image and text classification to identify self harm contents in Flickr, Coppersmith et al. and Schwartz et al. relied on only language models to identify mental conditions in Twitter. But a key aspect of detecting depression in social media is the speed of detection: the longer we wait to intervene, the greater the risk of self harm. Hence, predicting depression

early in a user's lifecycle is paramount. This argues for models that don't just look at one snapshot of a user's activities, but instead track the user's activities over time. It also argues for evaluation metrics that consider not only the precision and recall of detecting depressed users, but also the speed of that detection. Our main focus in this research was to address this issue; and we introduced a novel metric named latency-weighted F1 or F_{latency}, for measuring the quality and speed at which a model identifies whether a user is depressed given a series of their social media posts, and showed how it addresses some of the drawbacks of the current state-of-the-art mtric. We propose a general approach for improving the latency of detection models based on checking the consistency of a model's predictions over a risk window.

5.1.1 Related Works

The pioneering work in depression detection was done back in 2013 by De Choudhury et al., 2013. In their paper, they asked Amazon Mechanical Turk users to take the CES-D depression screening test and provide their Twitter handle. They then constructed support vector machine classifiers to distinguish between depressed and non-depressed users. Their models incorporated features such as posts per day, replies per day, shared interactions with other users, use of emotion words from the Linguistic Inquiry and Word Count (LIWC) lexicon, and use of a list of depression-related words mined from Yahoo! Answers Mental Health. Their model achieved almost 70% accuracy, but was not evaluated for the speed at which it could make a prediction.

Wang et al. (Wang et al., 2017) studied images posted in Flickr[1] to identify self harm contents. They started with a set of posts tagged with *selfharm* and *selfinjury*, collected tags that occur more frequently with these two tags and then collected posts tagged with one or more of these tags. They only selected those users with more than five posts with these tags, and then manually examined whether the user is tagged correctly as someone with history of self harm. They used convolutional neural networks to classify images, and took advantage of the image titles for a better prediction result. Their best model achieved a 71 F1 score on the test set. This task was mostly inclined towards identifying users with self harm history, rather than users in risk of future self harm, and early detection was not an issue with the task.

For the 2016 CLPsych shared task (Milne et al., 2016), the mental health forum ReachOut annotated a set of posts with how urgently they needed moderator attention (red/amber/green). Systems competed to take a post of interest and the user's preceding history of posts, and classify the post of interest as red, amber or green. The most successful system in the shared task used various weightings of n-grams: Mac Kim et al., 2016 used TF-IDF weightings of unigrams along with post-level and sentence level embeddings using sent2vec (Le and Mikolov, 2014), whereas Malmasi, Zampieri, and Dras, 2016 went through lexical features like n-grams ranging from 1 to 8 and syntactic features like parts of speech tags and dependencies. Both of these works implemented ensemble classification over sets of simpler classification models, and in both cases the ensemble model came out

[1] `http://www.flickr.com`

as the best. While this shared task considered prediction given a series of social media posts, it did not attempt to evaluate the speed of detection.

A model's speed of detection depends on two forms of latency- observational latency and computational latency. Observational latency represents how many instances (frames, posts etc.) a model needs to observe before manifesting a decision; and computational latency describes the speed of a model's prediction computation. In our research, we focus on this observational latency.Observational latency has been occasionally considered in fields outside of social media analysis. For example, in the field of computer vision, observational latency has been used as a parameter to facilitate early detection of events (Hoai and Torre, 2014; Ellis et al., 2013). Hoai and Torre, 2014 used the number of frames a model requires to detect a facial expression as a parameter for the loss function of their prediction model. Ellis et al. went in a similar direction(Ellis et al., 2013), using Microsoft Kinect data to detect human movement using the minimum number of frames possible. They showed how reducing the number of frames below a certain threshold can adversely affect the accuracy of the detection model.

5.2 CLEF eRisk 2017 Shared task

One problem with predicting depression lies in the annotation process of the data– at least one trained professional who can identify a depressed person from the texts (s)he has written is required for the annotation task. The 2017 CLEF eRisk pilot task– which was dedicated to detect depressed users early in their lifecycle using their social media texts–

had a clever solution to this problem: they identified Reddit users as depressed if they have any sort of self-declaration within their texts (Losada and Crestani, 2016). We participated in this shared task, built our own depression prediction models and observed how these models stack up in an early detection of depression scenario.

The first phase of the task, when the entire text collection of 486 users was released with their user-level annotations of depression, was for the training purpose of the models. The testing stage started two months after that, when the first 10% of texts written by 401 previously unobserved users were released. For the next 9 weeks, new chunks were released, with each chunk including the next 10% of each user's text. After each release, within a week's time, a system had to make one of three decisions for each users: tag the user as depressed, tag the user as non-depressed, or wait to see the next chunk of data. Among these three decisions, the first two are non-reversible- if a user is tagged as depressed or non-depressed by the system, it is not allowed to change the decision even if the next chunks reveal other things. After the release of 10th chunk (end of testing phase), the systems were required to select either depressed or non-depressed for all the remaining undecided users. Models were evaluated based on their precision, recall, and how many chunks they required to detect depressed users. For the last measure, the shared task creators came up with a new metric called Early Risk Detection Error or ERDE. Formally,

ERDE is defined as:

$$ERDE_o(U, sys) = \frac{1}{|U|} \sum_{u \in U} uERDE_o(u, sys)$$

$$uERDE_o(u, sys) = \begin{cases} c_{fp} & \text{if } ref(u){=}- \wedge sys(u){=}+ \\ c_{fn} & \text{if } ref(u){=}+ \wedge sys(u){=}- \\ c_{tp} \cdot \left(1 - \frac{1}{1 + e^{time(sys,u)-o}}\right) & \text{if } ref(u){=}+ \wedge sys(u){=}+ \\ 0 & \text{if } ref(u){=}- \wedge sys(u){=}- \end{cases}$$

where U is the set of users, $ref(u)$ is the reference label ('+' or '-') assigned to the user, $sys(u)$ is the system's earliest non-'?' prediction, $time(sys, u)$ is the time (i.e., number of posts observed) for that earliest prediction, and where o, c_{fp}, c_{fn}, and c_{tp} are parameters of the model that must be set manually.

A number of social media websites were considered as potential data sources for this shared task. Twitter[2] was discarded because it provided little to no context about the user, is highly dynamic and did not allow them to collect more than 3200 tweets per user, which, in 140-character microblogs, represents only a small amount of text. MTV's A Thin Red Line (ATL)[3], a platform designed to empower distressed teens to respond to issues ranging from sexting to cyberbullying, was also considered, but discarded as there were concerns about redistribution and problems regarding obtaining user history. Eventually, Reddit, a

[2] http://www.twitter.com

[3] http://www.athinline.org

social media and news aggregation website, was selected because of its organization of contents among specific subreddits, and the ease of collecting data using the API provided by Reddit itself. The organizers collected the maximum number of submissions they could find from each user and were allowed to download through the API (maximum 2000 posts and comments per user). Users with less than 10 submissions were discarded. Original redditor IDs were replaced with pseudo user IDs for anonymization, and published along with the title, time and text of the posts.

After the data collection, the users were divided into two cohorts: an experimental depressed group and a control (non-depressed) group. For the depressed group, the organizers searched for phrases associated with self-declaration of depression, such as *diagnosed with depression*, and then manually examined the posts to filter down to just those redditors who explicitly said they were diagnosed with depression by a physician. These self declaration posts were omitted from the dataset to avoid making the detection trivial. For the non-depressed group, organizers collected redditors who had participated in depression forums but had no declaration of depression, as well as redditors from other random subreddits. Their final collection contained 531,453 submissions from 892 unique users, of which 486 users were used as training data, and 401 were used as test data. Statistics for that dataset are shown in table 5.1.

We built our models based on two feature sets: depression lexicon and Metamap features. Depression lexicon is a set of unigrams that has high probability of appearing in depression-related posts. The list was collected from (De Choudhury et al., 2013),

	Train		Test	
	Depressed	Control	Depressed	Control
# of subjects	83	403	53	349
# of submissions	30,851	264,172	18,706	217,665
Avg. # of submissions/subject	371.7	655.5	359.7	623.7
Avg. # days from first to last submission	572,7	626.6	608.3	623.2
Avg. # of words per submission	27.6	21.3	26.9	22.5

Table 5.1: Summary of the task data

where the authors compiled a list of words that are most associated with the stem "depress" in the Yahoo! Answers *Mental Health* forum using pointwise mutual information and log-likelihood ratio and kept the top words based on their TF-IDF in Wikipedia articles. We used the top 110 words that were presented in the paper. For each post, we generated a 110 element feature vector: each element contained the counts of how many times that word occurred in the post.

Metamap (Aronson and Lang, 2010) is a highly configurable tool to discover concepts within a text from the Unified Medical Language System (UMLS) Metathesaurus[4]. In our preliminary experiments we found out that Metamap produced a lot of incorrect concept matches in social media texts (as it was mainly built to run on clinical texts), but with some tuning, it was possible to use this effectively on social media. We restricted Metamap to only one source (SNOMEDCT-US) and to only two semantic types (Mental or Behavioral Dysfunction, Clinical Drugs). We passed each post through the restricted Metamap and

[4] `https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/`

collected all the predicted concept unique identifiers (CUIs). We ended up with a set of 404 CUIs. We generated 404 features for each post: the counts of how many times each CUI occurred in the post. Although this task was formed as a sequential classification task, it was possible to consider it as non-sequential as the data was released as chunks, and we could consider one chunk (or a collection of them) as one single discreet activity of a user and decide upon that. Hence, along with a sequential model, we explored much simpler non-sequential ones too.

As per the shared task definition, classifiers were given the user's history in chunks (the first 10% of the user history, then the first 20%, etc.) and after each chunk, the classifiers were asked to make a prediction of "depressed" (+), "not depressed" (-), or "wait" (?). As there was no penalty in identifying non-depressed users later in their lifecycle, we trained all our models to make two-way predictions, "depressed" vs. "wait", and if a classifier predicted a user as depressed after seeing the first $n\%$ of the history, that prediction was considered final and the remaining $100 - n\%$ of the history was ignored. We only predicted "not depressed" for those users who had been predicted as "wait" after the task ended. Note that our models never made post-by-post decisions; they always observed the entirety of the $n\%$ of the history they were given and then made a single prediction for the entire $n\%$.

For our non-sequential model, we used a support vector machine classifier or SVM (figure 5.1). For this model, the feature vectors needed to summarize the entire history of the user. We converted the post-level raw count features to user-level proportion features (e.g., converting the number of times *depression* was used in each post to the proportion of

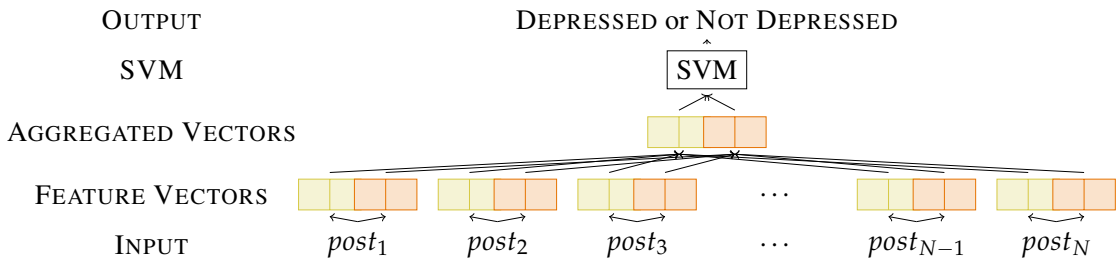


Figure 5.1: General architecture of the non-sequential models for predicting the user's depression status.

all words in a all of a user's posts that were *depression*).

We used two out-of-the-box implementations of support vector machines:

- Weka's implementation of the sequential minimal optimization algorithm (Platt, 1998) for training support vector machine classifiers (Witten and Frank, 1999). The model was set to output probability estimates and it normalizes all attributes by default. Other parameters were set to their defaults. We used a degree-1 polynomial kernel and a cache size of 250007 as it performed better in preliminary experiments on the training data.

- LibSVM's implementation of support vector machines (Chang and Lin, 2011) using C-support vector classification (Boser, Guyon, and Vapnik, 1992). Apart from tuning the model for probability estimate outputs, we used the default parameter settings. We used the radial basis function kernel for this one as it performed better in preliminary experiments on the training data.

Due to the sequential property of the data, we opted for machine learning techniques

that take advantage of this. We used Recurrent Neural Networks (RNN) for our sequential model, which have been successful in other natural language modeling problems (Mikolov et al., 2010). Our RNN was trained to take a sequence of feature vectors, each representing a single post, and predict whether the user is depressed or not. Figure 5.2 shows the general architecture of the model. We used Gated Recurrent Units (GRU) (Cho et al., 2014) to build recurrent layers.

Feature vectors representing each post are first concatenated, and then fed as input to the first recurrent layer. A second GRU layer is stacked on top of the first one for more expressive power, and its output is fed through a sigmoid to produce binary output. To make the experiments with different input features comparable, we fixed the size of the GRU units to 32 for all experiments. To avoid overfitting, we used dropout (Srivastava et al., 2014) with probability 0.2 on the first input-to-hidden layer. Models were trained with RMSProp optimization (Tieleman and Hinton, 2012) on mini-batches of size 200, with all hyperparameters set to default except the learning rate, which was set to 0.001. Each model is trained for at most 800 epochs. The training time for each experiment was around two hours using two Graphics Processing Units (GPUs).

We also implemented an ensemble learning technique using the probability outputs of the nine individual models (3 from Weka, 3 from LibSVM and 3 from RNN: models used as features either the depression lexicon, Metamap outputs, or both). We used 5-fold cross validation for each model to calculate the probability of each user being depressed and then fed these probabilities to a Naive Bayes classifier, which served as the ensemble classifier.

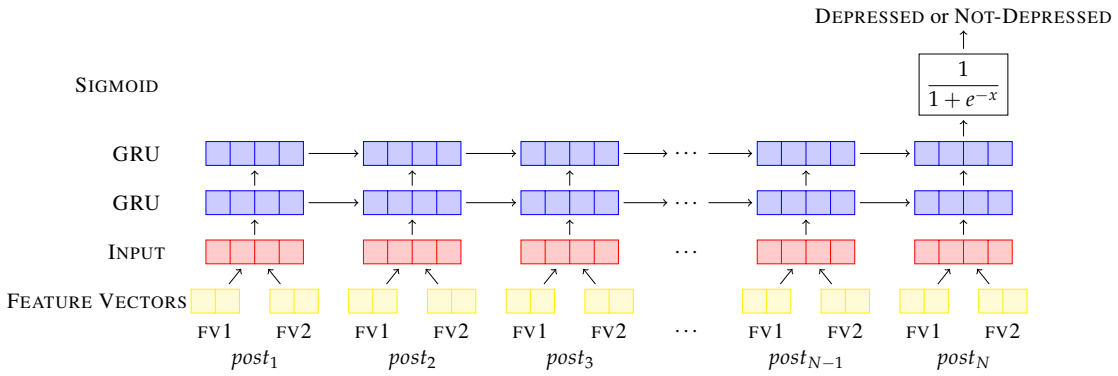


Figure 5.2: Architecture of the model for reading the sequence of a user's posts and predicting the user's depression status.

We used Weka's naive Bayes implementation with the default parameter settings.

We submitted five different models for the task:

UArizonaA An SVM model trained using LibSVM with the depression lexicon and Metamap outputs as features.

UArizonaB An SVM model trained using Weka with the depression lexicon as features.

UArizonaC An RNN model with both the depression lexicon and Metamap outputs as features.

UArizonaD The ensemble model.

UArizonaE An RNN model with the same structure as UArizonaC, but that always predicts "wait" until 60% of the test data is released.

All of these models were selected for their individual properties. UArizonaA was our most restrictive model, as it vigorously tried to not tag someone depressed, whereas UArizonaC

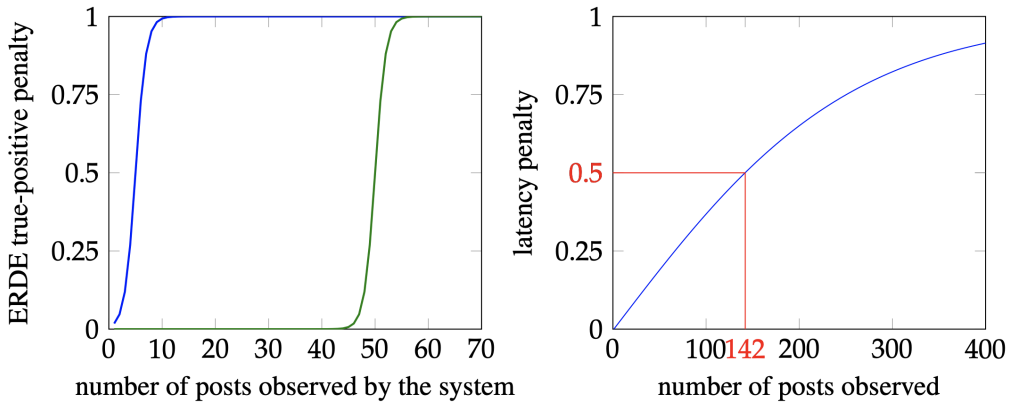
Model	Brief description	E_5	E_{50}	F_1	P	R
UArizonaA	LibSVM + lexicon + UMLS	14.62	12.68	0.40	0.31	0.58
UArizonaB	WekaSVM + lexicon	13.07	11.63	0.30	0.33	0.27
UArizonaC	RNN + lexicon + UMLS	17.93	12.74	0.34	0.21	0.92
UArizonaD	Ensemble	14.73	10.23	0.45	0.32	0.79
UArizonaE	RNN + lexicon + UMLS + 60%-wait	14.93	12.01	0.45	0.34	0.63

Table 5.2: Performance of the models. E_5 and E_{50} are the shared-task-defined Early Risk Detection Error (ERDE) percentages, P is precision, R is recall, and F_1 is the harmonic mean of precision and recall.

was the most open as it tagged more users as depressed than any other models. The other 3 sat in between these 2. To make UArizonaA a little more open towards depression tagging, we combined its 10th chunk output with UArizonaE's 10th chunk output.

The performance of our models are given in table 5.2. The models were evaluated based on 5 performance measures: precision, recall and F_1, and 2 Early Risk Detection Error (ERDE) (Losada and Crestani, 2016) variants, with cutoff parameter o set to 5 and 50 posts. ERDE penalizes systems that take many posts to predict depression. For precision, recall, and F_1, a high score is good, while for ERDE, a low score is good.

Our models were competitive with others in the shared task. UArizonaC ranked 1st out of 30 for recall, UArizonaD ranked 3rd for $ERDE_{50}$, and UArizonaB ranked 4th for $ERDE_5$. For precision and F_1, our models were less impressive; both UArizonaD



(A) Plot of how the ERDE true-positive penalty increases with the number of posts observed by the system, for $o = 5$ (blue) and $o = 50$ (green)

(B) Plot of how the latency penalty increases in our proposal with the number of posts observed by the system

Figure 5.3: ERDE penalty chart vs. our proposed penalty chart

and UArizonaE ranked 11$^{\text{th}}$ for F_1 and UArizonaE ranked 14$^{\text{th}}$ for precision. Overall, UArizonaD is the best of our models: it has the highest F_1, the lowest $ERDE_{50}$, and the second-best recall.

Our models fell short of the best system in the task for two main reasons. First, we attempted to predict depressed users from the beginning, even though the number of posts varies dramatically from user to user (from only 1 post per chunk to over 200 per chunk). A better strategy would have been to start making predictions after observing some threshold n posts, allowing us to predict early for users with many posts, while waiting until we have more information for users with few posts. Second, we did not sufficiently explore the broad range of possible features. For example, we could have built a domain-specific depression lexicon and used it instead of a previously collected lexicon, or we could have used more sophisticated techniques to represent posts as post-level feature vectors.

5.3 Problems with ERDE, and Introduction of Latency and Latency-weighted F1

While participating in the task, we were concerned about how well ERDE actually represent the temporal performance of a model. ERDE penalizes systems that take too long to make a prediction, but since it relies on a standard sigmoid centered at o, the transition between no penalty and 100% penalty is extreme. Figure 5.3(A) shows what the $ERDE_o$ penalty looks like for $o = 5$ and $o = 50$, the two values of o used in the eRisk 2017 evaluation.

With $ERDE_5$, even a perfect system that correctly classified every user after only a single post would be penalized, since $1 - \frac{1}{1+e^{1-5}} > 0$. With $ERDE_{50}$, there is essentially no penalty for a system that takes 45 posts to predict depression, while a system that takes only 10 more posts to predict depression (55 posts) gets essentially no credit at all. We argue that such behavior is undesirable for a measure of speed of detection when, as was the case for eRisk 2017, there is no clear answer to the question "how many posts *should* it take to detect depression?"

ERDE has several additional drawbacks. Beyond the o parameter that we have just discussed, ERDE has 3 other parameters that must be manually set. In eRisk 2017, the organizers defined $c_{fn} = 1$, $c_{fp} = 0.1296$, and $c_{tp} = 1$, but these values were set heuristically, and it is not clear whether such values are appropriate or meaningful for other types of early detection tasks. ERDE is also not easily interpretable. The top system in eRisk 2017 achieved $ERDE_5 = 12.70\%$. Is that fast or slow? How many posts should one expect such a system to take to predict depression? ERDE is unable to answer such questions. Hence, as an alternative to ERDE, we propose a simple, interpretable way of

measuring how long it takes a system to predict a depressed user. We define the *latency* of a system to be the median number of posts that the system observes before making a prediction on a depressed user. Formally:

$$latency(U, sys) = \underset{u \in U \wedge ref(u) = +}{\text{median}} time(sys, u)$$

where, as above, U is the set of users, $ref(u)$ is the reference label ('+' or '-') assigned to the user, and $time(sys, u)$ is the time in number of posts observed for the system's earliest non-'?' prediction. Latency directly answers our earlier question: how many posts should one expect system sys to take to predict depression?

Latency, a measure of speed, should be coupled with measures of accuracy, like precision and recall, to give a complete picture of a system's performance. To produce a single overall measure that combines latency and accuracy, we introduce another metric *latency-weighted F1*. We define latency-weighted F1, or F_{latency}, as the product of a model's F1-measure (the harmonic mean of precision and recall) and the median of a set of penalties in the range [0, 1), which are determined by the model's time to predict each user. The penalty is 0 if a prediction is made after exactly 1 post is observed, and approaches 1 as the number of observed posts increases. Formally, we define:

$$P_{\text{latency}}(u, sys) = -1 + \frac{2}{1 + e^{-p \cdot (time(u, sys) - 1)}}$$

$$F_{\text{latency}}(U, sys) = F_1(U, sys) \cdot \left(1 - \underset{u \in U \wedge ref(u) = +}{\text{median}} P_{\text{latency}}(u, sys)\right)$$

where $F_1(U, sys)$ is the F-measure of the system.

F_{latency} has a single parameter that must be set, p, which defines how quickly the

penalty should increase. We suggest that p should be set such that the latency penalty is 0.5 (i.e., 50%) at the median number of posts of a user. With this approach, p can be determined by fitting the $P_{latency}$ curve to two points: (0, 1) and (0.5, median-posts). In the eRisk 2017 data, the median number of posts of a user is 142, and fitting the $P_{latency}$ curve to (0, 1) and (0.5, 142) results in $p = 0.0078$. Figure 5.3(B) shows a plot of the resulting penalty.

We argue that the shape of this penalty curve is much more appropriate than ERDE for measuring the speed of depression detection: models that predict correctly on the first post are unpenalized, and the penalty gradually increases as the number of posts observed increases. (In the early part of this curve, each post after the first that the model observes applies roughly a 0.5% penalty to F-measure.)

We believe that latency and F_{latency} improve over ERDE by (1) being more interpretable, (2) having fewer parameters that must be manually tuned, and (3) using a penalty that gradually increases with the number of posts observed.

5.3.1 Analysis of Latency-weighted F1

To study how expressive our proposed measure is, we used improved versions of our submitted models, which we created by studying other systems that participated in the task. We replaced the depression lexicon features with our own *depression word features (DepWords)*: count-based features capturing the number of times that unigrams and bigrams commonly associated with depression, e.g., *depressed* or *anxiety*, or *my depression* or

panic attacks, appeared in the text. We first collected two sets of texts, one representing language commonly used when talking about depression, and one representing more general language. For the depression language, we drew the most recent posts (not comments) in the "top" and "hot" section of the *depression* subreddit, resulting in 1987 posts. Posts in this subreddit are not necessarily posted by users who are depressed, but are generally topically related to depression. For the general language, we drew the most recent posts in the "top" and "hot" section of the *textventures* subreddit, resulting in 1082 posts. Posts in this subreddit tell the beginning of a story (which commenters further develop), and cover a wide range of topics. We then used pointwise mutual information (Church and Hanks, 1990) to identify the top unigrams and bigrams most associated with the Depression subreddit. We ended up with 200 features: 100 top unigrams and 100 top bigrams.

We also introduced *Depression embedding features (DepEmbed)*: numeric features from a recurrent neural network that was trained to distinguish between depression-related language and other language. The network treats an entire post as a sequence so that it can capture linguistic phenomena that stretch over many words (e.g., *I just hit rock-bottom*), which cannot be captured by the previous features that treat a post as a bag of n-grams. We use a recurrent neural network in which the words in a post are fed into an embedding layer (128 dimensions), a Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997a) recurrent layer combines this sequence of embedded words into a dense vector (64 dimensions), and the result is fed through a sigmoid layer to produce a binary output. The

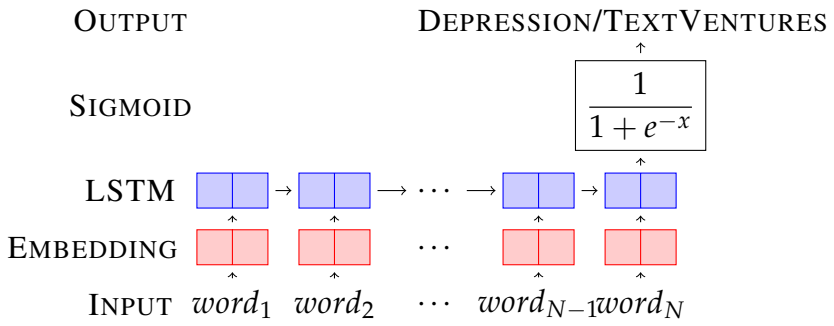


Figure 5.4: Architecture for training a model that can semantically summarize the contents of a post as a dense vector.

architecture is shown in figure 5.4. We train this model on the depression/textventures data from the DepWords features, asking the model to classify whether a post is from the depression subreddit or the textventures subreddit. We use an Adam optimizer (Kingma and Ba, 2015) for training and dropout (Srivastava et al., 2014) with probability 0.15 to avoid overfitting. Once the model is trained, we discard the sigmoid layer, run the model on the posts from the eRisk 2017 data, and use the dense vector produced by the LSTM layer as the features. We thus ended up with 64 features: the 64 dimensions of the model's LSTM layer.

Risk Window

In preliminary analysis of the models above, we found it was common for a model to make occasional mistakes. But recall that in early depression prediction, the first "+" or "-" prediction is considered final, so occasional mistakes will force an early detection model to abort entirely, even if they have seen only a small number of posts so far. This can have

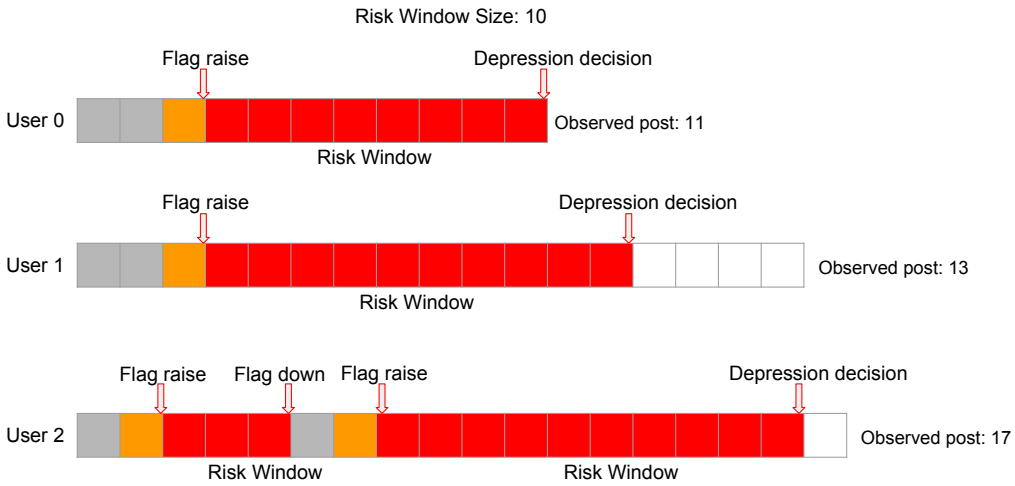


Figure 5.5: Example of post-by-post depression prediction with a risk window of size 10. Each block represents 1 post: gray is observed, orange is where the flag was raised, red is in the risk window, and white is unobserved. User 0 is an example where there are fewer remaining posts than the risk window, and user 2 is an example of restarting after a broken streak.

a significant impact on their performance.

We thus introduce a technique, which can apply generally to any model, that trades off between latency and precision. If the model makes a prediction that the user is depressed after post p (we refer to this as *raising the flag*), we only confirm that prediction if the model continues to make the same (depressed) prediction for the next n posts (we refer to this as the *risk window*), or, if the user has fewer than n posts remaining, continues to make the same (depressed) prediction for all of their remaining posts. Figure 5.5 demonstrates the process with a risk window of size 10.

Our two model architectures (non-sequential and sequential) and four feature sets

Model	Features	Precision	Recall	F_1
SVM	Words	53.3	38.6	44.8
SVM	DepWords	**77.3**	20.5	32.4
SVM	DepWords+Metamap	49.0	30.1	37.3
SVM	DepEmbed	69.4	30.1	42.0
SVM	DepEmbed+DepWords	66.7	45.8	54.3
SVM	DepEmbed+DepWords+Metamap	53.9	66.3	59.5
GRU	Words	72.8	51.8	60.6
GRU	DepWords	62.0	68.7	65.1
GRU	DepWords+Metamap	67.0	**75.9**	**71.2**
GRU	DepEmbed	65.8	60.2	62.9
GRU	DepEmbed+DepWords	60.0	61.4	60.7
GRU	DepEmbed+DepWords+Metamap	60.0	62.7	61.2

Table 5.3: Comparison of different models and feature sets in five-fold cross-validations on the training set when considering the entire posting history (window=∞).

(Words, DepWords, DepEmbed, and MetaMap) can be combined to create a large number of models. We used five-fold cross-validations on the training data to explore which model/feature combinations look most promising, so that those can be evaluated on the test set. We focused on the simpler setup where the model observes a user's entire posting history (window=∞), and is evaluated just in terms of precision and recall.

Table 5.3 shows the cross-validation performance of a variety of models on the training data.

The best F_1, 71.2, is achieved by the sequential (GRU) model with depression words (DepWords) and UMLS medical concept (MetaMap) features. Comparing across types of models, the sequential models are the clear winners: even the worst sequential (GRU)

model had a higher F_1 than the best non-sequential (SVM) model (60.6 vs. 59.5). This finding is intuitive, given that early detection is a sequential prediction problem. Comparing across types of features, adding medical concepts (MetaMap) always improved F_1, but the results for other types of features were more mixed. Depression embeddings (DepEmbed) always improved the non-sequential (SVM) models, but always hurt the sequential (GRU) models. And using all words (Words) was better than just the depression words (DepWords) for the non-sequential (SVM) model, but the reverse was true for the sequential (GRU) model.

Looking across all the models, we selected two models for evaluation on the test set: the best non-sequential (SVM) model (DepEmbed+DepWords+ Metamap) and the best sequential (GRU) model (DepWords+Metamap). For each of these models, we apply a risk window, considering all possible risk windows between 0 and the maximum number of posts, and optimizing the window size to maximize cross-validation F_{latency} on the training set. For the SVM model, an 11-post risk window yields the highest F_{latency} (67.1, with an F_1 of 82.0), while for the GRU model, a 23-post risk window yields the highest F_{latency} (52.6, with an F_1 of 65.7).

Table 5.4 evaluates the best models on the eRisk 2017 test set. For contrast, we also show each model with a risk window of 0 (i.e., the first '+' or '-' prediction is final) and a risk window of ∞ (i.e., the model always waits for all of a user's posts and decides at the final post).

Comparing ERDE to F_{latency}, we see that F_{latency} is better at discriminating between

Model	Risk window	$ERDE_5$	$ERDE_{50}$	F_1	Latency	F_{latency}
SVM	0	13.1	9.7	51.3	63.5	38.9
SVM	11 (best)	13.6	10.1	51.4	75	36.8
SVM	∞	13.2	11.7	45.4	199	16.0
GRU	0	12.5	9.4	33.5	9	32.3
GRU	23 (best)	15.2	11.5	44.4	69.5	32.7
GRU	∞	15.0	13.6	45.0	199	15.8

Table 5.4: Comparison of the top non-sequential and sequential models (SVM: DepEmbed + DepWords + Metamap and GRU: DepWords + Metamap) on the test set. For contrast, the same models are also shown with risk windows of 0 and ∞.

models. For example, the non-sequential (SVM) and sequential (GRU) models with risk window 0 have given very similar values for ERDE, with their $ERDE_5$s differing by only 0.6 points and their $ERDE_{50}$s differing by only 0.3 points. Yet these two models have hugely different performance characteristics: the GRU is extremely fast (latency 9) at a significant cost to accuracy (F_1 of 33.5), while the SVM is much more cautious (latency 63.5) and much more accurate (F_1 of 51.3). Table 5.4 also shows the challenge of setting the ERDE o parameter: with $o = 5$ as in eRisk 2017, ERDE can't distinguish (only a 0.1 point difference) between a non-sequential (SVM) model that sees a median of 63.5 posts (window=0) and one that sees a median of 199 posts (window=∞), despite the latter being much, much slower to make predictions. We see these empirical results as a strong indication that F_{latency} better captures the important evaluation characteristics of early detection problems.

We found that the models with risk windows optimized on the training set (SVM:window=11 and GRU:window=23) did not always outperform other simple choices of risk window (window=0 or window=∞) on the test set. While the 23-window GRU model indeed outperformed the F_{latency} of the other GRUs (GRU:0 and GRU:∞), the 11-window SVM model did not have a better F_{latency} than the 0-window SVM; the tiny improvement in F_1 achieved by SVM:11 over SVM:0 was outweighed by its larger jump in latency.

Despite the training set results where sequential models substantially out-performed non-sequential models, on the test set the no-risk-window non-sequential (SVM) model outperformed all sequential (GRU) models, in terms of both F_{latency} and F_1. But note that on the training set, we compared systems with access to the entire posting history (window=∞), and, as can be seen in table 5.4, the performance of the SVM model is much worse with such a large risk window. Probably the simple way that the non-sequential model aggregates feature vectors makes it easy to lose the signal of a single depressed post in a sea of many non-depressed posts.

5.4 Discussion and Future Works

$F_{latency}$ combines F1 and latency under the assumption that systems generally want to optimize F1. However different applications may need to optimize different evaluation measures. For example, if the goal is to have a human intervene when a risk of depression is detected in a social media user, then probably a high recall even at the expense of precision

would be preferred, so that the human would be able to intervene wherever possible. On the other hand, if the goal is to have an automatic intervention when a depression risk is detected, then probably a high precision is needed so that the automatic intervention is only applied when the model is very certain of the depression risk. Future work may need to extend $F_{latency}$ to such scenarios. $F_{latency}$ is a general metric, applicable to any problem where systems must examine a sequence of items associated with an object, and make a prediction about that object's class as rapidly as possible. In this research, we only explored $F_{latency}$ as applied to early detection of depression on social media. Future work will need to investigate the utility of $F_{latency}$ on other kinds of problems: detecting drug discontinuation, churn prediction, etc.

As we started this research, there were not many research groups that were working on detecting depression in social media. Now there are research groups around the world who are investing their efforts in this research. We have published multiple papers in premiere venues as one of the early proponents in this field (Sadeque, Xu, and Bethard, 2017; Sadeque, Xu, and Bethard, 2018), and we believe we have had some significant contributions- as other researchers have started to use our idea of the combination of observational latency and performance accuracy (Trotzek, Koitka, and Friedrich, 2018a; Trotzek, Koitka, and Friedrich, 2018b).

CHAPTER 6

CONCLUSION

In this research, we have attempted to predict user behavior in social media platforms. We focused on user engagement in social networks, whether they maintain a civil environment and whether their state of mental health is revealed in these platforms. For all of these tasks, we followed similar procedures- we collected data, we analyzed the contents, we explored the feature space, and built prediction models that can predict these behaviors. We experimented with various machine learning techniques- we used regressions, support vector machines, neural networks, ensemble models and so on. We used data from a wide range of social media platforms- we used data from support groups, tweets, Reddit posts, newspaper comments- and we could establish that it is possible for the machine learning models we have built to cross the boundary of the domain and be useful in a completely different platform. We have contributed to the innovations of how performance of machine learning models is measured, and have come up with more expressive and more comprehensive performance metrics.

While this dissertation tells the stories of the successes we had during the course of the research, it also shares the disappointments and failures we had to endure. Not every step was the right step, not every hypothesis was proven correct, not every model we built crushed the problem. We acknowledged our mistakes, learned from our failures and moved

on with our research. This entire research was a massive learning process- and I believe I have learned from successes we had as well as from failures we endured.

There are still a lot of open-ended questions that require attention- how does an uncivil environment affect user engagement, whether user-user interaction networks provide insights to an individual's continued participation, whether we can generalize our $F_{latency}$ to machine learning problems that are not specific to depression detection, and so on. Given enough time, I would have loved to answer all these questions, and I will continue trying to at least address them as I go on with my career. My dream is that this thesis, and the papers we have published during the course of this thesis, will encourage other researchers to put their best efforts in these fields and improve upon our work. For me, peer recognition is the ultimate form of appreciation, and as long as these publications help others, I will be happy.

Bibliography

Agrawal, Sweta and Amit Awekar (2018). "Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms". In: *CoRR* abs/1801.06482. arXiv: 1801.06482. URL: http://arxiv.org/abs/1801.06482.

Akbik, Alan, Duncan Blythe, and Roland Vollgraf (2018). "Contextual String Embeddings for Sequence Labeling". In: *COLING 2018, 27th International Conference on Computational Linguistics*, pp. 1638–1649.

Arguello, Jaime et al. (2006). "Talk to Me: Foundations for Successful Individual-group Interactions in Online Communities". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '06. Montréal, Québec, Canada: ACM, pp. 959–968. ISBN: 1-59593-372-7. DOI: 10.1145/1124772.1124916. URL: http://doi.acm.org/10.1145/1124772.1124916.

Aronson, Alan R and Francois-Michel Lang (2010). "An overview of MetaMap: historical perspective and recent advances". In: *Journal of the American Medical Informatics Association 17(3)*, pp. 229–236. DOI: 10.1136/jamia.2009.002733.

Binns, Reuben et al. (2017). "Like Trainer, Like Bot? Inheritance of Bias in Algorithmic Content Moderation". In: *Social Informatics*. Ed. by Giovanni Luca Ciampaglia, Afra Mashhadi, and Taha Yasseri. Cham: Springer International Publishing, pp. 405–415. ISBN: 978-3-319-67256-4.

Boser, Bernhard E, Isabelle M Guyon, and Vladimir N Vapnik (1992). "A training algorithm for optimal margin classifiers". In: *Proceedings of the fifth annual workshop on Computational learning theory*. ACM, pp. 144–152.

Burez, Jonathan and Dirk Van den Poel (2009). "Handling class imbalance in customer churn prediction". In: *Expert Systems with Applications* 36.3, pp. 4626–4636.

Cachola, Isabel et al. (2018). "Expressively vulgar: The socio-dynamics of vulgarity and its effects on sentiment analysis in social media". In: *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 2927–2938.

Chang, Chih-Chung and Chih-Jen Lin (2011). "LIBSVM: A library for support vector machines". In: *ACM TIST* 2.3, 27:1–27:27. DOI: 10.1145/1961189.1961199. URL: http://doi.acm.org/10.1145/1961189.1961199.

Chen, Jilin and Peter Pirolli (2012). "Why You Are More Engaged: Factors Influencing Twitter Engagement in Occupy Wall Street". In: *Sixth International AAAI Conference on Weblogs and Social Media*.

Cho, Kyunghyun et al. (2014). "On the Properties of Neural Machine Translation: Encoder-Decoder Approaches". In: *Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8)*.

Church, Kenneth Ward and Patrick Hanks (1990). "Word Association Norms, Mutual Information, and Lexicography". In: *Computational Linguistics* 16.1, pp. 22–29.

Coe, Kevin, Kate Kenski, and Stephen A. Rains (2014). "Online and Uncivil? Patterns and Determinants of Incivility in Newspaper Website Comments". In: *Journal of Communication* 64.4, pp. 658–679. DOI: 10.1111/jcom.12104. eprint: /oup/backfile/content_public/journal/joc/64/4/10.1111_jcom.12104/2/jjnlcom0658.pdf. URL: http://dx.doi.org/10.1111/jcom.12104.

Constant, David, Lee Sproull, and Sara Kiesler (1996). "The kindness of strangers: The usefulness of electronic weak ties for technical advice". In: *Organization science* 7.2, pp. 119–135.

Coppersmith, Glen, Mark Dredze, and Craig Harman (2014). "Quantifying Mental Health Signals in Twitter". In: *Associational for Computational Linguistics Workshop of Computational Linguistics and Clinical Psychology*.

Danescu-Niculescu-Mizil, Cristian et al. (2013). "No Country for Old Members: User Lifecycle and Linguistic Change in Online Communities". In: *Proceedings of the*

22Nd International Conference on World Wide Web. WWW '13. Rio de Janeiro, Brazil: International World Wide Web Conferences Steering Committee, pp. 307–318. ISBN: 978-1-4503-2035-1. URL: http://dl.acm.org/citation.cfm?id=2488388.2488416.

Dasgupta, Koustuv et al. (2008). "Social ties and their relevance to churn in mobile telecom networks". In: *Proceedings of the 11th international conference on Extending database technology: Advances in database technology*. ACM, pp. 668–677.

De Choudhury, Munmun et al. (2013). "Predicting Depression via Social Media." In: *ICWSM*, p. 2.

Demidov, Vladimir (2018). *Kernel Submission for Kaggle Toxic Classification Challenge*. https://www.kaggle.com/yekenot/pooled-gru-fasttext?. Last Accessed: 2018-12-02.

Detels, Roger (2009). *The Scope and Concerns of Public Health*. New York: Oxford University Press Inc.

Ellis, Chris et al. (2013). "Exploring the Trade-off Between Accuracy and Observational Latency in Action Recognition". In: *Int. J. Comput. Vision* 101.3, pp. 420–436. ISSN: 0920-5691. DOI: 10.1007/s11263-012-0550-7. URL: http://dx.doi.org/10.1007/s11263-012-0550-7.

Fan, Rong-En et al. (2008). "LIBLINEAR: A library for large linear classification". In: *The Journal of Machine Learning Research* 9, pp. 1871–1874.

Faucher, Chantal, Margaret Jackson, and Wanda Cassidy (2014). "Cyberbullying among University Students: Gendered Experiences, Impacts, and Perspectives". In: *Education Research International* 2014. DOI: 10.1155/2014/698545.

Fauman, Michael A. (2008). "Cyber Bullying: Bullying in the Digital Age". In: *American Journal of Psychiatry* 165.6, pp. 780–781. DOI: 10.1176/appi.ajp.2008.08020226.

Gauthier, Michael et al. (2015). "Text Mining and Twitter to Analyze British Swearing Habits". In:

Gerpott, Torsten J, Wolfgang Rams, and Andreas Schindler (2001). "Customer retention, loyalty, and satisfaction in the German mobile cellular telecommunications market". In: *Telecommunications Policy* 25.4, pp. 249 –269. ISSN: 0308-5961. DOI: http://dx.doi.org/10.1016/S0308-5961(00)00097-5. URL: http://www.sciencedirect.com/science/article/pii/S0308596100000975.

Goodwin, Frederick K. and Kay Redfield Jamison (1990). *Manic-Depressive Illness: Bipolar Disorder and Recurring Depression*. New York: Oxford University Press Inc.

Gustafsson, Anders, Michael D Johnson, and Inger Roos (2005). "The effects of customer satisfaction, relationship commitment dimensions, and triggers on customer retention". In: *Journal of Marketing*, pp. 210–218.

Habernal, Ivan et al. (2018). "Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation". In: *arXiv preprint arXiv:1802.06613*.

Hadden, John et al. (2007). "Computer assisted customer churn management: State-of-the-art and future trends". In: *Computers & Operations Research* 34.10, pp. 2902–2917.

Halfin, Aron (2007). "Depression: the benefits of early and appropriate treatment". In: *The American journal of managed care* 13.4 Suppl, S92—7. ISSN: 1088-0224. URL: http://europepmc.org/abstract/MED/18041868.

Hamilton, William L et al. (2017). "Loyalty in Online Communities". In: *arXiv preprint arXiv:1703.03386*.

Hinduja, Sameer and Justin W. Patchin (2008). "Cyberbullying: An Exploratory Analysis of Factors Related to Offending and Victimization". In: *Deviant Behavior* 29.2, pp. 129–156. DOI: 10.1080/01639620701457816. eprint: https://doi.org/10.1080/01639620701457816. URL: https://doi.org/10.1080/01639620701457816.

Hoai, Minh and Fernando De la Torre (2014). "Max-Margin Early Event Detectors". In: *International Journal of Computer Vision* 107.2, pp. 191–202. ISSN: 1573-1405. DOI:

10.1007/s11263-013-0683-3. URL: https://doi.org/10.1007/s11263-013-0683-3.

Hochreiter, Sepp and Jürgen Schmidhuber (1997a). "Long short-term memory". In: *Neural computation* 9.8, pp. 1735–1780.

Hochreiter, Sepp and Jürgen Schmidhuber (1997b). "Long Short-Term Memory". In: *Neural Computation* 9.8, pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735. eprint: https://doi.org/10.1162/neco.1997.9.8.1735. URL: https://doi.org/10.1162/neco.1997.9.8.1735.

Holgate, Eric et al. (2018). "Why Swear? Analyzing and Inferring the Intentions of Vulgar Expressions". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4405–4414.

Joulin, Armand et al. (2016). "Bag of Tricks for Efficient Text Classification". In: *CoRR* abs/1607.01759. arXiv: 1607.01759. URL: http://arxiv.org/abs/1607.01759.

Joyce, Elisabeth and Robert E. Kraut (2006). "Predicting Continued Participation in Newsgroups". In: *Journal of Computer-Mediated Communication* 11.3, pp. 723–747. ISSN: 1083-6101. DOI: 10.1111/j.1083-6101.2006.00033.x. URL: http://dx.doi.org/10.1111/j.1083-6101.2006.00033.x.

Karan, Mladen and Jan Šnajder (2018). "Cross-Domain Detection of Abusive Language Online". In: *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. Brussels, Belgium: Association for Computational Linguistics, pp. 132–137. URL: https://www.aclweb.org/anthology/W18-5117.

Karnstedt, Marcel et al. (2011). "The Effect of User Features on Churn in Social Networks". In: *Proceedings of the 3rd International Web Science Conference*. WebSci '11. Koblenz, Germany: ACM, 23:1–23:8. ISBN: 978-1-4503-0855-7. DOI: 10.1145/2527031.2527051. URL: http://doi.acm.org/10.1145/2527031.2527051.

Katz, S. (1987). "Estimation of probabilities from sparse data for the language model component of a speech recognizer". In: *Acoustics, Speech and Signal Processing, IEEE*

Transactions on 35.3, pp. 400–401. ISSN: 0096-3518. DOI: 10.1109/TASSP.1987.1165125.

Kawale, J., A. Pal, and J. Srivastava (2009). "Churn Prediction in MMORPGs: A Social Influence Based Approach". In: *Computational Science and Engineering, 2009. CSE '09. International Conference on*. Vol. 4, pp. 423–428. DOI: 10.1109/CSE.2009.80.

Keaveney, Susan M. (1995). "Customer Switching Behavior in Service Industries: An Exploratory Study". English. In: *Journal of Marketing* 59.2, pp. 71–82. ISSN: 00222429. URL: http://www.jstor.org/stable/1252074.

Kenski, Kate, Kevin Coe, and Stephen A. Rains (2017). "Perceptions of Uncivil Discourse Online: An Examination of Types and Predictors". In: *Communication Research* 0.0, p. 0093650217699933. DOI: 10.1177/0093650217699933. eprint: https://doi.org/10.1177/0093650217699933. URL: https://doi.org/10.1177/0093650217699933.

Kim, Hee-Su and Choong-Han Yoon (2004). "Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market". In: *Telecommunications Policy* 28.9–10, pp. 751 –765. ISSN: 0308-5961. DOI: http://dx.doi.org/10.1016/j.telpol.2004.05.013. URL: http://www.sciencedirect.com/science/article/pii/S0308596104000783.

Kingma, Diederik and Jimmy Ba (2015). "Adam: A method for stochastic optimization". In: International Conference on Learning Representation.

Le, Quoc V. and Tomas Mikolov (2014). "Distributed Representations of Sentences and Documents". In: *CoRR* abs/1405.4053. URL: http://arxiv.org/abs/1405.4053.

Leavitt, Alex (2015). ""This is a Throwaway Account": Temporary Technical Identities and Perceptions of Anonymity in a Massive Online Community". In: *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. CSCW '15. Vancouver, BC, Canada: ACM, pp. 317–327. ISBN: 978-1-

4503-2922-4. DOI: 10.1145/2675133.2675175. URL: http://doi.acm.org/10.1145/2675133.2675175.

Lin, Liu yi et al. (2016). "Association between Social Media Use and Depression among U.S. Young Adults". In: *Depression and Anxiety, 33(4)*, pp. 323–331. DOI: 10.1002/da.22466.

Linvill, Darren L. and Patrcik L. Warren (2018). *Troll Factories: The Internet Research Agency and State-Sponsored Agenda Building*. https://www.rcmediafreedom.eu/Publications/Academic-sources/Troll-Factories-The-Internet-Research-Agency-and-State-Sponsored-Agenda-Building.

Losada, David E and Fabio Crestani (2016). "A Test Collection for Research on Depression and Language Use". In: *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer International Publishing, pp. 28–39.

Mac Kim, Sunghwan et al. (2016). "Data61-CSIRO systems at the CLPsych 2016 Shared Task." In: *CLPsych@ HLT-NAACL*, pp. 128–132.

Mahmud, Jalal, Jilin Chen, and Jeffrey Nichols (2014). "Why Are You More Engaged? Predicting Social Engagement from Word Use". In: *CoRR* abs/1402.6690. URL: http://arxiv.org/abs/1402.6690.

Malmasi, Shervin, Marcos Zampieri, and Mark Dras (2016). "Predicting Post Severity in Mental Health Forums". In: *The 3rd Workshop on Computational Linguistics and Clinical Psychology*, pp. 133–137.

Manning, Christopher et al. (2014). "The Stanford CoreNLP Natural Language Processing Toolkit". In: *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Baltimore, Maryland: Association for Computational Linguistics, pp. 55–60. URL: http://www.aclweb.org/anthology/P14-5010.

Mikolov, Tomas et al. (2010). "Recurrent neural network based language model." In: *Interspeech*. Vol. 2, p. 3.

Milne, David N. et al. (2016). "CLPsych 2016 Shared Task: Triaging content in online peer-support forums". In: *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*. San Diego, CA, USA: Association for Computational Linguistics, pp. 118–127. URL: http://www.aclweb.org/anthology/W16-0312.

Milosevic, Miloš, Nenad Zivic, and Igor Andjelkovic (2017). "Early churn prediction with personalized targeting in mobile social games". In: *Expert Systems with Applications* 83, pp. 326 –332. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2017.04.056. URL: http://www.sciencedirect.com/science/article/pii/S0957417417303044.

Moreno, MA et al. (2011). "Feeling bad on Facebook: depression disclosures by college students on a social networking site". In: *Depression and Anxiety 28(6)*, pp. 447–455.

Mozer, Michael et al. (1999). "Churn Reduction in the Wireless Industry." In: *NIPS*, pp. 935–941.

Ngonmang, Blaise, Emmanuel Viennet, and Maurice Tchuente (2012). "Churn Prediction in a Real Online Social Network Using Local CommunIty Analysis". In: *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*. ASONAM '12. Washington, DC, USA: IEEE Computer Society, pp. 282–288. ISBN: 978-0-7695-4799-2. DOI: 10.1109/ASONAM.2012.55. URL: http://dx.doi.org/10.1109/ASONAM.2012.55.

Olah, Christopher (2015). *Understanding LSTM Networks*. http://colah.github.io/posts/2015-08-Understanding-LSTMs/. Last Accessed: 2019-02-04.

Patchin, Justin W. and Sameer Hinduja (2006). "Bullies Move Beyond the Schoolyard: A Preliminary Look at Cyberbullying". In: *Youth Violence and Juvenile Justice* 4.2, pp. 148–169. DOI: 10.1177/1541204006286288. eprint: https://doi.org/10.1177/1541204006286288. URL: https://doi.org/10.1177/1541204006286288.

Paul, Michael J and Mark Dredze (2011). "You are what you Tweet: Analyzing Twitter for public health." In: *Icwsm* 20, pp. 265–272.

PE, Greenberg et al. (2003). "The economic burden of depression in the United States: how did it change between 1990 and 2000?" In: *J Clin Psychiatry 64(12)*, pp. 1465–1475.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning (2014). "GloVe: Global Vectors for Word Representation". In: *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543. URL: http://www.aclweb.org/anthology/D14-1162.

Platt, J. (1998). "Fast Training of Support Vector Machines using Sequential Minimal Optimization". In: *Advances in Kernel Methods - Support Vector Learning*. Ed. by B. Schoelkopf, C. Burges, and A. Smola. MIT Press. URL: http://research.microsoft.com/\~jplatt/smo.html.

Primack, Brian A. et al. (2017). "Use of multiple social media platforms and symptoms of depression and anxiety: A nationally-representative study among U.S. young adults". In: *Computers in Human Behavior* 69, pp. 1 –9. ISSN: 0747-5632. DOI: http://doi.org/10.1016/j.chb.2016.11.013. URL: http://www.sciencedirect.com/science/article/pii/S0747563216307543.

Rains, Stephen A. et al. (2017). "Incivility and Political Identity on the Internet: Intergroup Factors as Predictors of Incivility in Discussions of News Online". In: *Journal of Computer-Mediated Communication* 22.4, pp. 163–178. DOI: 10.1111/jcc4.12191. eprint: /oup/backfile/content_public/journal/jcmc/22/4/10.1111_jcc4.12191/2/jjcmcom0163.pdf. URL: http://dx.doi.org/10.1111/jcc4.12191.

Reynolds, Kelly, April Kontostathis, and Lynne Edwards (2011). "Using machine learning to detect cyberbullying". In: *2011 10th International Conference on Machine learning and applications and workshops*. Vol. 2. IEEE, pp. 241–244.

Rheingold, Howard (2000). "The Virtual Community: Homesteading on the Electronic Frontier". In:

Roberts, Sarah T. (2014). *Behind the screen: the hidden digital labor of commercial content moderation*. University of Illinois at Urbana-Champaign.

— (2016). *Commercial Content Moderation: Digital Laborers' Dirty Work*. Peter Lang Publishing.

Rojas-Galeano, Sergio (2017). "On Obstructing Obscenity Obfuscation". In: *ACM Trans. Web* 11.2, 12:1–12:24. ISSN: 1559-1131. DOI: 10.1145/3032963. URL: http://doi.acm.org/10.1145/3032963.

Sadeque, Farig, Dongfang Xu, and Steven Bethard (2017). "UArizona at the CLEF eRisk 2017 Pilot Task: Linear and Recurrent Models for Early Depression Detection". In:

— (2018). "Measuring the Latency of Depression Detection in Social Media". In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. WSDM '18. Marina Del Rey, CA, USA: ACM, pp. 495–503. ISBN: 978-1-4503-5581-0. DOI: 10.1145/3159652.3159725. URL: http://doi.acm.org/10.1145/3159652.3159725.

Sadeque, Farig et al. (2015). "Predicting Continued Participation in Online Health Forums". In: *SIXTH INTERNATIONAL WORKSHOP ON HEALTH TEXT MINING AND INFORMATION ANALYSIS (LOUHI)*, p. 12.

Sadeque, Farig et al. (2016). "Why do they leave: Modeling participation in online depression forums". In: *Proceedings of the 4th Workshop on Natural Language Processing and Social Media*, pp. 14–19.

Schwartz, H ANDREW et al. (2016). "Predicting individual well-being through the language of social media". In: *Biocomputing 2016: Proceedings of the Pacific Symposium*, pp. 516–527.

Shandwick, Weber (2018). *Civility in America 2018: Civility at Work and in our Public Squares*. https://www.webershandwick.com/wp-content/uploads/2018/06/Civility-in-America-VII-FINAL.pdf. Last Accessed: 2018-06-11.

Sinha, Tanmay et al. (2014). "Capturing "attrition intensifying" structural traits from didactic interaction sequences of MOOC learners". In: *Proceedings of the 2014 Empirical Methods in Natural Language Processing Workshop on Modeling Large Scale Social Interaction in Massively Open Online Courses*.

Socher, Richard et al. (2013). "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank". In: *EMNLP*.

Srivastava, Nitish et al. (2014). "Dropout: a simple way to prevent neural networks from overfitting." In: *Journal of Machine Learning Research* 15.1, pp. 1929–1958.

Srivastava, Saurabh, Prerna Khurana, and Vartika Tewari (2018). "Identifying Aggression and Toxicity in Comments using Capsule Network". In: *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pp. 98–105.

Tieleman, Tijmen and Geoffrey Hinton (2012). "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude". In: *COURSERA: Neural networks for machine learning* 4.2.

Trotzek, Marcel, Sven Koitka, and Christoph M. Friedrich (2018a). "Early Detection of Depression Based on Linguistic Metadata Augmented Classifiers Revisited". In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Ed. by Patrice Bellot et al. Cham: Springer International Publishing, pp. 191–202. ISBN: 978-3-319-98932-7.

— (2018b). "Utilizing Neural Networks and Linguistic Metadata for Early Detection of Depression Indications in Text Sequences". In: *CoRR* abs/1804.07000. arXiv: 1804.07000. URL: http://arxiv.org/abs/1804.07000.

W3C (2015). *Activity Streams 2.0: Working Draft 15 December 2015*. Ed. by James M Snell and Evan Prodromou. http://www.w3.org/TR/2015/WD-activitystreams-core-20151215/.

Wang, Wenbo et al. (2014). "Cursing in English on Twitter". In: *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*. CSCW '14. Baltimore, Maryland, USA: ACM, pp. 415–425. ISBN: 978-1-4503-2540-

0. DOI: 10.1145/2531602.2531734. URL: http://doi.acm.org/10.1145/2531602.2531734.

Wang, Yilin et al. (2017). "Understanding and Discovering Deliberate Self-harm Content in Social Media". In: *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, pp. 93–102.

Waseem, Zeerak and Dirk Hovy (2016). "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter". In: *Proceedings of the NAACL Student Research Workshop*. San Diego, California: Association for Computational Linguistics, pp. 88–93. DOI: 10.18653/v1/N16-2013. URL: https://www.aclweb.org/anthology/N16-2013.

West, Sarah Myers (2018). "Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms". In: *New Media & Society* 20.11, pp. 4366–4383. DOI: 10.1177/1461444818773059. eprint: https://doi.org/10.1177/1461444818773059. URL: https://doi.org/10.1177/1461444818773059.

WHO, World Health Organization (2001). *The world health report 2001- Mental Health: New Understanding, New Hope*. http://www.who.int/whr/2001/en/whr01_en.pdf?ua=1. Last Accessed: 2016-04-02.

— (2003). *Global Burden of Disease (GBD) 2000: version 3 estimates*. http://www.who.int/entity/healthinfo/gbdwhoregionyld2000v3.xls?ua=1. Last Accessed: 2016-04-08.

Wiegand, Michael et al. (2018). "Inducing a Lexicon of Abusive Words – a Feature-Based Approach". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 1046–1056. DOI: 10.18653/v1/N18-1095. URL: https://www.aclweb.org/anthology/N18-1095.

Witten, Ian H and Eibe Frank (1999). *Data mining: practical machine learning tools and techniques with Java implementations*.

Wulczyn, Ellery, Nithum Thain, and Lucas Dixon (2017). "Ex machina: Personal attacks seen at scale". In: *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, pp. 1391–1399.