A FRAMEWORK FOR TRANSPARENT DEPRESSION CLASSIFICATION IN COLLEGE SETTINGS VIA MINING INTERNET USAGE PATTERNS

by

RAGHAVENDRA KOTIKALAPUDI

A THESIS

Presented to the Faculty of the Graduate School of

MISSOURI UNIVERSITY OF SCIENCE AND TECHNOLOGY

In Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE IN COMPUTER SCIENCE

2011

Approved by

Dr. Sriram Chellappan, Advisor
Dr. Donald Wunsch, Co-advisor
Dr. Sanjay Madria

ABSTRACT

Depression is a serious mental health problem affecting the mind and body of people in our society. As of today, depression affects about 5 to 7% of the American population, which roughly translates to 14 million people! A particularly worrying concern these days is the increasing incidence of depression among college students. The problem is so severe that many US campuses have dedicated centers with mental health professionals catering to students needs. Although there are very effective treatments for depression, studies report that ∼80% of depressed students do not seek any help due to the lack of perception of the problem, low self esteem, and loneliness.

The aim of this thesis is to develop a transparent depression identification/classification framework that can be deployed in college settings. The premise of this thesis stems from surveys in the mental health community revealing extensive known correlations between Internet use and Depression. The thesis investigates the feasibility of identifying and classifying depression via mining real Internet usage patterns derived from Cisco Netflow data.

To the best of the authors knowledge, this thesis is the first to study depression using *real* Internet data. As a result, several new statistical results correlating Internet use with depression is presented. Additionally, this thesis takes a quantum leap forward in transparent depression detection by developing a classifier that can predict depression with ∼74% accuracy, demonstrating the feasibility of the approach. The performance was further improved to ∼84% by considering deviations in baseline Internet activity of a user.

The strength of the proposed framework lies in its practicality, extensibility, transparency, and privacy preserving nature.

ACKNOWLEDGMENT

My first thanks and immense gratitude goes to my advisor *Dr. Sriram Chellappan* for giving me the freedom to pursue my own interests and for trusting me before I thought I could do any research at all. He has always encouraged me to live intensively and taught me how to appreciate good scientific work and build on it. During my work with him, I had the opportunity to meet and collaborate with several interesting people, and had a jolly time doing so. The idea for this thesis emerged through a series of dialogues with my advisor. I would like to thank him for suggesting such an interesting problem, mentoring, and guiding with several aspects of this thesis.

I would also like to thank my co-advisor *Dr. Donald Wunsch* for introducing me to the wonderful world of *clustering* and *reinforcement learning*. I have learned a great deal about machine learning from him, thanks to his broad knowledge and deep intuitions about where things should go and what is necessary to get there. I would also like to thank *Dr. Sanjay Madria* for willing to be on my thesis committee.

This thesis would not be possible without the generous help of *Dr. Karl Lutzen* for coordinating and helping me with netflow data acquisition, *Levy Mallot* for helping me with initial scripting and coding, and *Dr. Frances Montgomery* for conducting surveys and for all the insightful discussions. Though not involved with my thesis directly, I am greatly indebted to *Dr. Daniel Tauritz*, particularly for his wonderful course "*Advanced Topics in AI*", which in many ways was instrumental in improving my writing skills, and for making me a better researcher.

It was a pleasure to share Masters studies and life with wonderful people like *Anup*, *Avinash*, *Veeresh*, and *Makarand*, who are very close friends now. I specially wish to thank *Veeresh* and *Makarand* for making *chai* regularly ☺, *Aanchal Shah* for literally providing me with *khana aur makan*, and (soon to be Dr.)*Neelanjana Dutta* for helping me cope with the last minute chaos and assistance with thesis slides. I also wish to acknowledge some of my friends for reviewing my thesis.

I saved the last for people closest to my heart – my family. Without their constant encouragement and support, I wouldn't be in this position today. I specially wish to acknowledge my sister for encouraging me to do my masters. If it wasn't for her, I would have missed out on this amazing experience.

TABLE OF CONTENTS

APPENDICES

LIST OF ILLUSTRATIONS

Figure Page

LIST OF TABLES

1. INTRODUCTION

The Chapter begins by stating the problem description and motivation for conducting this research. This is followed by the main research question(s) and a very brief outline of the proposed research approach. The chapter closes with an outline of this thesis along with the major research contributions.

1.1. PROBLEM DESCRIPTION

Depression is a serious mental health problem affecting the mind and body of people in our society. It affects about 5 to 7% of the American population, which roughly translates to 14 million people! [63, 9] Depression not only impacts personal aspects of everyday life such as eating, sleeping, working, and relationships, but also exacts an economic cost of over $83 billion each year in lost productivity and increased medical expenses.

A particularly worrying concern today is the increasing incidence of depression among college students. In the US alone, surveys have shown that 14.3% of female and 7.3% of male students have been diabolized with depression [63], which is much higher when compared to the national averages. The problem is so severe that many US campuses have dedicated centers with mental health professionals catering to students needs. Although there are very effective treatments for depression, studies indicate that over two-thirds of depressed people do not seek treatment simply because they do not recognize the symptoms until the damage is done. In most cases, depressed people tend to keep to themselves and are extremely reluctant to seek help [75]. In fact, more than 80% of depressed college students do not seek any help at all [30, 8].

If left undetected and untreated, depression can cause severe health impacts like loss of appetite, sleep disorders, chronic fatigue, anxiety and panic attacks, which could ultimately lead to an increased propensity for suicide and violence among students. Depressed students also suffer from poor academics, decreased work place performance and high drop out rates, all of which have severe debilitating impacts to the overall academic environment in our colleges. The problem is so severe that depression has been identified by a variety of studies as being among the leading causes for death amongst the US college population [50, 63, 35]. **Therefore, there is a critical need to detect depression in a transparent and simple manner.**

1.2. INTERNET USAGE AS A MARKER FOR DEPRESSION

In order to identify depression, the first step is to look at attributes that are known to be correlated with depression. While this opens up a plethora of possibilities, the list can be narrowed down by focusing on attributes that can be monitored *transparently* and are *available in college settings*. Internet usage is one such possibility. This section begins with the motivation as to why Internet usage was chosen. This is followed up by a brief literature review of interesting correlations between Internet usage and depression.

1.2.1. Motivation. There are several reasons for choosing Internet usage over other attributes:

- **Availability:** College students extensively use the Internet. In fact, it is expected that more than 95% of college students use it actively [5, 4], making it a widely available feature.

- **Extensive known correlations:** There is a vast body of psychology literature correlating internet usage activity with depression. Not only is depression linked to internet use, but other factors such as academic performance, loneliness, eating disorders etc., which are correlated with depression are also known to be correlated with internet usage (a detailed description is provided in section 1.2.2). Therefore, internet usage implicitly accounts for a vast number of other attributes that are correlated with depression and can, therefore, prove to be instrumental in classifying depression.

- **Transparent and negligible deployment costs:** Internet usage can easily be monitored in college settings. In fact, most universities typically have an on-campus IT department that already logs internet flow level information for identifying attacks, p2p downloads etc., thereby alleviating the need for special infrastructure.

- **Privacy preserving:** Compared to other behavioral attributes, Internet usage remains as the *only* attribute that can easily be monitored in a hassle free, transparent and privacy preserving manner[1]

[1] Privacy preserving aspect is detailed in chapter 2

1.2.2. Correlations between Internet use, Depression and related ancillary factors. While the benefits of Internet use for academic learning, research and social networking are well known, several recent studies are exploring downsides of Internet use among college students. The foundations for such explorations mainly stems from the fact that the Internet provides an alternative platform for students to replace their otherwise normal activities in the real world by an online virtual world that is secretive, and where repercussions are not immediate. Under such situations, it is increasingly likely for an already vulnerable student body to be easily victimized. Recent studies show that increased Internet use leads to loneliness and social isolation in the real world [59, 16], which are known to increase the formation of mal-adaptive attitudes particularly among adolescents.

Studies conducted in [16, 52, 77] demonstrate that students with excessive Internet use have poor academic performance, and higher drop out rates. More seriously, excessive Internet use has also been linked to increased Obesity and eating disorders [69, 77], Cyberstalking [44] and even harassment [84] among college students, all of which exhibit a positive correlation with depression.

In the last decade, one issue that has received a considerable amount of attention is the impact of Internet usage among college students on their Mental Health. A pioneering study by Young and Rogers in 1998 [86] demonstrated that students with depressive symptoms used the Internet five times more likely for non-academic purposes than those without symptoms, which is also validated by a 2006 study conducted by Campbell et al. [23]. Another study by Morgan and Cotton also concluded that the type of activity engaged on the Internet influenced various levels of depression among college students. When Internet was utilized for non-communication oriented activities such as shopping, their study showed that levels of depressive symptoms among students tended to increase [60].

More recently, studies conducted by Lam and Peng [54], and by Morrison and Gore at the University of Leeds [31] also conclude that teenagers who spend significant amounts of time online are much more likely to be depressed than casual users. Researchers found striking evidence that students are developing compulsive Internet habits, whereby they replace real-life social interactions with online chat rooms and social networking sites, leading to increased isolation and anxiety. Their claim is further validated by Stevens and Morris' study [73].

Recent studies on excess online gambling by students [53, 39, 65], and frequent visits to health websites [21] have also shown to be indicators of depressive behavior. From the perspective of usage time, studies in [7, 37, 86] show an increasing incidence of depression among subjects who are are active online at late nights, and are consequently sleep deprived and fatigued.

From the aforementioned literature review, it is clear that there are many aspects of internet usage that are *correlated* with depression. Whether or not internet usage *causes* depression is debatable; it is, however, immaterial to this study as this thesis focuses on *predicting* depressive behavior based on internet usage patterns. Figure 1.1 summarizes various correlations exhibited between depression, internet usage and related ancillary factors. In summary, it indicates how four basic Internet attributes, namely, Category of Internet use, Repeat visits to same sites, Excessive Internet usage duration and Night use have debilitating impacts to various related aspects of the mental health of college students, which indirectly correlate with depressive behavior.

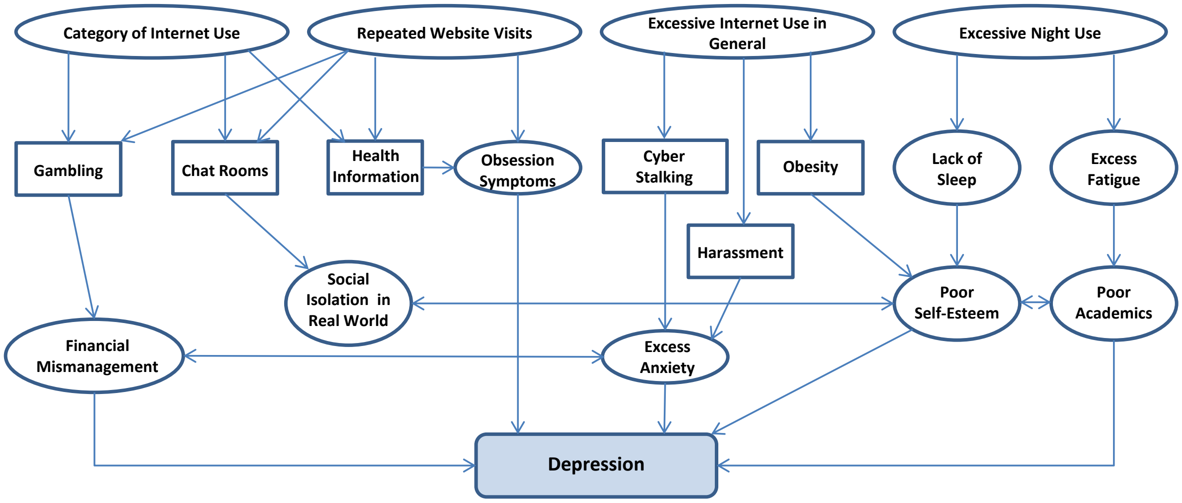


Figure 1.1. Correlations between Internet Use, Depression and related ancillary factors in College Students

1.3. RESEARCH QUESTION AND MAJOR CONTRIBUTIONS

In light of reasons correlations described in section 1.2, the main research question addressed by this thesis is as follows:

> *Is it possible to detect and classify depression based on internet usage activity in a transparent and privacy preserving manner?*

While there is a lot of existing work on correlations between internet use and depression, to the best of the authors knowledge, an attempt to identify and classify depression based on internet use has never been done before. Considering the fact that depression is an extremely complex mental health disorder, coupled with the lack of related research, this is not an easy task. The difficulty is further exacerbated by the fact that sensitive internet attributes such as email messages, keystrokes etc., cannot be used due to privacy concerns. This thesis, therefore, presents an original work with the following as its major contributions:

- **Privacy Preserved Depression classification:** As stated earlier, it is critical to use Internet attributes that preserve the privacy of a user. By using Cisco Netflow data (internet flow level statistics) as a representative of users internet activity, a classification accuracy of ∼73% was achieved, demonstrating the feasibility of the approach. The accuracy was further improved to ∼85% by considering netflow baseline deviation patterns. As netflow data represents flow level statistics that are already monitored by most universities on-campus IT department, user privacy is not compromised.

- **Deeper Understanding of Internet and Depression:** While there is existing work on Internet usage and Depression, they are based on surveys and are susceptible to self reporting bias. For example, it is not very hard to imagine why people might be reluctant to report gambling usage. To the best of the authors knowledge, this study is the first to use *actual* internet data. While some existing hypothesis were validated, several new correlations were discovered. In particular, this is the first work to uncover correlations between Internet usage entropy and depression, which is previously undocumented in psychology literature.

- **Foundational work for identifying other disorders:** The work lays a foundation for understanding the relationship between Mental Health and Internet use in an increasingly cyber networked world. The author believes that the developed framework can be applied to study and detect other disorders like Anorexia, Bulimia, ADHD etc., all of which are shown to be correlated with Internet use [73, 85, 46].

- **Positive Impact on College Life:** Early detection/intervention for depressed students will lead to enhanced student productivity, minimizing drop out rates. The immediate future work is to collaboratively use results from this project to also design tutorials for all our students on both healthy and un-healthy Internet usage, and its impacts on mental health.

1.4. THESIS ORGANIZATION

Excluding this chapter, the remainder of the thesis is organized in the form of five chapters.

Chapter 2 describes the research methodology along with the necessary infrastructure and processes used to obtain relevant data. In particular, details on participants, survey process, and acquisition of internet usage data is described. Internet usage is obtained in the form of Cisco Netflow data by collaborating with Missouri S&T on-campus IT department. Privacy and related concerns are also addressed in this chapter.

Chapter 3 describes netflow preprocessing and feature extraction process in detail. In particular, three levels of features are extracted, each with increasing level of information content. The chapter closes by presenting a detailed account of various statistical insights obtained.

Chapter 4 is solely devoted to the main research task of building a depression classifier. First, the research task is formalized to simplify discussion. The chapter begins by describing a new similarity metric for computing the distance between netflow data sets. The latter portions of this chapter describes the mathematical similarities with other problem domains. By leveraging on these similarities, some methods are directly applied while others are specialized for the task of depression classification. The chapter concludes by presenting performance metrics such as the accuracy, precision and recall for the proposed classification models.

Chapter 5 builds on insights gained in Chapter 4 and describes the development of an improved classifier that is able to predict the likelihood of increase or decrease in depression by examining baseline deviation patterns in netflow data. Additionally, new statistical results are also presented in this chapter.

Finally, Chapter 6 concludes the thesis by summarizing all the major results. In addition, several potential approaches are described for improving the classification accuracy, intended as a guide to future researchers who wish to extend and build on this thesis.

2. RESEARCH METHODOLOGY AND SETUP

This chapter begins with a description on what exactly constitutes the "Internet usage". This is followed by a brief description of the proposed research methodology along with the detailed description of the necessary research infrastructure and data collection process. Privacy and related concerns are addressed towards the end of the chapter.

2.1. CISCO NETFLOW™AS INTERNET USAGE DATA

Internet usage is a very broad term. Part of the challenge is to identify internet attributes that *do not* compromise a persons privacy. As an example, consider email messages. One approach for classifying depression is to look for words like "unhappy", "disappointed", "gloomy" etc. in email messages. Based on the word frequencies, it is reasonable to assume that a probabilistic classifier can be built based on word frequencies. One can even go to the extent of identifying words that separates a set of depressed people with non-depressed people by monitoring email messages of the participants. However, even if such a classifier is successful, it is unlikely to be adopted by universities due to privacy concerns. Similarly, the possibility of monitoring keystrokes, instant messages etc., are also ruled out.

In this thesis, Cisco Netflow™data was used to characterize the internet usage of an individual. Very briefly, Cisco Netflow technology [1] is a proprietary protocol for collecting IP traffic information and is widely accepted as the de-facto standard. The netflow data consists of several "flows". Cisco defines a flow as follows [2]:

> *A flow is identified as a unidirectional stream of packets between a given source and destination - both defined by a network-layer IP address and transport-layer source and destination port numbers. Specifically, a flow is identified as the combination of the following seven key fields:*
>
> - *Source IP address*
> - *Destination IP address*
> - *Source port number*
> - *Destination port number*

- *Layer 3 protocol type*
- *ToS byte*
- *Input logical interface (ifIndex)*

One of the immediate benefits of using netflow is with regard to privacy. As flows represent traffic level information, a fair amount of privacy is ensured. In some sense, it can be considered as the equivalent of the calling information of a phone conversation, not the conversation *per se*. However, one critical question remains: *"Is flow level information sufficient to classify depression?"*.

While there is no previous research in this direction, the use of netflow is motivated by its recent success in building robust spam classifiers [71, 33] and Intrusion detection systems [89, 64, 36, 12]. If one can identify spam and network intrusions, simply by looking at patterns in netflow statistics, it is at least worth investigating if depressive users can be identified in a similar manner. One might argue that depression is more of a behavioral attribute when compared to spamming and intrusions, both of which are network centric activities; however, there are increasingly large number of applications where netflow data was used to identify behavioral attributes. For example, Chen et al. showed that users can be identified with 90% accuracy solely based on their active and idle time distributions in an online game-play activity [28]. More interestingly, a recent study showed that a persons internet usage activity represented by Cisco netflow data can be used as a behavioral biometric, achieving a classification accuracy of up to 90% [57].

In section 1.2.2, it was shown that depression is correlated with four basic Internet attributes: 1) Category of Internet use; 2) Obsessive behavior (repeated visits to same websites); 3) Duration; 4) Time of usage (day vs. night usage). While flow level statistics do not capture these attributes directly, a surprisingly large number of related features can be derived. For example, one can track obsessive behavior by looking at the frequency of visits to the same destination ip-address. Alternatively, the type of activity, i.e., chat, gaming etc., can be identified by filtering flows based on destination port and protocol fields. In fact, it will be shown in latter chapters that most of the relevant features (as mentioned in section 1.2.2) can be derived from Netflow data.

2.1.1. Netflow versions. There are several versions of Netflow. Versions 1, 5, 6, 7 and 8 export packets over UDP. This means that packets can be lost, because UDP doesn't guarantee packet delivery. Lost packets can be detected by going over the sequence numbers in the Netflow packet header. Netflow versions 5 and 9 are very popular among research communities. Version 9 is designed to be transport protocol independent, but UDP is still commonly used. It is designed to be extensible and makes use of templates that can be defined by the user.

Even though netflow v9 provides greater flexibility, netflow v5 was used because of its simplicity. It provides a simple and clear way to gain insight into the network and is able to give answers to questions like who is (has been) communicating with whom, how, when, for how long and how much data they are transferring. A detailed description of a v5 flow record is given in Table 2.1.

Table 2.1. A Netflow v5 Record

Bytes	**Contents**	**Description**
0-3	srcaddr	Source IP address
4-7	dstaddr	Destination IP address
8-11	nexthop	IP address of next hop router
12-13	input	SNMP index of input interface
14-15	output	SNMP index of output interface
16-19	dPkts	Packets in the flow
20-23	dOctets	Total number of Layer 3 bytes in the packets of the flow
24-27	first	SysUptime at start of flow
28-31	last	SysUptime at the time the last packet of the flow was received
32-33	srcport	TCP/UDP source port number or equivalent
34-35	dstport	TCP/UDP destination port number or equivalent
36	pad1	Unused (zero) bytes
37	tcp_flags	Cumulative OR of TCP flags
38	prot	IP protocol type
39	tos	IP type of service
40-41	src_as	Autonomous system number of the source, either origin or peer
42-43	dst_as	Autonomous system number of the destination, either origin or peer
44	src_mask	IP protocol type
45	dst_mask	IP type of service
46-47	pad2	Unused (zero) bytes

2.1.2. Flow Exports.

Netflow operates by creating cache entries. This means that for every connection between hosts, an entry is made in the cache. For each combination of key fields, the stored information includes octets send, packets send, start time, end time and TCP flags. The information included depends on the Netflow version and set-up and is accessible by using the Command Line Interface (CLI) for an immediate view of the Netflow cache or by exporting Netflow data packets to a Netflow Collector (used in this research) [3].

Flows are exported from the cache when the following events occur:

- A TCP flow is ended by a FIN or RST flag
- When a flow has been inactive for a specified time (the inactive timer)
- When a flow has has been active for a specified time (the active timer)
- When the Netflow cache is full

The inactive and active timer can split flows, resulting in the same flow, broken into several smaller flows. These records will have to be merged during the analysis. Another thing to keep in mind when using Netflow data is that flows are defined as a unidirectional stream of packets. For a successful connection between two hosts, Netflow will export two flows, one for each direction.

2.2. RESEARCH APPROACH

This thesis is accomplished via a three phase procedure.

- **Phase I - Privacy preserved data collection:** In order to classify depression based on internet usage, both depression score and internet usage data associated with a person is required. This data is obtained from the student population of Missouri S&T, volunteering to participate in the study. Depression score is obtained on a scale of 1-60 and is based on Center for Epidemiologic Studies Depression Scale (CES-D) survey. Netflow statistics on the other hand is monitored by the on-campus IT department and is logged to a secure server for subsequent use in analysis. To mask the association between the participating students and their internet usage data, additional anonymization procedures are enforced during the data collection process.

- **Phase II - Data preprocessing and statistical analysis:** In its raw form, netflow approaches several millions of rows of data when aggregated over a month. Additionally, it contains several unnecessary flows pertaining to DNS lookup requests, System updates (OS, Antivirus etc.), and other auxiliary services that are clearly activities not associated with a user. Therefore, before any analysis can begin, unnecessary flows are filtered out from the logs. Thereafter, the data is preprocessed to extract features that are thought to be relevant to depression. Not only does this reduce/compress the data to a manageable size, but also facilitates the conversion of data to a relevant format where statistical techniques can operate. As there is no related research, statistical analysis will be performed to get a basic feel for the data, test existing hypotheses and possibly discover new ones.

- **Phase III - Depression classification using Machine Learning Techniques:** Leveraging on the statistical insights from phase II, several machine learning techniques (Time series classification, multi-instance learning, Fuzzy and Ellipsoid ARTMAPs and support vector machines) will be employed to perform depression classification. Performance metrics such as precision, recall, and accuracy will be used to the validate the performance of the final classifier against unseen data to assess its reliability and correctness. Finally, additional fine-tuning may be done, based on the insights gained during the process.

The entire process flow is summarized in Figure 2.1.

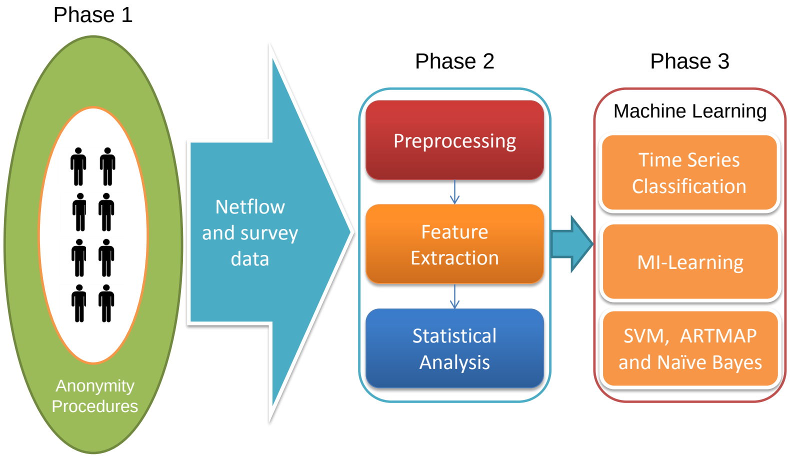


Figure 2.1. Summary of the proposed research plan

2.3. DATA COLLECTION

In this section, the infrastructure and process required to collect necessary foundational data is described. In particular, students from Missouri S&T are used as the subject pool for acquiring depression score and netflow data. Depression score is quantified based on standardized surveys while netflow data is collected in collaboration with the on-campus IT department. The section closes with a detailed account of privacy preserving mechanisms used.

2.3.1. Participants. In this study, the participant pool comprised of the undergraduate student population from Missouri S&T. In particular, three undergraduate classes, namely Psych 50, CS 284 and CS 150 volunteered to participate in this research. Psych 50 was chosen as the students had to fulfill their general psychology course research requirement. A majority of students in CS 284 and CS 150 volunteered since these were computer science courses. All in all, Psych 50 represents the general student population as students from all majors are required to take that course. CS students, being very technical and privacy conscious, asked a lot of questions regarding privacy and the research in general. Their 100% participation shows that the other majors can be convinced to participate in the future. A detailed demographic description of participants is shown in Table 2.2.

Table 2.2. Demographic statistics of the participant students

	CS	**Psych**
Male	120	68
Female	8	20
Totals	128	88

Of the 216 participants, data from only 165 participants were utilized due to difficulties with process coordination and errors on the part of the IT department.

2.3.2. Depression Measurement. Once the subject pool was determined, participants depression levels were quantified based on Center for Epidemiologic Studies Depression Scale (CES-D). CES-D was originally developed by Lenore Radloff of Utah State University and is used to measure depression levels in a general population [66].

It consists of 20 questions rated on a 4-point Likert scale that ranges from 0 (rarely or none of the time) to 4 (most or all of the time). Possible scores range from 0 to 60 with higher scores indicating greater levels of depression symptoms. In general, a score of 16 or above is considered to be indicative of depression. Internal consistency for the general population is in the good range with Cronbachs α of 0.85 [42]. A detailed demographic description of participants depression scores is shown in Table 2.3.

Table 2.3. Demographic statistics of participants depression levels

	CS	**Psych**	**Depressed**	**Non-Depressed**
Male	120	68	54	132
Female	8	20	10	18
Totals	128	88	64	152

In order to minimize demand characteristics in the study (where participants form an interpretation of the experiment's purpose and unconsciously change their behavior accordingly) [58], the survey was entitled "*Attitudes, Feelings and Perceptions of College Life*", with the CES-D question items embedded among a variety of other items asking about participants' attitudes and feelings about being a college student.

In addition to depression scores, students also completed a survey on *overall college adjustment* along with the Background Information Sheet where demographic and academic information is voluntarily provided. This information is collected in order to derive additional features in the event when netflow data is not sufficient to categorize depression. Additionally, this data was used to check if the proposed framework can be applied to predict behavioral attributes other than depression. Please see the Appendix A for a listing of all the questionnaires and consent forms used.

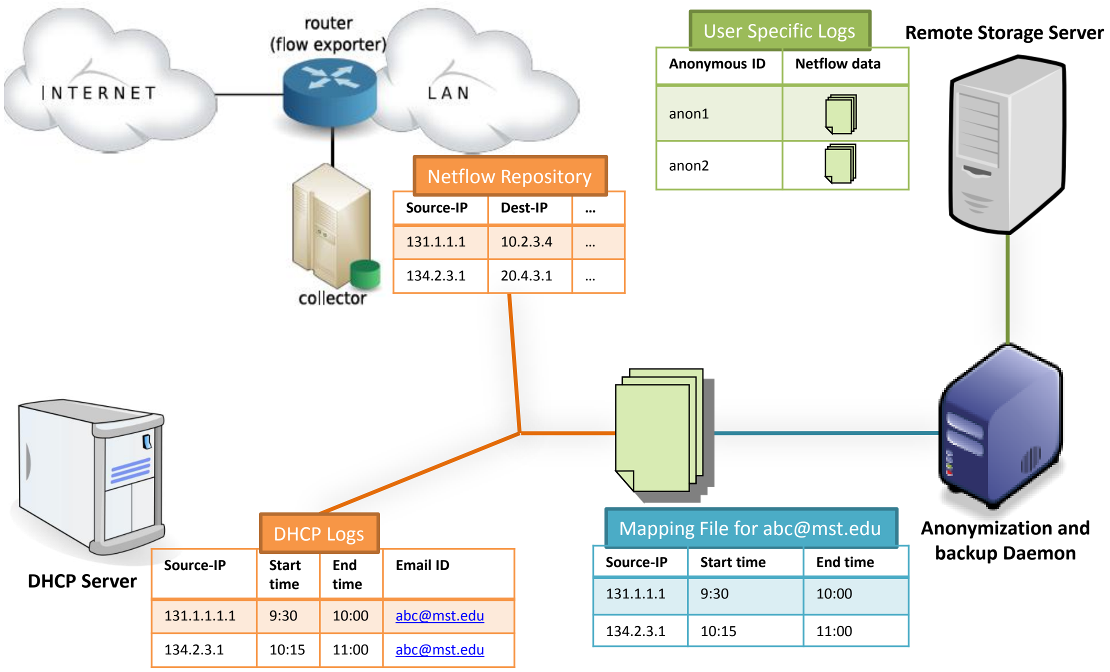
2.3.3. Internet Data Collection. The main data-source of "*Internet use*" for this research is Netflow. Netflow represents a relatively light-weight network monitoring tool. Network monitoring can be performed by packet level inspection with the use of tools like TCPDUMP [47]. However, for large packet switching facilities the processing time and disk-space requirements will become infeasible. SNMP [72] is another popular alternative for network monitoring. It, however, aggregates information on a very high level (i.e. network interface throughput or device uptime) and a lot of information is lost. Between these two extremes (recording every packet or aggregating high level data) is the Cisco's Netflow protocol.

In this study, participants' internet usage on S&T campus machines is monitored with the help of the on-campus IT department. Note that a subject's Internet use may not be restricted to data collected from Campus machines. Students also tend to use Internet extensively off campus and also on their mobile phones. Collecting the entire gamut of a subject's Internet use is difficult, if not impossible. Therefore, the current study restricts itself to on-campus internet usage.

Typically, every on-campus IT department monitors the netflow data of all its users in order to determine attacks, bottlenecks, network abuses etc. The Missouri S&T campus has a connection to both the standard commodity Internet and the Internet 2 education research network. Both Internet and Internet 2 traffic pass through the same router where Netflow statistics recording and exporting is enabled. Every 5 minutes, these flows are exported from the router to a collector where they are stored for a period of around 45 days for network and security analysis purposes and are then discarded automatically.

In order to obtain participants netflow data, the flows pertaining to student participants are identified based on the `source-ip` field and is subsequently filtered and logged to a secure remote server at the end of every month. As S&T campus uses DHCP (Dynamic Host Configuration Protocol) to provide IP address, DNS, and gateway information for all its workstations, the IP address that one individual is using at any given point of the day, could easily be used by someone else later in the day or possibly the next day. Therefore, the extraction process begins by creating a mapping file, associating each user with a set of assigned ip addresses, along with the start and end time stamps. This information is used by the backup daemon to extract user specific netflow information by filtering flows based on `source-ip` field. The mapping file is created by analyzing DHCP logs that includes a person's `userid`, which is also their email address on S&T campus. The entire process is summarized in Figure 2.2.

Potential Errors: There are some potential errors with the aforementioned method. As the granularity of netflow capture is a 5 minute window, a person's IP address changing within that particular window will go unnoticed. Furthermore, as S&T campus uses a traffic shaper to optimize performance, improve latency, and in-



Netflow Repository

Source-IP	Dest-IP	...
131.1.1.1	10.2.3.4	...
134.2.3.1	20.4.3.1	...

User Specific Logs

Anonymous ID	Netflow data
anon1	
anon2	

DHCP Logs

Source-IP	Start time	End time	Email ID
131.1.1.1.1	9:30	10:00	abc@mst.edu
134.2.3.1	10:15	11:00	abc@mst.edu

Mapping File for abc@mst.edu

Source-IP	Start time	End time
131.1.1.1	9:30	10:00
134.2.3.1	10:15	11:00

Figure 2.2. Illustration of the netflow data logging process. At the end of every month, participant specific netflow data identified through DHCP logs are anonymized and copied to a secure remote storage

crease usable bandwidth by blocking certain protocols (Ex: peer-to-peer traffic), the recorded traffic information might be different from the actual activity. However, the use of traffic shaper is very common. Depression classifier, if deployed, should be able to perform the classification by implicitly accounting for traffic shapers distortions.

2.3.4. Student Rights and IRB Approval. A descriptive handout was given to all student participants prior to the start of the Phase I, explaining their benefits, obligations and rights (including the right to withdraw at any time) in this study. An Opt-in form is also included with the handout (see the last page of Appendix A). The project was approved by IRB (see Appendix B) under Exemption Category 4 stated below:

> *Research involving the collection or study of existing data, documents, records, pathological specimens, or diagnostic specimens if these sources are publicly available or if the information is recorded by the investigator in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects.*

2.4. PRIVACY AND RELATED CONCERNS

Privacy is very critical for research of this nature. During the survey process, several students inquired about privacy preserving mechanisms, demonstrating its importance. Student participation and IRB approval would not be possible without the mechanisms described in this section.

From a privacy standpoint, there are two potential concerns:

1. Identifying and associating user specific attributes such as gender, academic performance and college adjustment scores based on the participants userid (in this case, the email id)

2. Identifying and associating internet usage activity of the participants

The first concern is addressed by distributing data among multiple parties such that no particular group has enough information to associate a personal attributes with student participants. In particular, data is distributed among three groups: 1) Survey and information collection group (SIG); 2) Research group (RG); 3) IT department. SIG is responsible for conducting surveys and for entering the information into a secure database. In this table (`table1`), every participant is ID'ed by their userid.

The IT department is responsible for collecting participants netflow data. Initially, they generate a "*link table*" (`table2`), associating a participants userid with a randomly generated pseudonym. This table is restricted to IT and cannot be accessed by SIG or RI. When netflow data is logged to a secure remote server, the link table is used to replace userid's with pseudonyms. The anonymized data is accessible by the research group. Since IT has access to userid's and pseudonyms, one can argue that IT can associate internet usage with participants; however, this is not truly a violation of privacy because IT, by default, has access to students netflow data and are required to maintain user privacy as per the dictated terms and conditions.

Finally, the anonymized survey data is stored in `table4`, which is generated by replacing `userid` field from `table1` by using the link table. While `table1` access is only permitted to SIG, `table4` can be accessed by RI, but is restricted for IT. As IT does not have access to the survey data, they cannot associate survey specific attributes such as gender, grades etc. with the pseudonym. On the other hand, RI cannot associate the pseudonym with the participant as they don't have access to the link table. SIG is only restricted to survey data and they cannot deduce anything regarding students internet activity. Table 2.4 summarizes the resulting access control matrix.

Table 2.4. Access control matrix illustrating read and write access for different groups

Sub/Obj	**Survey Data**	**Link Table**	**Anonymized Flow Data**	**Anonymized Survey Data**
SIG	R, W	-	-	-
RI	-	-	R	R
IT	-	R, W	W	-

As netflow represents packet level meta information, some degree of privacy is already ensured. The only sensitive information embedded in netflow data is the source and destination ip-address as the website URL can be deduced by performing a DNS reverse lookup operation. One way to avoid this issue is to discard the source and destination ip-address fields. However, the *type* of browsing activity engaged such as news reading, information seeking (health, academic etc.), social networking usage etc. which are known to correlate with depression, can only be deduced via the website URL.

To resolve this issue, the `dest-ip` field is preprocessed to represent the URL category. For instance, URLs www.facebook.com and www.orkut.com can be represented as "social networking" category. Similarly, other categories such as "media", "information site", "Academic" etc. can be used instead of the raw URL. As each category is traced to at most K unique URLs, the proposed approach is consistent with K-anonymity model [74]. Not only does this reinforce security, but it also serves as a preprocessing step for use in designing the classifier. The details of pre-processing URLs to categories is described in section 4.2.2.

3. FEATURE EXTRACTION AND STATISTICAL ANALYSIS

This chapter describes the first experimental steps taken in order to associate netflow data with depression. As there is no related literature in this direction, exploratory and statistical analysis was performed in order to get a better feel for the data. Furthermore, statistical analysis was performed as “real” Internet data was involved. This is the first study to do so.

Since netflow data is represented as a vector of flows, it is not straight forward to apply conventional statistical and machine learning techniques directly. This chapter, therefore, opens with a discussion on feature extraction. In particular, three types of feature vectors are derived, each with an increasing amount of information granularity. The remainder of the chapter describes several statistical insights gained by analyzing these features. A summary of conclusions are presented towards the end of this chapter.

3.1. DERIVING FEATURE VECTORS FROM NETFLOW DATA

Netflow data in its natural form is not suitable for statistical analysis. An example of sample netflow data for a single participant is shown in Table 3.1.

Table 3.1. Sample Netflow data

srcIP	**dstIP**	**prot**	**srcp**	**dstp**	**oct**	**pkts**	**dur**
131.151.211.200	208.78.158.10	6	65055	80	1187	13	158
131.151.211.200	208.78.158.10	6	65058	80	1141	12	166
131.151.211.200	208.78.158.10	6	65042	80	402	5	67
131.151.211.200	208.78.158.10	6	65062	443	1533	9	196
131.151.211.200	98.156.50.166	6	65072	57460	587	7	98

Each row in the table is termed as a flow. As netflow v5 is used, the flow record is defined as an 8-tuple key given by:

$$\overrightarrow{flow} = \langle src\_ip, dest\_ip, src\_port, dest\_port, protocol, octets, packets, duration\rangle \quad (1)$$

In order to derive meaningful statistics, one has to preprocess netflow data $N = \{\overrightarrow{flow_i}\}_{i=1}^{k}$ into an n-dimensional feature vector. Furthermore, as the number of flows associated with a participant approaches the order of magnitude of millions when aggregated over a month, preprocessing also serves the purpose of compressing the data to manageable proportions. As the space of all possible feature vectors $\mathcal{F} = 2^{\langle flow \rangle}$ is typically very large; care must be taken to extract features that are thought to be relevant to depression. In this regard, correlations described in section 1.2.2 form a good starting point.

In this thesis, three types of feature vectors are explored, each with an increasing amount of information granularity as illustrated in figure 3.1.

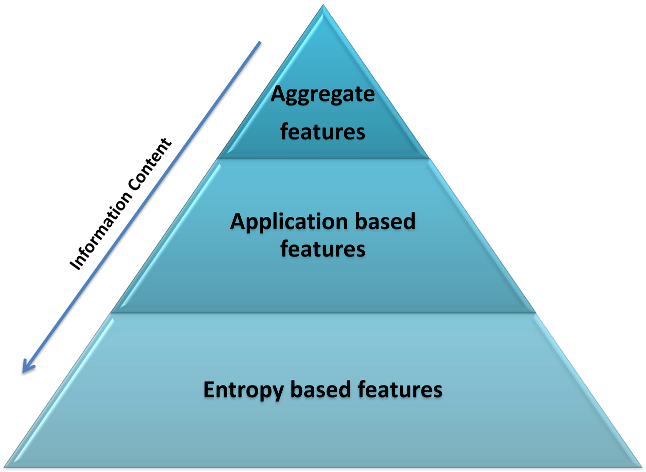


Figure 3.1. Three levels of feature vectors with increasing information granularity

1. **Aggregate traffic statistics:** At the first and the most abstract level is the overall aggregate traffic statistics such as total packets, flows, octets etc. Though abstract, this kind of a feature vector can be used to test hypothesis such as: *"Does more internet usage correlate with greater likelihood of depression?"*. Aggregate flow statistics are derived by using `flow-report` in the `flow-tools` suite. In addition, bash scripting was used to extract the data and convert it into a feature vector, one per participant. In total, 14 features were derived and are summarized in Table 3.2.

Table 3.2. Features vector for aggregate traffic statistics

Feature	**Description**
flows	Total Flows
oct	Total Octets
pkts	Total Packets
timeflows	Total Time (1/1000 secs) (flows)
durreal	Duration of data (realtime)
durdata	Duration of data (1/1000 secs)
aftime	Average flow time (1/1000 secs)
apsize	Average packet size (octets)
afsize	Average flow size (octets)
apflow	Average packets per flow
afsec	Average flows / second (flow)
afsecreal	Average flows / second (real)
akbits	Average Kbits / second (flow)
akbitsreal	Average Kbits / second (real)

2. **Application level statistics:** Traffic aggregation alone loses a lot of information. For example, high email usage and low gaming usage cancel each other and when internet usage is considered as a whole. Correlations described in section 1.2.2 show that application level usage patterns, such as the one described above, may be crucial in understanding depression. Hence, the second level feature vector attempts to capture more information by sub-categorizing aggregate traffic features by application. i.e., traffic features such as total octets, packets and duration are derived per application such as emailing, gaming, chatting etc.

   Based on the `dest-port` and `dest-protocol` fields, the netflow data was categorized into 61 different applications. Please see Appendix C for details.

   Since this study was conducted in college, the netflow data was only logged for usage activity on-campus. As a result, several application categories showed little or no activity. This is likely due to the fact that universities block certain protocols (`torrent`), and also due to the fact that students tend to limit their activities on-campus. In any case, applications with little or no activity were discarded. In total, 25 application categories were retained. These applications were further grouped into 8 logical categories and is described in Table 3.3. For each of the

25 applications, total octets, packets and duration was determined. The resulting feature vector is therefore of 75-dimensions.

Table 3.3. Application groups and categories

Group	**Applications**
P2P	*File-sharing applications based on peer-to-peer architecture (Edonkey, neomodus, winmx)*
HTTP	*HyperText Transfer Protocol applications (http, https)*
Streaming	*Stream media applications (Shoutcast, real, winmedia, streamworks, audiogalaxy)*
Chat	*Instant messaging applications (Aim, irc, carracho)*
Email	*Email traffic (IMAP, POP3, SMTP)*
FTP	*File transfer applications (snmp, ftp)*
Games	*Massively multiplayer online games (battlenet, quake, starseige, portzero, halflife, gamespyarcade, directx)*
Remote Access	*Remote file system access (afs, nfs)*

3. **Entropy based features:** As every flow is associated with a start and end timestamp, they can naturally be expressed as a time series. More particularly, each of the 8 attributes defined in Equation 1 can be represented as a time-series in itself. Level 1 and Level 2 features are based on aggregation and therefore fail to capture time-series specific characteristics such as periodicity, trends, randomness etc, which *may* prove to be crucial in understanding depression.

   One way to quantify trends and periodicity is to use spectral analysis techniques such as the discrete wavelet transform (DWT) or Discrete fourier transform (DFT). With these techniques, periodic basis/component signals are extracted in the form of low frequency coefficients and are typically used in time series analysis/classification [61]. Nevertheless, DFT was not considered in this phase as they provide little or no statistical intuition.

   Hence, *randomness* was chosen as a metric to extract useful information from the time-series. In particular, information regarding time-series frequency distribution was captured in terms of shannon entropy [70]. Intuitively, entropy estimates the average uncertainty of a series of discrete events. Consider H(dest-port); It estimates an average of how much the next destination port visit will surprise us. If all destination ports are equally visited, it is rather difficult to *guess* the next port visit. As more number of bits is required to represent this information, the entropy is high. On the other hand, if the distribution is highly skewed, it is rather trivial to make a guess (probabilistically); Entropy is therefore low.

Given a discrete random variable X, shannon entropy $H(X)$ is given by equation 2, where $p(x) = Pr\{X = x\}, x \in \chi$

$$H(X) = -\sum_{x \in \chi} p(x) \log p(x) \tag{2}$$

Note that $Pr\{X = x\}$ is the probability of occurrence of x with respect to some event. In case of netflow, one can compute the entropy of any of the 8 attributes defined in Equation 1 with respect to flows. For example, when entropy of *ports* is calculated, the probability of occurrence of an event x (say port 80) is given by the number of flows with port 80 divided by the total number of flows. Alternatively, when continuous variable such as the bytes, octets or duration is used, entropy is given by:

$$H(x) = -\int_{-\infty}^{\infty} p(x) \log p(x) dx \tag{3}$$

It is also possible to compute the entropy of flow attributes with respect to discrete variables such as the ip-address or port. Consider $H(duration\_relative\_ip)$; it gives randomness in the amount of duration time spent on different ip-addresses. However, the current study limits itself to flow distributions due to the time constraints on the project.

3.2. BACKGROUND ON STATISTICAL METHODS USED

This section begins with a very brief description of the statistical techniques and methods used to analyze the feature vectors. In particular, correlation coefficients, and t-test analysis is described.

3.2.1. Correlations. Correlations are used to describe the degree of association between two variables. As correlations between depscore (continuous variable) and various feature vectors are to be computed, Pearson product moment correlation is used. Pearson's correlation, denoted by r, is defined as the covariance of the two variables divided by the product of their standard deviations:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{4}$$

However, Pearson correlation only measures the *linear* relationship between two variables. For this purpose, ranked correlation coefficients such as Spearman's rank correlation coefficient (ρ) and Kendall's rank correlation coefficient (τ) are used. Spearman correlation is similar to Pearson's correlation except that the *rank* of variables are used. On the other hand, Kendall tau correlation (tau-b[2]) is given by:

$$\tau_B = \frac{n_c - n_d}{\sqrt{(n_0 - n_1)(n_0 - n_2)}} \tag{5}$$

where,

$$n_0 = \binom{n}{2}, n_1 = \sum_i \binom{t_i}{2}, n_2 = \sum_j \binom{u_j}{2}$$

t_i = Number of tied values in the i^{th} group of ties for the first quantity

u_j = Number of tied values in the j^{th} group of ties for the second quantity

Correlation coefficient, ranging from -1 to +1 is both a measure of the strength of the relationship and the direction of the relationship. A correlation coefficient of 1 describes a perfect linear relationship in which every change of +1 in one variable is associated with a change of +1 in the other variable. A value of -1 is analogously associated with perfect negative linear relationship. Zero correlation describes a situation in which a change in one variable is not associated with any particular change in the other variable.

3.2.2. Students T-test. As depscore can be used to deduce two depression categories, namely, depressed (depscore $\geq$ 16) or non-depressed (depscore $<$ 16), stu-

[2] makes adjustments for ties and is suitable for square tables

dents t-test can be used to check whether the means of internet usage features differ across both groups. The insight gained from this analysis can be used to deduce conclusions such as: *"On average, people playing online games are less prone to depression"*. More importantly, it can help identify prominent features that can be used to build a classifier.

t-test assumes normal data distribution and homogeneity of variance. Normality can be verified by observing P-P plot (probability-probability plot) while Levene's test can be used to assess the equality of variances. In case the data deviates from normal distribution, non-parametric Mann-Whitney U test will be used. On the other hand, if homogeneity of variances assumption is violated, it is corrected by avoiding the pooled estimate for the error term for the t-statistic and also by making adjustments to the degrees of freedom using the Welch-Satterthwaite method.

3.3. RESULTS

In this section, correlation and t-test results for all three levels of feature vectors is presented.

3.3.1. Aggregate Traffic Features.

Correlations (pearson, spearman and tau) between depscore and aggregate traffic feature are shown in Table 3.4. As spearman and tau correlations are more reliable and less sensitive to outliers, it is concluded that the feature `apflow` (Average packets per flow) exhibits a significant positive correlation with depression, r(163) = .137, .198 ($\rho < 0.05$).

Examination of P-P plots revealed that none of the traffic features follow a normal distribution. This was further verified by Kolmogorov-Smirnov test. As normality assumption is violated, independent samples Mann-Whitney U Test was conducted to check if any of the features differentiate between the depressed and non-depressed groups.

The test showed that there is a statistically significant difference in the mean values of `apflow` across depressed and non-depressed groups (U(163) = 2231, Z = -2.384, ρ(2-tailed) = 0.017). It can be further concluded that depressed people exhibit higher `apflow` (μ = 168.47, σ = 46.11) when compared to non-depressed group (μ = 110.91, σ = 14.51). The result is summarized in Figure 3.2.

3.3.2. Application usage statistics.

Correlations (pearson, spearman and tau) between depscore and application usage categories (Table 3.5) revealed several interesting associations. All three coefficients showed a significant correlation between

Table 3.4. Correlations between depscore and aggregate traffic features

Traffic Features	*depscore*		
	Pearson	**Spearman rho**	**tau-b**
flows	.091	.022	.035
oct	.195*	.078	.116
pkts	.170*	.101	.149
timeflows	.130	.060	.089
durreal	.095	-.002	-.001
durdata	.108	.035	.048
aftime	.045	.095	.140
apsize	.078	-.001	-.006
afsize	.140	.104	.152
apflow	.056	.137*	.198*
afsec	.104	.034	.049
afsecreal	.091	.021	.031
akbits	.193*	.080	.118
akbitsreal	.195*	.082	.120

***. Correlation is significant at 0.01 level (2-tailed)*
**. Correlation is significant at 0.05 level (2-tailed)*

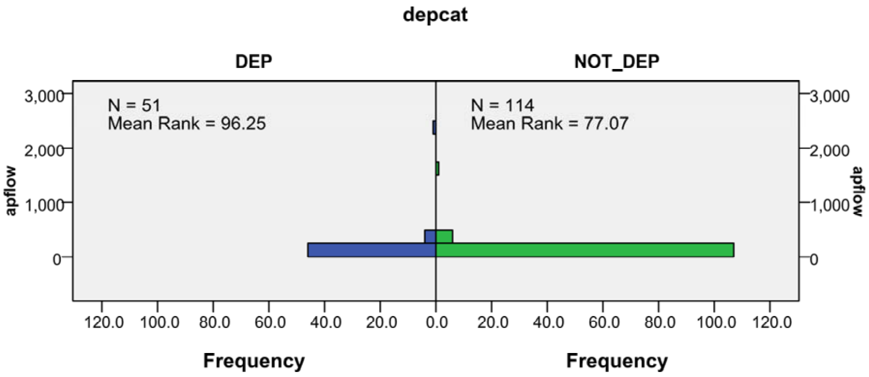


Figure 3.2. Independent Mann-Whitney U test results demonstrating higher `apflow` in depressed people

p2p_packets and depscore. Additionally, pearson's correlation also indicated a strong correlation between p2p_octets, p2p_duration with depscore. Overall, the results indicate a *strong association between excessive p2p usage and depression.*

Table 3.5. Correlations between depscore and application usage categories

Application Features	*depscore*		
	Pearson	**Spearman rho**	**tau-b**
p2p_octets	.173*	.075	.111
p2p_packets	.236**	.106*	.160*
p2p_duration	.265**	.098	.143
http_octets	.039	.005	.006
http_packets	.057	.074	.112
http_duration	.094	.044	.067
streaming_octets	.110	.001	-.002
streaming_packets	-.023	-.004	-.012
streaming_duration	-.019	-.001	-.007
chat_octets	.267**	.100	.145
chat_packets	.053	.007	.012
chat_duration	-.023	-.004	-.012
mail_octets	-.071	.010	.011
mail_packets	.164*	.050	.068
mail_duration	.202**	.048	.064
ftp_octets	-.023	-.004	-.012
ftp_packets	-.021	-.002	-.008
ftp_duration	.267**	.100	.145
game_octets	.095	.027	.043
game_packets	.104	.042	.063
game_duration	.094	.038	.058
remote_octets	.281**	.117*	.172*
remote_packets	-.023	-.004	-.012
remote_duration	.023	.018	.023

***. Correlation is significant at 0.01 level (2-tailed)*
**. Correlation is significant at 0.05 level (2-tailed)*

In addition to p2p usage, remote_octets is another feature where all three coefficients showed a significant correlation with depscore. Therefore, it can be inferred that *heavy duty remote application usage is significantly correlated with higher levels of depression, at least amongst college students.*Pearson's correlation also revealed a significant correlation between chatting, email and ftp usage duration with depscore. Therefore, it can be concluded that *Heavy chatting, email checking, and ftp usage is correlated with higher levels of depression.*

Examination of P-P plots revealed that none of the application usage features follow a normal distribution. This was further verified by Kolmogorov-Smirnov test. As normality assumption is violated, independent samples Mann-Whitney U Test was conducted to check if any of the features differentiate between the depressed and non-depressed groups.

The test showed that there is a statistically significant difference in the mean values of `remote_octets` across depressed and non-depressed groups (U(163) = 2343, Z = -1.989, ρ(2-tailed) = 0.047). It can be further concluded that depressed people exhibit higher `remote_octets` ($\mu = 1.17 \times 10^{10}$, $\sigma = 1.88 \times 10^{10}$) when compared to non-depressed group ($\mu = 5.90 \times 10^{9}$, $\sigma = 5.97 \times 10^{9}$). The result is summarized in Figure 3.3.

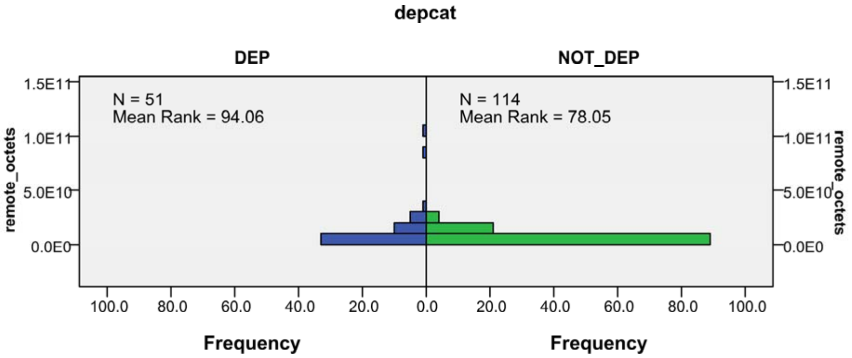


Figure 3.3. Independent Mann-Whitney U test results demonstrating higher remote octets in depressed people

Additionally, when applications were directly analyzed (instead of using groups), Mann-Whitney U Test revealed a significant difference in the mean values of `irc_octets` (U(163) = 2602, Z = -2.225, ρ(2-tailed) = 0.026), packets (U(163) = 2596, Z = -2.269, ρ(2-tailed) = 0.023) and duration (U(163) = 2608, Z = -2.182, ρ(2-tailed) = 0.029), indicating that `irc_usage` is significantly higher in depressed people.

3.3.3. Entropy based features. Correlations (ρ and τ) revealed a strong correlation between destination port entropy and depression scores. A complete summary of correlation coefficients are presented in Table 3.6

Table 3.6. Correlations between depscore and entropy based features

Entropy Features	*depscore*		
	Pearson	**Spearman rho**	**tau-b**
duration_ent	.123	.090	.141
bps_ent	.094	.005	.010
pps_ent	-.077	-.041	-.060
octets_ent	.098	.037	.056
packets_ent	.017	.048	.069
destip_ent	.081	.020	.029
destport_ent	.118	.095	.142
destprot_ent	.149	0.117*	.167*

***. Correlation is significant at 0.01 level (2-tailed)*
**. Correlation is significant at 0.05 level (2-tailed)*

Examination of P-P plots revealed that none of the entropy based features follow a normal distribution. This was further verified by Kolmogorov-Smirnov test. As normality assumption is violated, independent samples Mann-Whitney U Test was conducted to check if any of the features differentiate between the depressed and non-depressed groups.

The test showed that there is a statistically significant difference in the mean values of `duration_entropy` across depressed and non-depressed groups (U(163) = 2337.5, Z = -2.008, ρ(2-tailed) = 0.045). Figure 3.4 summarizes the result. Higher mean rank in depressed group indicates that *depressed people have higher flow duration entropy.*

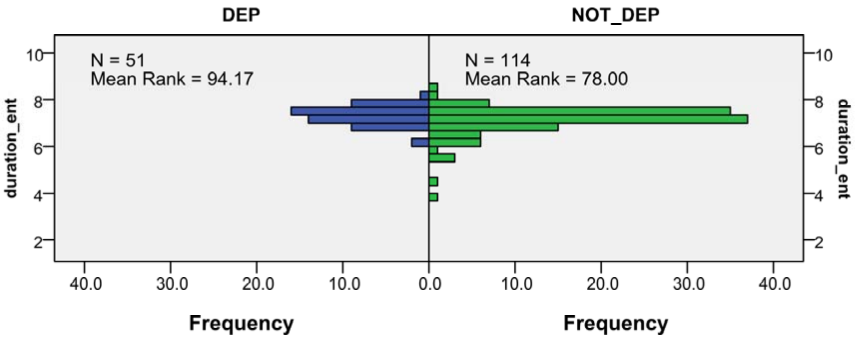


Figure 3.4. Independent Mann-Whitney U test results demonstrating higher internet usage duration entropy in depressed people

3.4. SUMMARY OF CONCLUSIONS

This chapter described the necessary pre-processing of netflow data in order to identify compressed, yet meaningful features. In particular, three levels of features were derived: Aggregate traffic statistics that are simply aggregates of individual traffic attributes, Entropy based features that quantify amount of information contained in netflow time, and finally, Application usage statistics that are aggregate traffic statistics further categorized by Internet application. Based on the analysis, it can be concluded that depressed people tend to exhibit:

1. Higher average packets per flow
2. High P2P and remote application usage
3. High chat activity, frequent email checking and high ftp activity
4. High flow duration entropy

High average packets per flow could mean a lot of things. One way to interpret *apflow* is in terms of application switching behavior. High *apflow* can occur when there is frequent switching between various internet applications. Frequent switching is synonymous with the lack of attention and exhibits similarities to Attention deficit hyperactivity disorder (ADHD). This result indicates a possible connection between ADHD and depression.

High P2P and remote application usage in depressed people is rather curious. As far as remote application usage is concerned, one may hypothesize its association with assignments and school work. As the subject pool mainly comprised of Computer Science majors, the assumption might not be far off. Nevertheless, additional experiments are required to interpret the result correctly.

Frequent email could indicate high levels of anxiety. The interpretation of high flow duration entropy and high chat activity is also not clear at this point. Nevertheless, the study opens a whole pandora's box of questions that might be of interest to mental health community.

4. DEPRESSION CLASSIFICATION

This chapter is solely devoted towards the main research task - to build a classifier that can identify depression based on internet usage features. First, the research task is formalized to simplify the discussion in the remainder of this chapter. As most machine learning methods are based on the notion of *similarity* between input vectors, the chapter begins by defining a new similarity metric based on the information context of $\overrightarrow{flow}$.

The chapter proceeds by highlighting and leveraging similarities to other problems domains such as the document classification, video recognition etc. While some techniques are directly applicable, a few modifications are described for specializing the existing machine learning techniques for depression classification. More particularly, context aware flow similarity is utilized for building custom support vector machine kernels. This chapter also provides a brief description of potential supervised machine learning techniques that may be applied to feature vectors derived in chapter 3.

As verification and validation is very important to assess the reliability of the classifier, training and testing strategies, along with description of commonly used evaluation metrics such as the accuracy, precision and recall are presented. The chapter closes by presenting the evaluation results for all the proposed methods.

4.1. FORMAL PROBLEM STATEMENT

To make the remainder of the discussion in this chapter succinct and precise, the formal problem statement, along with the necessary mathematical notation is described. The problem of depression classification can be formalized as follows:

> *Given a set of examples* n *examples* $\mathcal{T} = \{(\mathcal{N}_i, D_i) \mid \mathcal{N}_i = \{\overrightarrow{flow_j}\}_{j=1}^{k_i},\ D_i \in \{0,1\}\}_{i=1}^{n}$, *where* $\mathcal{N}_i$ *represents the netflow data and* $D_i = \{0,1\}$ *indicates whether the* i^{th} *sample corresponds to a non-depressed or depressed participant; the task is to build a model* $M : \mathcal{N} \to D$

Note that the number of flows associated with each participant may be variable. More particularly, the number of flows for the i^{th} participant is denoted by k_i.

4.2. CONTENT AWARE NETFLOW DISTANCE METRIC

The similarity between examples of a set of data often gives much information about the patterns that may be present in that data. Following this premise, there are quite a few machine learning algorithms that are based on the notion of *similarity* between input vectors. One of the best known examples is the use of distance metric in clustering algorithms. In such exploratory analysis techniques, the choice of distance metric is very important as it will have a direct influence on how clusters are formed. Therefore it is important to choose a metric that is relevant to the characteristics of application under consideration.

In this research, data is obtained in the form $\mathcal{T} = \{(\mathcal{N}_i, D_i)\}_{i=1}^{n}$. Hence, one needs a meaningful notion of distance between netflow data; i.e., $d(\mathcal{N}_i, \mathcal{N}_j)$ needs to be defined. As $\mathcal{N}$ represents a set, hausdroff set distance metric is used. Formally, given two finite point sets $A = \{a_1, a_2, \dots, a_p\}$ and $B = \{b_1, b_2, \dots, b_q\}$, the hausdroff distance is defined as:

$$H(A, B) = \max(h(A, B), h(B, A)) \tag{6}$$

where

$$h(A, B) = \max_{a \in A} \min_{b \in B} \|a - b\| \tag{7}$$

$\|.\|$ is some underlying norm on points of A and B (usually Euclidean or L_2 norm). The function $h(A, B)$ is called the *directed* hausdroff distance from A to B. It identifies the point $a \in A$ that is farthest from any point of B and measures the distance from a to its nearest neighbor in B (using the appropriate $\|.\|$). Thus, the hausdroff distance $H(A, B)$ measures the mismatch between two sets by measuring the distance of point A that is farthest from any point of B and vice-versa (see Figure 4.1 for an illustration).

While $H(A, B)$ can trivially be computed in $O(pq)$ for two point sets of size p and q respectively, the implementation utilized an efficient algorithm, reducing the complexity to $O((p + q)\log(p + q))$ [14].

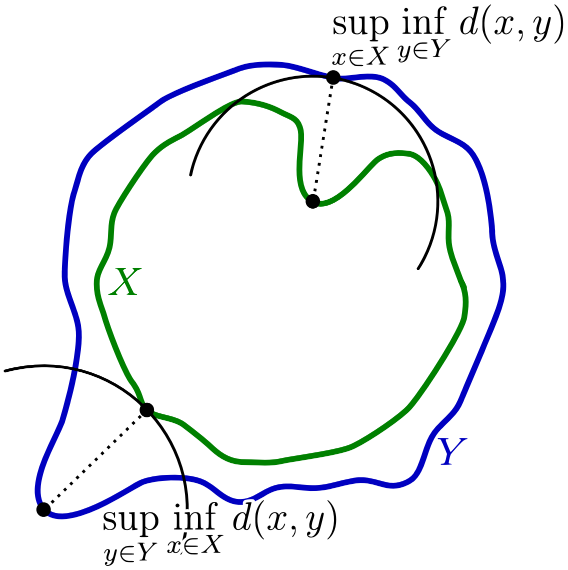

$\sup_{x\in X}\inf_{y\in Y} d(x,y)$
X
Y
$\sup_{y\in Y}\inf_{x\in X} d(x,y)$

Figure 4.1. Components of the calculation of the Hausdorff distance between the green line X and the blue line Y (source: `http://en.wikipedia.org/wiki/Hausdorff_distance`

4.2.1. Context aware flow distance metric. While computing $H(A,B)$, the underlying distance $\|.\|$ between individual $\overrightarrow{flow} \in \mathcal{N}$ needs to be defined. Equation 1 describes $\overrightarrow{flow}$ as an 8-tuple vector with a mix of continuous and categorical attributes. While the distance between continuous attributes such as `flow`, `octets`, `packets`, and `duration` can be computed in terms of Euclidean distance, categorical attributes such as the `ip-address`, `protocol` and `port` needs special attention. As depression is known to be correlated with the *type* of internet activity and the *type of content* browsed, the distance metric somehow needs to account this information.

One way to accomplish this task is to compute the distance between ip_i and ip_j in terms of information content similarity. More particularly, the idea is to express $d(ip_i, ip_j)$ in terms of $d_{content}(p_i, p_j)$, where p_k denotes the webpage obtained by traversing the URL determined using the DNS reverse lookup operation on ip_k. In order to express $d_{content}(p_i, p_j)$, a few definitions are in order.

Definition 4.1 (Categorical Tree)**.** Let $T = (\mathcal{C}, E)$ represent a hierarchical n-ary tree, where C represents a world wide web content category and $e(a,b) \in E$ represents a hierarchical relationship between categories a and b.

Definition 4.2 (WWW Category)**.** A WWW category $c \in \mathcal{C}$ represents a collection of webpages $\{p_i\}_{i=1}^{k}$ that are *similar* to each other in terms of information content or function. For example, `www.facebook.com` and `plus.google.com` can be considered similar in terms of function while `xkcd.com` and `abstrusegoose.com` are similar in terms of content (both are comic strips).

Given a categorical tree $T = (\mathcal{C}, E)$, representing the entire WWW, $d_{content}(p_i, p_j)$ is given by:

$$d_{content}(p_i, p_j) = |d(pred, c_i) + d(pred, c_j)| \tag{8}$$

where, p_k belongs to the WWW category c_k, $pred$ is the common predecessor of nodes $c_i, c_j \in \mathcal{C}$; i.e., $PREDECESSOR(c_i) = PREDECESSOR(c_j) = pred$, and $d(c_i, c_j)$ is the count of the number of edges encountered while traversing from node c_i to c_j.

For intuition, consider an example categorical tree shown in Figure 4.2. Suppose that p_1 belongs to the category `Top/Movies/Action/Comedy` and p_2 belongs to `Top/Movies/Drama/Political`. Given this information, the common predecessor of both nodes is `Movies`. Therefore, $d_{content}(p_1, p_2) = |d(\texttt{Movies}, \texttt{Comedy}) + d(\texttt{Movies}, \texttt{Political})| = 4$, which is the same as the distance required to traverse from `Top/Movies/Action/Comedy` to `Top/Movies/Drama/Political` through the common predecessor.

Once the categorical distance is known, one way to compute $d(\overrightarrow{flow_i}, \overrightarrow{flow_j})$ is as follows:

$$d(\overrightarrow{flow_i}, \overrightarrow{flow_j}) \;=\; \left(\sqrt{\textstyle\sum_{a_i, a_j \in \overrightarrow{flow_i}\overrightarrow{flow_j};\ \forall a \in \{flows, duration, packets, octets\}} (a_i - a_j)^2}\right)^{d} \tag{9}$$

Where,

$$d = 1 + d_{content}(LUP(ip_i), LUP(ip_2))$$

and $LUP(.)$ signifies the DNS reverse lookup operation.

Let us take a moment to understand Equation 9. It is basically computing the distance between two $\overrightarrow{flow}$ by considering the distance in two different sub-spaces. First, the continuous attributes in $\overrightarrow{flow}$ is used to compute the Euclidian distance. It is then magnified by a factor d, where d represents the distance in the discrete `ip-address`

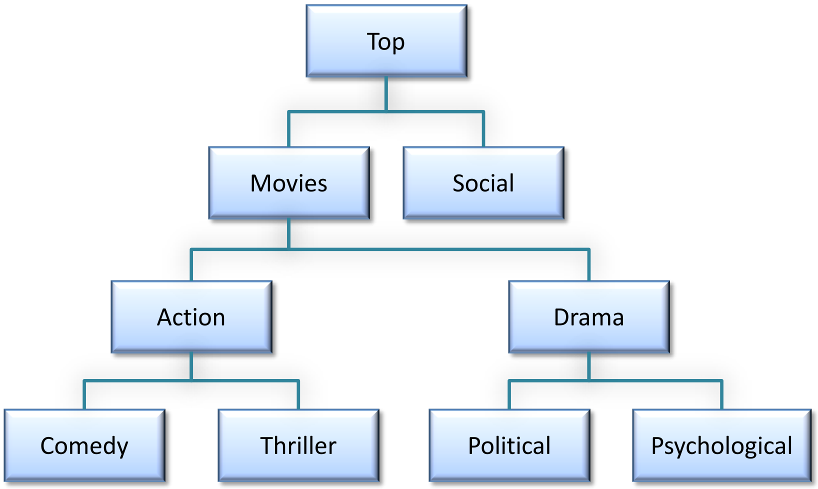


Figure 4.2. A sample categorical tree representing the categorical hierarchy. Every category encompasses a set of related webpages.

space. If ip_1 and ip_2, both belong to the same category, then $d_{content} = 0$ and the magnification factor is 1. However, as $d_{content}$ increases, the distance is exponentially magnified as distance within the `ip-address` space is given more priority.

In order to avoid huge numbers, log factor is considered. Therefore, Equation 9 can be rewritten as:

$$\begin{aligned} d(\overrightarrow{flow_i}, \overrightarrow{flow_j}) &= 1 + d_{content}(LUP(ip_i), LUP(ip_2)) \\ &\log\left(\sqrt{\sum_{a_i, a_j \in \overrightarrow{flow_i}\overrightarrow{flow_j};\ \forall a \in \{flows, duration, packets, octets\}} (a_i - a_j)^2}\right) \end{aligned} \tag{10}$$

The consideration of log factor also adds others attractive properties. For instance, log is known to boost smaller values by a greater fraction when compared to larger inputs. This leads to a smoother distance interpolation in continuous attribute space.

Using Equation 10, Equation 7 can be rewritten as:

$$d(\mathcal{N}_i, \mathcal{N}_i) = \max_{\overrightarrow{flow_i} \in \mathcal{N}_i} \min_{\overrightarrow{flow_j} \in \mathcal{N}_j} \left\| d(\overrightarrow{flow_i}, \overrightarrow{flow_j}) \right\| \tag{11}$$

4.2.2. Building the categorical tree. The World Wide Web (WWW) is a huge resource of interlinked hypertext documents accessed via the Internet. What started in 1990 with as little as 50 of public URLs in 1992 [11], quickly grew to an estimated 1 trillion unique URLs today (Claimed by Google [10])! So, the question is

"*How can one build a WWW category tree*"?

There are several ways to do it. The simplest approach is to use a commercial service that provides access to URL category database. Websense[3] is one such organization that claims to provide more than 100 URL categories for millions of Web sites representing more than 50 languages. Another option is to use one of the many free webservices[4] that provide a public API key for accessing the master database. Unfortunately, most of these services limit the number of queries to 100 requests per day.

A more ambitious and academic approach is to build a custom crawler, mine the top x words along with their frequencies for all the pages on the Internet. By starting with a set of well-connected initial pages called *seeds*, the entire web can be traversed by following individual page links, recursively repeating the procedure for all the new links found on new webpages. Thereafter, there are several methods for classifying URLs ranging from using document clustering techniques [18, 19] to using URL names alone [48].

Though the author initially adopted a document clustering approach, it was later found that the resulting categorical tree was different, every time it was constructed. This is due to variable initial conditions (initial centroids, random number seed) in clustering algorithms. While different trees correspond to different ways of organizing the information, and is not necessarily a bad thing, the approach was abandoned as it introduces a lot of experimental variability, jeopardizing repeatability and empirical validation.

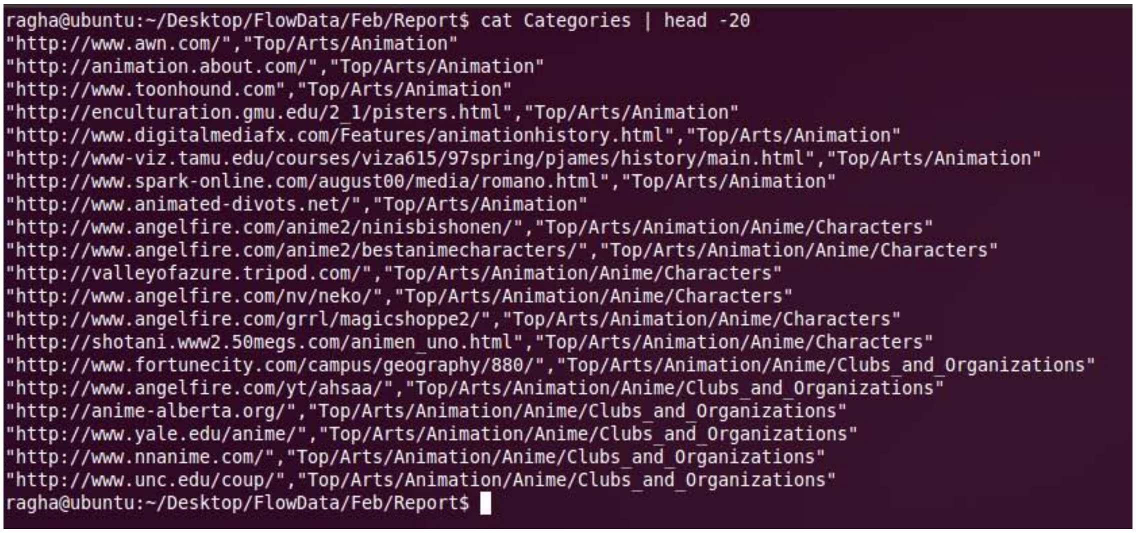
In order to maintain a consistent categorical tree, the URL categorization database available freely from the Open Directory Project (ODP) was utilized. The open directory initiative, accessible from `http://www.dmoz.org/`, describes itself as *the largest, most comprehensive human-edited directory of the Web, constructed and maintained by a vast, global community of volunteer editors.* The current open directory database, available in the form of an RDF (Resource Description Framework) dump can be downloaded from `http://www.dmoz.org/rdf.html`. An RDF file is an XML document containing several tags that needs to be parsed by a custom script.

In order to speed up the computations, the author wrote a full-fledged JAVA program to parse the RDF database onto a flat text file. In the output file, each line is of form $\langle URL\rangle, \langle Category\rangle$. In total, the file contains ∼4.3 million popular URLs along

[3] `http://www.websense.com/content/URLCategories.aspx`

[4] One such service is available at `http://www.trustedsource.org/en/feedback/url`

with the corresponding hierarchical category. In realtime operation, the entire tree is loaded onto the main memory to facilitate efficient distance computations. A snapshot of flat text database is shown in Figure 4.3.



```
ragha@ubuntu:~/Desktop/FlowData/Feb/Report$ cat Categories | head -20
"http://www.awn.com/","Top/Arts/Animation"
"http://animation.about.com/","Top/Arts/Animation"
"http://www.toonhound.com","Top/Arts/Animation"
"http://enculturation.gmu.edu/2_1/pisters.html","Top/Arts/Animation"
"http://www.digitalmediafx.com/Features/animationhistory.html","Top/Arts/Animation"
"http://www-viz.tamu.edu/courses/viza615/97spring/pjames/history/main.html","Top/Arts/Animation"
"http://www.spark-online.com/august00/media/romano.html","Top/Arts/Animation"
"http://www.animated-divots.net/","Top/Arts/Animation"
"http://www.angelfire.com/anime2/ninisbishonen/","Top/Arts/Animation/Anime/Characters"
"http://www.angelfire.com/anime2/bestanimecharacters/","Top/Arts/Animation/Anime/Characters"
"http://valleyofazure.tripod.com/","Top/Arts/Animation/Anime/Characters"
"http://www.angelfire.com/nv/neko/","Top/Arts/Animation/Anime/Characters"
"http://www.angelfire.com/grrl/magicshoppe2/","Top/Arts/Animation/Anime/Characters"
"http://shotani.www2.50megs.com/animen_uno.html","Top/Arts/Animation/Anime/Characters"
"http://www.fortunecity.com/campus/geography/880/","Top/Arts/Animation/Anime/Clubs_and_Organizations"
"http://www.angelfire.com/yt/ahsaa/","Top/Arts/Animation/Anime/Clubs_and_Organizations"
"http://anime-alberta.org/","Top/Arts/Animation/Anime/Clubs_and_Organizations"
"http://www.yale.edu/anime/","Top/Arts/Animation/Anime/Clubs_and_Organizations"
"http://www.nnanime.com/","Top/Arts/Animation/Anime/Clubs_and_Organizations"
"http://www.unc.edu/coup/","Top/Arts/Animation/Anime/Clubs_and_Organizations"
ragha@ubuntu:~/Desktop/FlowData/Feb/Report$ █
```

Figure 4.3. A snapshot of the flat text database. Every line in the file contains a unique URL, represented as $\langle URL\rangle, \langle Category\rangle$.

The author has open sourced the text database. It can be downloaded from `http://tinyurl.com/www-categories`. Interested researchers can download for free and use the file for any purpose needed.

4.3. LEVERAGING SIMILARITIES WITH EXISTING PROBLEMS

While the particular problem of depression classification has not been investigated before, there are several problems that bear resemblance to this problem at a mathematically abstract level and can therefore be adapted. In this section, three types of problem domains are introduced, along with details on necessary adaptations required to perform depression classification.

4.3.1. Similarities to Multi-Instance learning. In multi-instance (MI) learning, instead of receiving a set of feature vectors labeled positive or negative, the learner is provided with a set of *bags* – comprising of a sequence of feature vectors – labeled positive or negative [90]. In case of depression classification, netflow data $\mathcal{N}$ can be considered as a bag containing a set of flows $\{\overrightarrow{flow_1}, \overrightarrow{flow_2}, \ldots, \overrightarrow{flow_k}\}$.

The term multi-instance learning was first coined by Dietterich et al. [34] when they were investigating the problem of drug activity prediction. Given a set of bags labeled positive or negative, with each bag containing a set of instances, Dietterich was able to solve the multi-instance learning problem under standard MI assumption which is stated as follows:

> *A bag is labeled negative if all the instances in it are negative. On the other hand, a bag is labeled positive if there is at least one instance in it which is positive.*

Classification occurs by constructing an *axis-parallel hyper rectangle (APR)* in instance space such that it at least contains one instance from each positive bag, while excluding all the instances from negative bags. Initially, the algorithm begins with an "all-positive" APR that includes all instances of a positive bag. This APR, however, classifies many negative bags as positive. The task of the optimization algorithm is to shrink the all-positive APR until no more instances from negative bags are covered. The dotted box in Figure 4.4 indicates the shrunk APR. Shrinking is done by considering each bound and choosing a new bound that eliminates the highest number of instances from negative bags, thereby trying to leave as many instances from positive bags inside the APR as possible.

Multi-instance learning under standard MI assumption naturally applies to several problems, a few of which include image scene classification and video pattern recognition. In image scene classification, each image represented as a bag of pixels or subregions can be classified into positive or negative labels (indicating the image class). As the user labels an image as positive if the image somehow contains the concept, standard MI assumption applies [56, 91]. Application in video pattern recognition include kinematic action recognition [13] (similar to recognition performed by Microsoft Kinect and Nintendo Wii), labeling human faces in video feeds [83], and for incident retrieval in transportation surveillance video databases [29].

For the depression classification problem, the goal is to identify flow patterns exclusive to depressed people. i.e., the netflow bag is labeled positive (depressed) if it contains at least one positive $\overrightarrow{flow}$ instance and is, therefore, consistent with the standard MI assumption. However, for reasons mentioned in section 4.3.2, $\overrightarrow{cflow}$ is used instead. i.e., APR tries capture all concept instances $\overrightarrow{cflow}$ belonging to the depressed group.

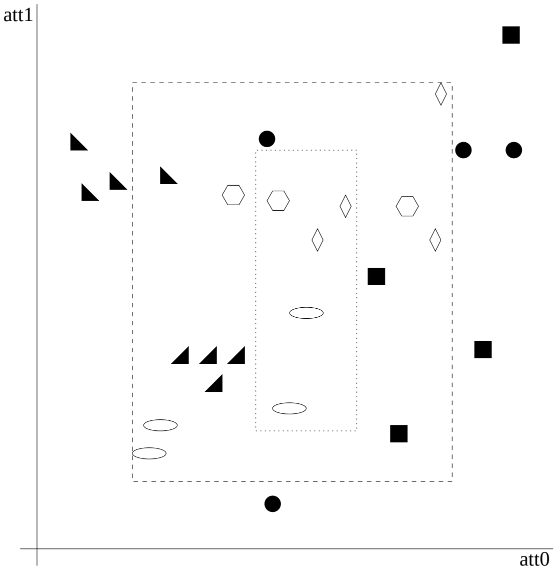


Figure 4.4. APRs in a two-dimensional instance space (adapted from [79]). Instances from the same bag appear in the same shape. Solid shapes are instances from negative bags. The dashed box denotes the initial all-positive APR. The dotted box is the final APR resulting after performing "shrinking" operation to exclude instances from negative bags.

Under standard MI assumption, a host of classification algorithms[5] such as Diverse Density [55], Citation-kNN and Bayesian-kNN [78], EM-DD [88], MI-optimal ball [20], and SVMs with MI-kernels [41] will be explored.

Another way to view the problem is to assume that a subset of $\overrightarrow{cflow} \in \mathbb{C}^f$ collectively constitutes towards depression. As MI-SVM kernels [17] do not make any assumptions, they can used to classify depression under collective assumption.

4.3.2. Adapting Document Classification for Depression classification. Documents contain a sequence of symbols conveying some information. The nature of information determines the topic, or topics pertaining to the document. The problem of document classification is to identify document categories (topics) automatically, given the sequence of words (symbols). In many ways, the problem of depression classification is very similar to document classification. In both cases, given a *sequence* of symbols (or inputs), the task is to map these sequences to a set of fixed classes $\mathbb{C} = \{c_1, c_2, \ldots, c_k\}$. In case of depression classification, the sequence of symbols is replaced by a sequence of

[5] It is beyond the scope of this thesis to present a detailed account of all these algorithms. However, interested readers may refer to section 4 in [90] for a unified view of all the approaches.

$\overrightarrow{flow}$ and $\mathbb{C} = \{0, 1\}$ corresponds to a depressed or non-depressed case. Thus, a naïve bayes classifier can be adapted to depression classification as follows:

The probability of occurrence of a flow $\overrightarrow{flow_i}$ in a depression class $D = \{0, 1\}$ is given by $Pr(\overrightarrow{flow_i} \mid D)$. Assuming that flows are independent of each other (which is not quite true) the probability that a given netflow data N contains all the flows $\overrightarrow{flow_i}$, given a depression class D is

$$Pr(N \mid D) = \prod_i Pr(\overrightarrow{flow_i} \mid D) \tag{12}$$

What is required is the probability that a given netflow data N belongs to a class in D, i.e., $Pr(D \mid N)$

$$\text{Using bayes theorem, } Pr(D \mid N) = \frac{Pr(D)}{Pr(N)} Pr(N \mid D) \tag{13}$$

Assuming mutually exclusive classes $D = 1$ (depressed) and $D = 0$ (non-depressed), such that every symbol ($\overrightarrow{flow_i}$) is either in one or another;

$$Pr(D = 1 \mid N) = \frac{Pr(D = 1)}{Pr(N)} \prod_i Pr(\overrightarrow{flow_i} \mid D = 1) \tag{14}$$

$$Pr(D = 0 \mid N) = \frac{Pr(D = 0)}{Pr(N)} \prod_i Pr(\overrightarrow{flow_i} \mid D = 0) \tag{15}$$

However, considering the fact that a $\overrightarrow{flow}$ represents a very tiny fraction of information regarding the user activity, coupled by the fact that $|N|$ approaches the order of magnitude of millions per participant, it is *highly unlikely* that $\overrightarrow{flow}$ alone sufficiently generalizes depression characteristics. i.e., it is very likely that $Pr(\overrightarrow{flow_i} \mid D = 1)$ and $Pr(\overrightarrow{flow_i} \mid D = 0)$ are very close to each other, thus rendering equations 14 and 15 void and useless.

Nevertheless, the problem can be avoided by clustering $\overrightarrow{flow}$ instances such that they represent a more general concept encompassing *similar* or *related* $\overrightarrow{flow}$ instances. Let $\overrightarrow{cflow}$ represent a general concept defined as follows:

Definition 4.3. Let $\overrightarrow{cflow}$ represent the general characteristics expressed by a set of $\overrightarrow{flow}$ F given by $\{\overrightarrow{flow_1}, \overrightarrow{flow_2}, \ldots, \overrightarrow{flow_k}\}$ such that $\frac{\sum_{i=1}^{k} \sum_{j=i}^{k} d(\overrightarrow{flow_i}, \overrightarrow{flow_j})}{\binom{k}{2}}$, where

$d(\overrightarrow{flow_i}, \overrightarrow{flow_j})$, referring to the context sensitive flow distance metric given by Equation **??**, is to be minimized while maintaining sufficient inter-cluster distance among the set of concepts $\mathbb{C}^f = \{\overrightarrow{cflow_1}, \overrightarrow{cflow_2}, \ldots, \overrightarrow{cflow_n}\}$

As N is now represented by the set $\mathbb{C}^f = \{\overrightarrow{cflow_1}, \overrightarrow{cflow_2}, \ldots, \overrightarrow{cflow_k}\}$, equations 14 and 15 can be re-expressed in terms of $\overrightarrow{cflow}$ given by:

$$Pr(D = 1 \mid N) = \frac{Pr(D = 1)}{Pr(N)} \prod_i Pr(\overrightarrow{cflow_i} \mid D = 1) \tag{16}$$

$$Pr(D = 0 \mid N) = \frac{Pr(D = 0)}{Pr(N)} \prod_i Pr(\overrightarrow{cflow_i} \mid D = 0) \tag{17}$$

Dividing equation 16 by 17 gives:

$$\begin{aligned} \frac{Pr(D = 1 \mid N)}{Pr(D = 0 \mid N)} &= \frac{Pr(D = 1) \prod_i Pr(\overrightarrow{cflow_i} \mid D = 1)}{Pr(D = 0) \prod_i Pr(\overrightarrow{cflow_i} \mid D = 0)} \\ &= \frac{Pr(D = 1)}{Pr(D = 0)} \prod_i \frac{Pr(\overrightarrow{cflow_i} \mid D = 1)}{Pr(\overrightarrow{cflow_i} \mid D = 0)} \end{aligned} \tag{18}$$

Expressing 18 as log likelihood ratio yields:

$$\ln \frac{Pr(D = 1 \mid N)}{Pr(D = 0 \mid N)} = \ln \frac{Pr(D = 1)}{Pr(D = 0)} + \sum_i \ln \frac{Pr(\overrightarrow{cflow_i} \mid D = 1)}{Pr(\overrightarrow{cflow_i} \mid D = 0)} \tag{19}$$

As $Pr(D = 1 \mid N) + Pr(D = 0 \mid N) = 1$, the netflow data N corresponds to a depressed case when

$$Pr(D = 1 \mid N) \quad > Pr(D = 0 \mid N) \tag{20}$$

i.e., when

$$\ln \frac{Pr(D = 1 \mid N)}{Pr(D = 0 \mid N)} \quad > 0 \tag{21}$$

Note that instead of $\overrightarrow{cflow}$, any of the eight attributes defined in equation 1 can also be used.

4.3.3. Leveraging similarities to time-series classification. Netflow data can be represented as a time-series since $N = \{\overrightarrow{flow_1}, \overrightarrow{flow_2}, \ldots, \overrightarrow{flow_k}\}$ represents an ordered sequence, where $\overrightarrow{flow_t}$ given by equation 1 is recorded during epoch t such that the observations in $\overrightarrow{flow_t}$ were recorded previous to those in $\overrightarrow{flow_{t+1}}$ for all $1 \leq t < k$.

In section 3.1, shannon entropy was used to capture one property of the time series - randomness. However, a very intuitive feature, namely, the *shape/trend* of time series is not captured by this metric. A popular approach to capture *shape* characteristics is to extract periodic sinusoidal basis functions from the signal, devoid of noise and unnecessary redundancies. In particular, given a uniform sample of a real signal $f(t) = (f(1), f(2), \ldots, f(k))$, fourier transform is used to express the signal as coefficients in a function space spanned by a set of complex exponential basis functions, representing sinusoidal components in the real domain. In some sense, DFT is analogous to principle components wherein a compressed representation of time series is obtained. The idea is illustrated in fig 4.5.

The Discrete Fourier Transform (DFT) is the projection of a signal from the time domain into the frequency domain by:

$$c_f = \frac{1}{\sqrt{n}} \sum_{t=1}^{k} f(t) \exp\left(\frac{-2\pi i f t}{k}\right) \tag{22}$$

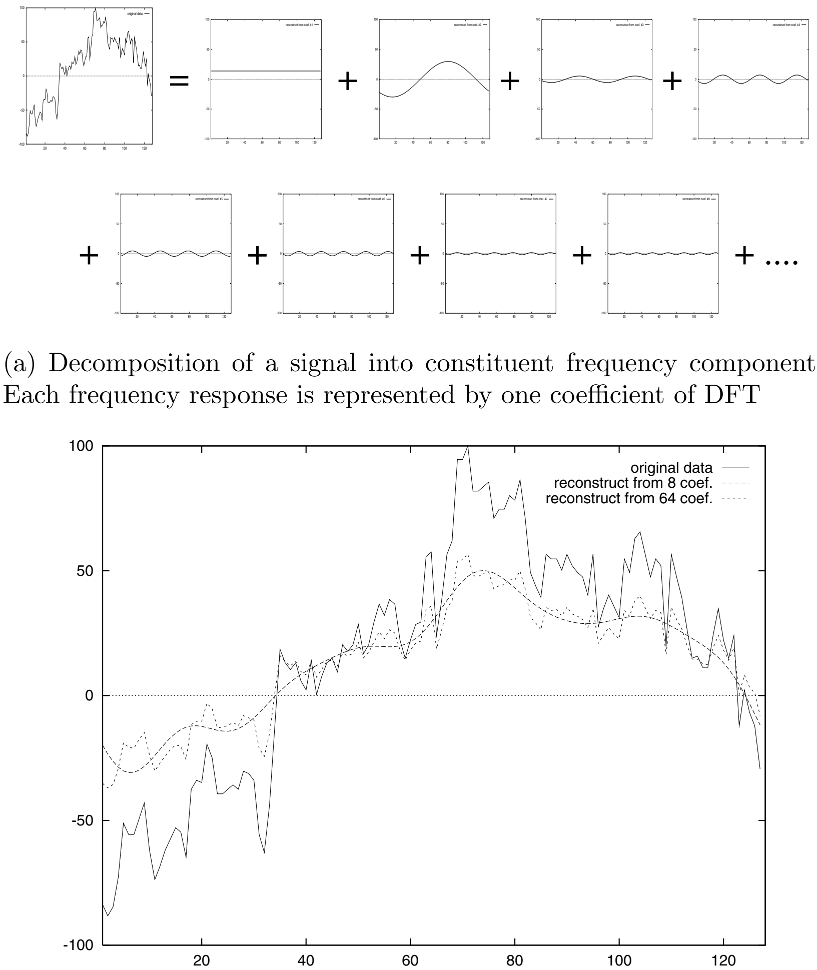
where, $f = 1, 2, \ldots, k$ and $i = \sqrt{-1}$. c_f is a complex number and represents the amplitudes and shifts of a decomposition of the signal into sinusoid functions. For real signals, c_i is the complex conjugate of $c_{n-i+1} (i = 2, 3, \ldots, \frac{k}{2})$. Further, the implementation utilized fast fourier transform (FFT), reducing the time complexity to $O(n \log n)$.

For each participant $i = [1, n]$, the netflow data N_i is used to derive eight time-series', one for each flow attribute given in equation 1. The number of coefficients for each participant $|c_f^i|$ is determined by preserving 80% energy of the signal. The energy of transformed signal is given by:

$$E^i(f(t)) = \sum_{j=1}^{|c_f^i|} a_j c_j \tag{23}$$

Where, a_j corresponds to appropriate scaling coefficients.

However, with this approach, the number of coefficients for each participant may be different and is dependent of the complexity of the flow attribute time-series. Therefore,



(a) Decomposition of a signal into constituent frequency components. Each frequency response is represented by one coefficient of DFT

(b) Reconstruction of the original signal using low frequency components. Note how reconstruction preserves the basic *shape* characteristics of the signal, in-spite of just using 8 coefficients

Figure 4.5. Illustration of Discrete Fourier Transform and the effects of signal reconstruction (adapted from [82])

in order to maintain a consistent size of the feature vector across all n participants, the total number of coefficients for each participant is given by:

$$|\text{Coefficients}| = \min(|c_f^1|, |c_f^2|, \ldots, |c_f^n|) \tag{24}$$

4.4. OVERVIEW OF MACHINE LEARNING TECHNIQUES USED

In this section, an overview of potential supervised learning techniques are described. As the input feature vector comprises of level 1,2,3 feature vectors (refer to 3.1) plus the DFT coefficients, the dimensionality is rather large. Therefore, instead of using Neural Networks, Support Vector Machines (SVM) were adopted. The theoretical properties of SVMs allows them to handle high dimensional spaces efficiently. Additionally, ARTMAPs are considered for their online and incremental learning property. The remainder of this section provides a brief introduction to both the methods.

4.4.1. Support vector machines. Given some training data $\mathcal{D}$ with n examples of form $\mathcal{D} = \{(x_i, y_i) \mid x_i \in \mathbb{R}^p, y_i \in \{-1, 1\}\}_{i=1}^{n}$, where x_i denotes the input and y_i denotes the class label $\{-1, 1\}$, the idea behind Support Vector Machines (SVMs)[6] is to build lines, planes or hyperplanes in input space for separating the classes $y_i = \pm 1$ with maximum possible margin. Intuitively, maximum margin guarantees the best generalization.

Consider two hyperplanes:

$$\vec{w} \cdot \vec{x}_i + b \quad \geq 1, \forall \vec{x}_i \ni y_i = +1 \tag{25}$$

$$\vec{w} \cdot \vec{x}_i + b \quad \leq -1, \forall \vec{x}_i \ni y_i = -1 \tag{26}$$

Given an unseen input x_i, the distance from the hyperplane to a point x_i is given by:

$$d(\vec{w}, b; x_i) = \frac{|\langle \vec{w}, x_i \rangle + b|}{\|w\|} \tag{27}$$

Consequently the margin between two hyperplanes can be written as :

$$\min_{x_i; y_i = 1} d(\vec{w}, b; x_i) + \min_{x_i; y_i = -1} d(\vec{w}, b; x_i) \tag{28}$$

Intuitively, high generalization performance is achieved when the margin between the hyperplanes separating positive and negative cases is maximized. i.e., $\|w\|$ needs to be minimized. This comes down to solving a quadratic optimization problem with linear constraints. Notice however that the data is assumed to be perfectly linear separable. In practice, this will often not be the case. By employing the so called *soft-margin* method

[6] An excellent survey of SVMs can be found in [22]

in contrast to the *hard-margin* method, the SVM is allowed to perform some misclassifications, while reducing an *upper bound* on the expected generalization error [76]. Omitting further details, the soft-margin optimization problem can be stated in the dual form, i.e., the task is to find the Lagrange multipliers $\alpha_i > 0$, $\forall i = [1, N]$, so that:

$$\text{Maximize} \quad : L(\alpha_1, \ldots, \alpha_N) = \textstyle\sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \tag{29}$$

$$\text{Subject to} \quad : \textstyle\sum_{i=1}^{N} \alpha_i y_i = 0,\ 0 \leq \alpha_i \leq C$$

Considering the dual problem in Equation 29, the maximum margin hyperplane as a linear combination of support vectors. By definition the vectors x_i corresponding with non-zero α_i are called the support vectors SV and this set consists of those data points that lie closest to the hyperplane and thus are the most difficult to classify (see Figure 4.6 for an illustration).

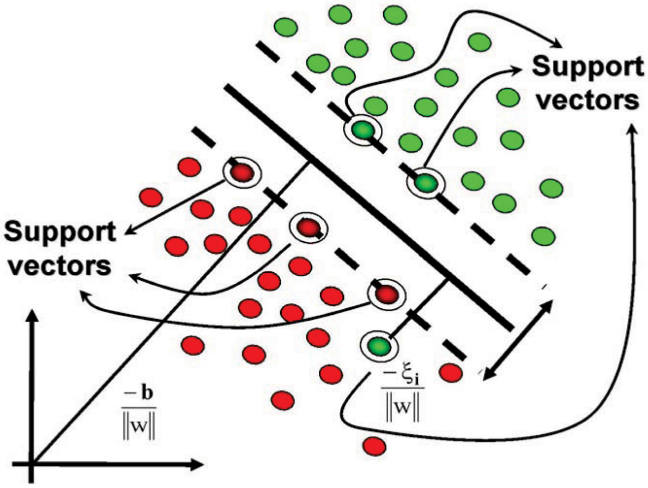


Figure 4.6. Illustration of support vectors and maximum margin hyperplane in 2D feature space. Red and green circles distinguish the data points belonging to two different classes.

In order to classify a new point x_{new}, one has to determine the sign of:

$$\sum_{x_i \in SV} \alpha_i y_i \langle x_i, x_{new} \rangle + b \tag{30}$$

If this sign is positive x_{new} belongs to class 1. It belongs to class -1 otherwise. Note that the summation is restricted to set SV of support vectors because the other α_i are zero anyway. This means that the mdoel complexity of an SVM is unaffected by the number of features encountered in the training data. For this reason, SVMs are well suited to deal with learning tasks where the number of features is large with relative to the number of training instances.

The Kernel trick: The discussion so far assumes that the data is linearly separable. In practice it will often be the case that the data can not be separated linearly by means of a hyperplane. In such situations, one the basic ideas behind SVMs is to use a function Φ to map the data onto a higher dimensional space such that non-linear separation in lower dimensional *input space* X is transformed into a linear separation in some arbitrarily higher dimensional space F (possibly infinite dimensional Hilbert space H $\Phi : \mathbb{R}^n \rightarrow H$) (see figure 4.7 for an illustration).

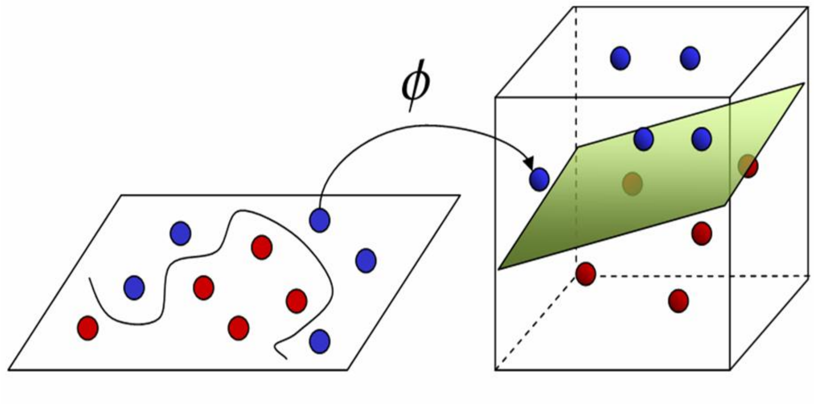


Figure 4.7. Illustration of Φ, mapping data points in 2-dimensional space onto a 3-dimensional transformed feature space.

However, transforming the input vectors in the into such a higher-dimensional space incurs computational problems. The high dimensionality of F makes it very expensive both in terms of memory and time to represent the feature vectors $\Phi(x_i)$ corresponding to the training vectors x_i. Moreover, it is not a trivial task to find the transformation Φ that linearly separates the transformed data.

Luckily, all computations in SVM involve dot products of from $\Phi(x_i) \cdot \Phi(x_j)$ as the objective function $L(\alpha_1, \ldots, \alpha_N)$ in Equation 29 and the definition of the hyperplane in

Equation 30 depend only on inner products between vectors. Therefore, one can avoid the need to determine Φ by using a kernel function of form $K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$. This idea is commonly termed as the *Kernel trick*[7].

Thus, kernels are a special class of functions that allow inner products to be calculated directly in feature space, without performing the mapping described above [68]. When a kernel is used, the decision rule in Equation 30 is modified to:

$$\operatorname{sgn}\left(\sum_{x_i \in SV} \alpha_i y_i K(x_i, x_{new}) + b\right) \tag{31}$$

Genton [38] described several classes of kernels, however, he did not address the question of which class is best suited to a given problem. It is common practice to estimate a range of potential settings and use cross-validation over the training set to find the best one. This thesis will explore the following kernels:

1. Linear: $K(x_i, x_j) = x_i^T \cdot x_j$
2. Polynomial: $K(x_i, x_j) = \left(\gamma x_i^T \cdot x_j + b\right)^d$
3. Radial basis function (RBF): $K(x_i, x_j) = \exp(-\gamma \left\|x_i - x_j\right\|^2)$, for $\gamma > 0$. In this case, the corresponding feature space is a Hilbert space of infinite dimensions.

4.4.2. Adaptive resonance theory. While SVM represents a good choice of classifier as the number of features are large relative to the number of samples, it is static in nature. i.e., if one were to obtain more data samples, the SVM will have to be re-trained from the start. As this study utilized relatively low number of samples, one of the immediate future task is to consider more data samples. Therefore, it is desirable to build a classifier that is *robust* and facilitates *incremental* learning, often resulting in lesser training time and yields better performance [43]. Incremental learning is particularly critical to depression classifier due to the sheer volume of netflow and due to the heavy computational overhead involved in pre-processing netflow.

A key prerequisite to achieving incremental learning is to solve the so-called "*stability-plasticity*" problem [62]. Plasticity is the ability of a network to learn new patterns in response to new stimuli whereas stability is the ability of a network to retain previously

[7] An excellent overview of kernel based methods can be found in [45]

stored patterns by preventing significant interactions among them. Solving the stability-plasticity problem leads to the property of incremental learning since the network is able to learn new patterns without interfering with stored representations of other patterns.

In this thesis, the ARTMAP network will be used as it is inherently capable of incremental learning. In particular, two types of ARTMAP architectures are explored - Fuzzy ARTMAPs (FAM) and Ellipsoidal ARTMAPs (EAM). The remainder of this section will introduce ARTMAPs, and provide a brief intuition behind FAM and EAM.

- **ARTMAPs:** The Adaptive Resonance Theory (ART), introduced by Grossberg in the 1970s, began with the analysis of human cognitive information processing [25, 40]. It led to the creation of a family of self-organizing neural networks for fast learning, pattern recognition, and prediction. Some popular members of the family are both unsupervised models (i.e., ART1, ART2, ART2-A, ART3, fuzzy ART, distributed ART) and supervised models (i.e., ARTMAP, instance counting ARTMAP, fuzzy ARTMAP, distributed ARTMAP, and default ARTMAP) [24].

  ARTMAP is a supervised neural network that consists of two unsupervised ART modules, ARTa and ARTb, and an inter-ART module called a map-field (see Figure 4.8) [27]. The ARTa module clusters patterns of the input domain and ARTb clusters the ones of the output domain. The ART modules form categorical representations in input and output domain and is therefore equivalent to finding clusters in the corresponding domains. The inter-ART module is responsible for mapping associations from input ARTa recognition categories onto the output ARTb categories. Thus, unlike neural networks which learn to associate every input exemplar a_i with output t_i, ARTMAP learn to identify invariant similarities among input vectors, and associates it with an output category. The information regarding the input-output associations is stored in the weights w_j^{ab} of the inter-ART module.

  An essential feature of the ARTMAP design is its ability to conjointly *maximize generalization* and *minimize predictive error*. During the learning process, a mismatch at the map field between the ARTa category activated by an input a and the ARTb category activated by the output t increases ARTa vigilance by a minimum amount needed for the system to search for, and if necessary, learn new ARTa categories whose prediction matches the ARTb category. The ARTMAP categories are stored in the template vectors w_j contained in the top-down weights of the

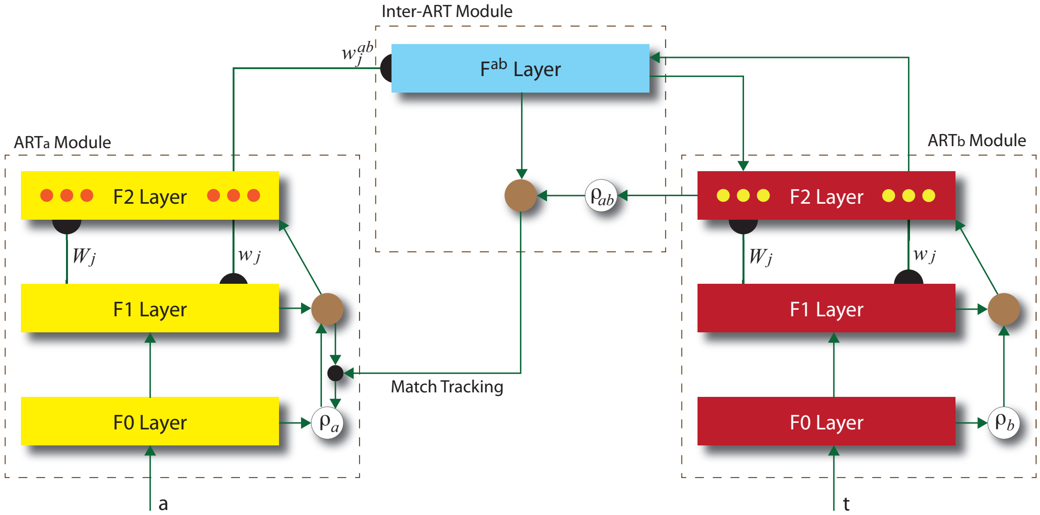


Figure 4.8. Block diagram of an ARTMAP neural network architecture (adapted from [27])

F2-layer nodes in each module. A detailed description of ARTMAP learning can be found in [27].

- **Fuzzy ART:** Fuzzy ARTMAP [26] is merely ARTMAP using fuzzy ART units, resulting in a corresponding increase in efficacy. The crisp non-fuzzy intersection ($\cup$) and union ($\cap$) that describe the ART1 dynamics are simply replaced by fuzzy OR ($\vee$) or AND ($\wedge$) operators of fuzzy set theory [87]. This allows the fuzzy ARTMAP to learn stable categories in response to either analog or binary patterns in contrast with the basic ARTMAP, which operates with binary patterns only.

- **Ellipsoidal ARTMAPS:** EAMs [15] came as an enhancement and generalization of ARTMAPs , which, in turn, follow the same learning and functional principles of Fuzzy ART (FA) and Fuzzy ARTMAP (FAM) [26]. Unlike ARTMAPs that use hyper-rectangles for categorical representation, EAM uses hyper-ellipsoids, a powerful geometrical generalization that allows them to capture and learn complex decision boundaries. A Hyper-ellipsoid embedded in the feature space represents a category which encodes whatever information the EAM classifier has learned about the presence of data and their associated class labels in the locality of its geometric representation. This information is encoded into the location and size (major and minor axis) of the hyper-ellipsoid. A comparison between 3-dimensional Hyper-rectangle and EA category is given in Figure 4.9.

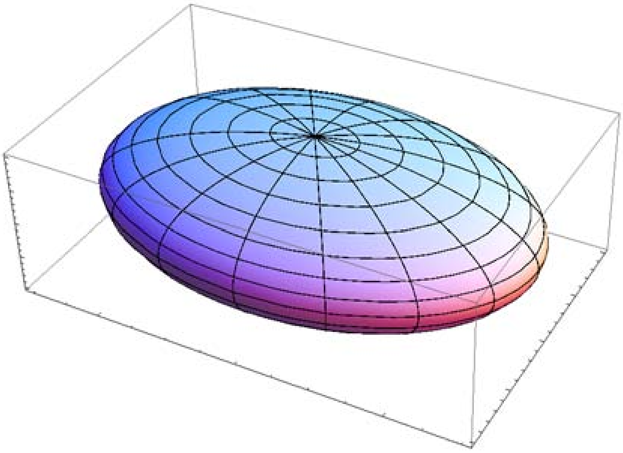


Figure 4.9. Illustration of 3-dimensional FAM and EAM categories. The gray cuboid represents the hyper-rectangle.

4.5. CONTENT AWARE SVM KERNEL

In this section, a specialized SVM kernel, leveraging on the context aware netflow distance metric given by Equation 11 is described.

In support vector machines, one of the crucial ingredients is the so-called kernel trick for the computation of dot products in high-dimensional feature spaces using simple functions defined on pairs of input patterns (please refer to section 4.4.1 for a concise overview of SVMs). In support vector machines, similarity information is implicitly contained in the kernel function. Consider the training data $\mathcal{D} = \{(x_i, y_i) \mid x_i \in \mathbb{R}^p,\ y_i \in \{-1, 1\}\}_{i=1}^{n}$, where x_i denotes the input and y_i denotes the class label $\{-1, 1\}$. When the input space $\mathcal{X} \in \mathbb{R}$, the dot product between two input vectors x_i, x_j is related to the geometric notion of length and orientation (angle). Clearly, more the angle between x_i and x_j, larger is the Euclidean distance between them. Similarly, length also affects the Euclidean distance between these two points. Since the Euclidian distance, or any other $\|.\|_p$ for that matter, is related to the notion of dissimilarity. Therefore, a kernel expresses prior knowledge about the patterns being modeled, encoded as a (dis)similarity measure between two input vectors.

Not all symmetric functions over $X \times X$ are kernels that can be used in a SVM. Since a kernel K is related to an inner product, it has to satisfy some conditions that arise naturally from the definition of an inner product and are given by Mercers theorem [32], which states that *the kernel function has to be positive definite (PD).* PD kernels are defined as:

Definition 4.4 (Positive Definite Kernel)**.** A symmetric function $K : X \times X \to \mathbb{R}$ which for all $m \in \mathbb{N}$, $x_i, x_j \in \mathcal{X}$ gives rise to a positive semi-definite (PSD) kernel matrix, i.e., for which for all $c_i \in \mathbb{R}$ satisfies:

$$\sum_{i,j=1}^{m} c_i c_j K_{ij} \geq 0, \text{ where } K_{ij} = K(x_i, x_j) \tag{32}$$

is a positive definite (PD) kernel.

When K is not PD the convexity of the optimization problem can no longer be guaranteed.

Although PD kernels define inner products, they can also use a form of dissimilarity in their calculation. A good choice for this type of kernel is the radial basis kernel $K_{rb} : X \times X \to \mathbb{R}$ that explicitly makes use of the distance between two points in the input space X. It is given by:

$$K_{rb}(x_i, x_j) = \exp(-\gamma \left\| x_i - x_j \right\|^2), \text{ for } \gamma > 0 \tag{33}$$

By using content aware netflow distance metric $d(\mathcal{N}_i, \mathcal{N}_j)$ given in Equation 11, the resulting RBF kernel can be specialized for depression classification. Figure 4.10 illustrates the gaussian function. As X-axis corresponds to $d(\mathcal{N}_i, \mathcal{N}_j)$, whenever the distance is zero, the resulting value is maximum, which is consistent with the notion of high similarity. Similarly, when $d(\mathcal{N}_i, \mathcal{N}_j)$ is large, K_{rb} is small. By changing σ^2, one can control the *width* of the curve. If σ^2 is small, $d(\mathcal{N}_i, \mathcal{N}_j)$ is only defined for extremely similar samples. With larger σ^2, it is possible to compare $\mathcal{N}_i, \mathcal{N}_j$, that are far apart.

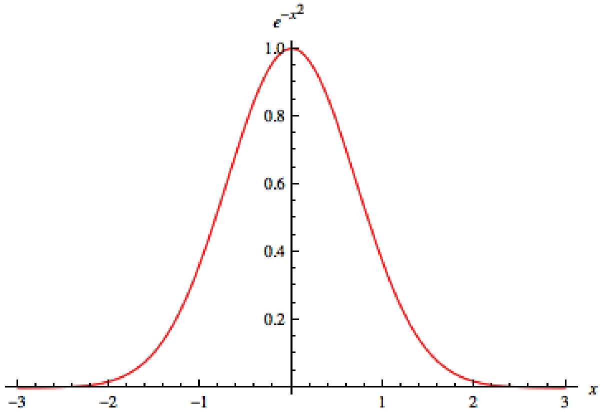


Figure 4.10. Illustration of the generalized radial basis kernel

4.6. PERFORMANCE EVALUATION

The performance of an inductive learning method is its predictive accuracy on unseen data. Typically, an inductive learning algorithm is trained on a set of training examples. The algorithm is considered good, if its results can successfully be *generalized*, i.e. if it makes correct predictions for unseen cases. One approach to assess the generalization capability of a classifier is by diving the data into training, testing and validation data sets. The training set is used to build a classification model, while the validation set is used to assess its generalization capabilities. This is done in order to avoid overfitting with respect to training data. Since the procedure can itself lead to some overfitting to the validation set, the performance of the classifier is confirmed by measuring its performance on test set.

However, due to the limited number of participants, the above mentioned train, validation and test procedure wont be reliable. For the purpose, 10-fold cross validation is used. First, the data is split into 10 disjunct, roughly equal-sized sets called *folds*. The performance of the classifier is determined by averaging its performance over 10 runs. In each run, one of the 10 sets is held out, and the classifier trained on the remaining 9 sets and evaluated on the held out set.

In a standard cross-validation, the examples are assigned randomly to folds and may not represent the class distribution of the dataset. For instance, one fold may have more positive examples than negative ones when compared to the class distribution on the entire dataset. This may lead to problems with classifiers that rely on apriori distribution of output classes - For instance, the naïve bayes classifier. For this reason, a variation of the standard cross-validation called "stratified" cross-validation is used. Here, the class distribution in each fold is roughly the same as the original dataset.

In this thesis, three types of performance measures, namely *accuracy*, *precision* and *recall* are used. These measures have been widely used in the data mining literature to evaluate data classification algorithms [81]. First, the output of the depression classifier is categorized into four groups:

1. **True Positive (tp)**: Total number of cases correctly classified as depressed.
2. **False Positive (fp)**: Total number of cases falsely classified as depressed (also known as Type I Errors).
3. **False Negative (fn)**: Total number of cases falsely classified as non-depressed (also known as Type II errors).
4. **True Negative (tn)**: Total number of cases correctly classified as non-depressed

These four categories are aptly summarized in table 4.1.

Table 4.1. Confusion matrix for the binary depression classification problem

True class label	Predicted class label	
	Depressed	Non-depressed
Depressed	*tp*	*fn*
Non-depressed	*fp*	*tn*

Precision is defined as the ratio of true positives to true and false positives. It measures the *exactness* of classifier. A higher precision means less false positives, while a lower precision means more false positives. This is often at odds with recall, as an easy way to improve precision is to decrease recall. It is given by:

$$Precision = \frac{tp}{tp + fp} \tag{34}$$

Recall is defined as the ratio of True Positives to True and False Negatives. It measures the *completeness*, or *sensitivity*, of a classifier. Higher recall means less false negatives, while lower recall means more false negatives. Improving recall can often decrease precision because it gets increasingly harder to be precise as the sample space increases. Recall is given by:

$$Recall = \frac{tp}{tp + fn} \tag{35}$$

Finally, overall accuracy is given as the ratio of all True Positives to all True and False Positives for all classes. Unlike precision and recall which are per-class measures, overall accuracy measures the average precision of *all classes* and is given by:

$$Accuracy = \frac{\sum_{i=1}^{n} tp_i}{\sum_{i=1}^{n}(tp_i + fp_i)} \tag{36}$$

where, n is the number of classes. For the depression classifier, n=2 (depressed/non-depressed classes).

In summary, accuracy is an overall measure of classifier performance. Precision deals with the classifier's tendency to make mistakes by over-generalizing, and recall is the ability to correctly classify all instances belonging to a certain class.

4.7. RESULTS

In this section, the performance of various classifiers listed in Table 4.2 is described. In total, nine classifiers, grouped into three learning categories, are evaluated.

Table 4.3 shows the average precision, recall and accuracy for all nine classifiers trained using the 10-fold cross validation method.

In general, it can be observed that the overall accuracy is roughly around 70% on average, which is a lot better when compared to random guessing, thereby demonstrating the feasibility of using nerflow for depression classification.

For the supervised learning group trained using three levels of feature vectors and the DFT coefficients, it is not surprising that the SVM had the best performance (good precision/recall tradeoff), considering the high dimensionality of the input feature space. Among the ARTMAP classifiers, EAM performed better when compared to FAM, demonstrating the efficacy of ellipsoidal categories over hyper-rectangles.

In the multi-instance learning group, MI-SVM performed a lot better when compared to the other classifier, especially in terms of recall for *DEP* class. The author attributes this improvement to collective assumption, as all MI classifiers with standard

Table 4.2. List of classifiers evaluated

Group	**Classifier (Input)**	**Symbol**
Supervised Learning (Three levels of feature vectors + DFT coefficients + gender)	Support Vector Machine	SVM
	Fuzzy ARTMAP	FAM
	Ellipsoid ARTMAP	EAM
Multi-Instance Learning (Netflow sets + gender)	Diversity Density	MI-DD
	Citation-KNN	MI-CKNN
	MI-Optimal Ball	MI-OB
	MI-SVM (Collective assumption)	MI-SVM
Custom Classifiers (Netflow sets + gender)	Probabilistic Classifier (Section 4.3.2)	PC
	Content aware SVM Kernel (Section 4.5)	ConSVM

Table 4.3. Performance metrics for all nine classifiers

Classifier	**Class**	**Precision**	**Recall**	**Overall Accuracy**
SVM	DEP	57.63%	66.67%	74.55%
	NOT_DEP	83.96%	78.07%	
FAM	DEP	58.14%	49.02%	73.33%
	NOT_DEP	78.69%	84.21%	
EAM	DEP	58.93%	64.71%	75.15%
	NOT_DEP	83.49%	79.82%	
MI-DD	DEP	51.92%	52.94%	70.30%
	NOT_DEP	78.76%	78.07%	
MI-CKNN	DEP	61.29%	37.25%	73.33%
	NOT_DEP	76.12%	89.47%	
MI-OB	DEP	63.33%	37.25%	73.94%
	NOT_DEP	76.30%	90.35%	
MI-SVM	DEP	52.38%	64.71%	70.91%
	NOT_DEP	82.35%	73.68%	
PC	DEP	66.67%	50.98%	76.97%
	NOT_DEP	80.16%	88.60%	
ConSVM	DEP	56.06%	72.55%	73.94%
	NOT_DEP	85.86%	74.56%	

MI assumption performed rather poorly. i.e., they failed to extract generalizations for *DEP* class, and over-learned *NOT_DEP* class.

For the custom classifier category, the probabilistic classifier incurred high generalization error. As PC involves bayesian learning involving probabilistic estimates, the error is likely due to small sample size. *ConSVM* on the other hand, had the best balance between precision and recall for the *DEP* class with an overall accuracy of ∼74%. This result illustrates the importance of content aware netflow distance metric and its relation to depression.

4.8. SUMMARY OF CONCLUSIONS

In this chapter, it was shown that depression can be classified with ∼74% accuracy by leveraging on patterns in netflow data. Being better than random guessing, the result demonstrates the feasibility of using netflow for depression classification.

By observing the results, two major conclusions can be drawn. For the depression classification, all $\overrightarrow{flow} \in \mathcal{N}$, influences the depression class. This was demonstrated by the fact that the MI-SVM algorithm had better performance when compared to other standard MI algorithms. Additionally, *ConSVM* demonstrated that the *type of browsing activity* is crucial for estimating the depression class.

5. CLASSIFICATION USING BASELINE DEVIATION PATTERNS

Although Chapter 4 presented depression classification models with up to 74% accuracy, it was developed under the simplifying assumption, considering depression scores to be static. In reality, however, depression is continuous time variable and is usually periodic in nature. In this chapter, the continuous nature of depression is accounted by taking repeated measures of depscore over fixed time-intervals. Additionally, the author hypothesizes that the classification accuracy can be improved by considering the deviation patterns in baseline Internet activity. By considering baseline deviations over absolute trends, one can account for person specific differences.

This chapter opens with a brief description of motivation for studying baseline deviation patterns. Analogous to chapter 3 and 4, the remainder of this chapter describes the statistical and classification results for the baseline deviation approach.

5.1. MOTIVATION FOR STUDYING BASELINE DEVIATIONS

The main premise of this chapter is the hypothesis that baseline deviations can better explain depression. For example, consider the chat activity of a depressed social and anti-social person. When analyzed by considering absolute chat activity, trends observed in both persons might cancel each other out. Instead, it is more beneficial to identify deviation trends such as the increase or decrease in *baseline* chat activity, which can then be correlated with the change in depscore. As depression is considered as a continuous time variable, the deviation trends for different time windows t_i–t_{i+1} can be correlated with $depscore_{i+1} - depscore_i$. While correlations are used to gain statistical insight, classification techniques are utilized to identify the invariant baseline deviation patterns that are common to $\Delta depscore > 0$, but not present in $\Delta depscore <= 0$.

In considering depression as a continuous time variable, it is important to note that this fact does not necessarily invalidate the credibility of work presented in chapter 4. Classification models presented in chapter 4 are built by assuming depscore to be static within a monthly window. In fact, when depression scores were reassessed for the subsequent months, scores of very few participants changed. Figure 5.1 shows the depression change in participants from the month of February to March. Overall, depression score difference `depscorediff` exhibited $\mu = -2.11$ and $\sigma = 8.72$

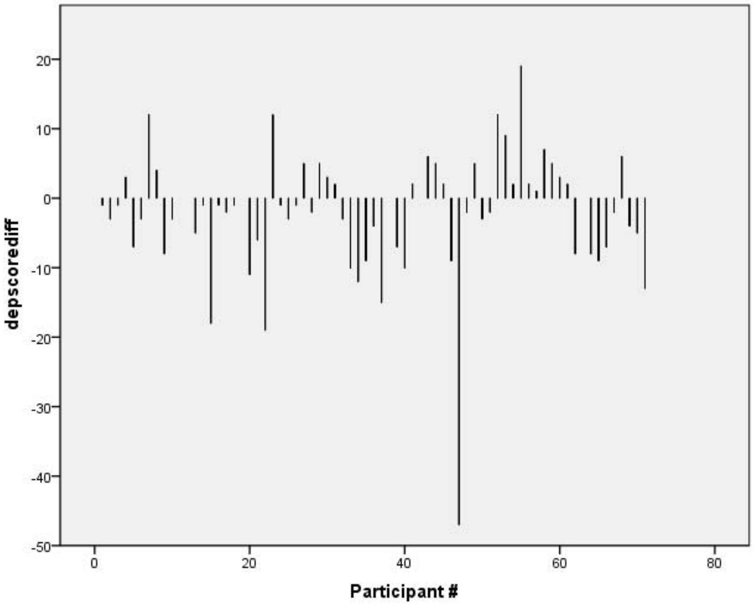


Figure 5.1. Depression change in participants. The change is computed by considering the difference in `depscore` for two consequent months.

The premise of this research is the hypothesis that the Internet usage patterns are correlated with depression. Results from chapter 3 and 4 confirms the hypothesis. By considering *depscore* as a continuous time variable, an attempt is made to uncover the relationship between the *change* in Internet activity and the *change* in depscore. As change is measured, an attempt is made to build a classification model that can predict the increase or decrease in `depscore` based on deviations in baseline Internet activity of a user.

5.2. CORRELATION RESULTS

In this section, the correlations between `depscorediff` and the change in Internet traffic features are presented. Due to time constraints at the time of writing this thesis, the discussion is limited to level 1 and level 2 feature vectors alone (See section 3.1 for a detailed description). Both `depscorediff` and the attributes of the feature vector are modeled as binary variables with the following meaning:

$$B(var) = \begin{cases} 1, & \text{if } var > 0 \\ 0, & \text{otherwise} \end{cases} \tag{37}$$

As ordinal variables are used, Gamma and Tau-b correlations are considered. Pearson correlation is not computed as it is better suited for continuous variables.

5.2.1. Aggregate Traffic Features. Correlations (gamma and tau) between `depscorediff` and the change in aggregate traffic features are shown in Table 5.1.

Table 5.1. Correlations between depression change and change in aggregate traffic features

Traffic Features	*depchange*	
	Gamma	**tau-b**
flows	-.103	-.031
oct	-.049	-.021
pkts	.170	.080
timeflows	.021	.017
durreal	.222	.103
durdata	.029	.015
aftime	.014	.013
apsize	.138	.063
afsize	-.456	-.107
apflow	-.211	-.043
afsec	-.400*	-.158*
afsecreal	-.500*	-.166*
akbits	-.054	-.022
akbitsreal	-.286	-.125

***. Correlation is significant at 0.01 level (2-tailed)*
**. Correlation is significant at 0.05 level (2-tailed)*

Both correlation coefficients revealed a significant negative correlation between the change in depression score and the change in the average number of flows per second. Therefore, it may be concluded that "*An increase in depression score is significantly correlated with a decrease in the average number of flows per second.* Decrease is the average number of flows can be attributed to a multitude of reasons. To provide a reasonable interpretation, additional experimentation needs to be performed.

5.2.2. Application usage statistics. Correlations (gamma and tau) between `depchange` and the change Internet application usage is described in Table 5.2. Both coefficients showed a significant negative correlation between `depchange` and change in http_packets, http_duration, streaming_duration. Conversely, a positive correlation was observed between `depchange` and change in ftp_packets, and ftp_duration.

Table 5.2. Correlations between depression change and change in Internet application usage

Application Type	*depchange*	
	Gamma	**tau-b**
p2p_octets	.312	.127
p2p_packets	.063	0.03
p2p_duration	.024	.011
http_octets	-.018	-.007
http_packets	-.368*	-.189*
http_duration	-.356*	-.182*
streaming_octets	-.267	-.099
streaming_packets	-.130	-.059
streaming_duration	-.438*	-.222*
chat_octets	.063	.029
chat_packets	-.029	-.014
chat_duration	-.154	-.076
mail_octets	.098	.047
mail_packets	.125	.062
mail_duration	.263	.133
ftp_octets	.328	.151
ftp_packets	.412*	.188*
ftp_duration	.417*	.198*
game_octets	.073	.025
game_packets	-.106	-.045
game_duration	.007	.003
remote_octets	.147	0.06
remote_packets	-.070	-.032
remote_duration	.025	.013

**. Correlation is significant at 0.01 level (2-tailed)
*. Correlation is significant at 0.05 level (2-tailed)

A negative correlation between depchange with http packets and duration indicates that there was a significant increase in depscore that co-occurred with the decrease in http usage. Curiously, http octets did not differ significantly. By considering duration into account, one way to interpret this result is to conclude that “*Increase in depression* *is correlated with a significant decrease in http browsing activity*". However, it is not completely clear as to how the combination of octets, packets, and duration relate to the change in user activity.

The negative correlation between depchange and streaming_duration leads to a straight forward conclusion that "*Increase in depression has a significant co-occurrence with the decrease in time spent on streaming activity*".

Another surprising result is the positive correlation between depchange with ftp packets and duration. Analogous to http activity, it is peculiar to note that ftp octets did not deviate significantly. As with http, one way to interpret this result is to conclude that the "*Increase in depression has a significant co-occurrence with an increase in ftp usage activity*". Since college students are involved, one way to account for this result is to interpret that the increased ftp usage is correlated with the increase in assignments or workload given to students. Nevertheless, these are mere hypotheses and needs to be validated with further experimentation.

5.3. CLASSIFICATION RESULTS

Chapter 4 described several classification techniques to build a model model $M : \mathcal{N} \rightarrow D$. In this section, the same set of classification techniques are applied to build a model $M : \Delta\mathcal{N} \rightarrow \Delta D$. Therefore, the task is to estimate the likelihood of increase or decrease in depression by analyzing deviations in user netflow activity. The list of classifiers evaluated can be referred from Table 4.2.

In the supervised learning group, change in all three levels of feature vectors were considered. Additionally, DFT computations were performed on difference in the netflow time series. i.e., given the netflow data for months m and $m+1$, DFT coefficients were extracted using the difference time-series $\mathcal{N}_{m+1} - \mathcal{N}_m$. The difference time series was also utilized in Multi-instance learning and custom classification groups.

Unlike the construction of an absolute depression classifier where 165 samples were used, this experiment was limited to 72 samples only. The sample size was reduced as all the participants were not available for the reassessment of depression scores. Amongst the 72 participants, the depscore for 24 participants increased while it decreased or remained unchanged for the remainder 48 participants. Due to the low sample size, 10 fold CV approach was not used as it divides the data into relatively few groups. Instead, leave one out cross-validation (LOOCV) was used. In LOOCV [49], a single observation is removed and the remaining data is used as the training data set. The process is repeated for all data samples and the results are averaged.

Table 5.3 shows the average precision, recall and accuracy for all nine classifiers.

Table 5.3. Performance metrics for all nine classifiers predicting the change in depression

Classifier	**Class**	**Precision**	**Recall**	**Overall Accuracy**
SVM	DEP	62.96%	70.83%	76.39%
	NOT_DEP	84.44%	79.17%	
FAM	DEP	65.22%	62.50%	76.39%
	NOT_DEP	81.63%	83.33%	
EAM	DEP	58.06%	75.00%	73.61%
	NOT_DEP	85.37%	72.92%	
MI-DD	DEP	68.18%	62.50%	77.78%
	NOT_DEP	82.00%	85.42%	
MI-CKNN	DEP	71.43%	41.67%	75.00%
	NOT_DEP	75.86%	91.67%	
MI-OB	DEP	73.33%	45.83%	76.39%
	NOT_DEP	77.19%	91.67%	
MI-SVM	DEP	66.67%	75.00%	79.17%
	NOT_DEP	86.67%	81.25%	
PC	DEP	78.57%	45.83%	77.78%
	NOT_DEP	77.59%	93.75%	
ConSVM	DEP	75.00%	75.00%	83.33%
	NOT_DEP	87.50%	87.50%	

In general, it can be observed that the overall accuracy is roughly around 76% on average, which is a stark improvement when compared to the performance of classification models presented in chapter 4. This shows that baseline deviations are indeed more general and useful when compared to absolute Internet activity.

The results from all three groups are more or less consistent with the conclusions claimed in chapter 4. For the supervised group, SVM, once again exhibited a good balance between precision and recall when compared to FAM and EAM. Perhaps, feature selection could have improved the performance of ARTMAP classifiers. Amongst the ARTMAP classifiers, EAM was able to derive better generalizations for positive depchange group when compared to FAM. The resulting efficacy is likely due to the use of hyper-ellipsoids, which can represent more complex boundaries when compared to FAMs hyper-rectangular categories.

In the multi-instance learning group, MI-SVM with the collective assumption performed a lot better in learning concepts from the positive depchange group (higher recall). Poor generalizations of MI-OB, MI-CKNN and MI-DD show that the standard MI assumption may not be suited for depression change prediction.

In the custom classifier category, the probabilistic classifier over-learned the unchanged depscore group and failed to derive generalizations for positive depchange group. In PC, considering the fact that bayesian learning is involved and only 24 positive samples were present, the result is perhaps not very surprising. Once again, *ConSVM* achieved the best balance between precision and recall. Additionally, it also exhibited the best overall accuracy. This shows that the content aware netflow distance metric is even more crucial for predicting the *change* in depression.

5.4. SUMMARY OF CONCLUSIONS

In this chapter, it was shown that a better classification accuracy can be achieved by considering baseline deviation patterns in netflow data. More particularly, the classification result improved from ∼74% to ∼84%. However, unlike the absolute classifier described in chapter 4, the baseline deviation classifier can only predict the likelihood of increase or decrease in depression. One possible way to utilize this classifier is to chain the results of the baseline deviation classifier with yet another classifier predicting the current depression state. i.e., given a sequence of predictions from the baseline derivation patterns, one might attempt to build a classifier for predicting the absolute depression state. Given the fact that depression tends to be periodic, the proposed approach is well worth trying.

From the statistical standpoint, several interesting observations were drawn. Overall, it was observed that an increase in depression tends to co-occur with:

1. A decrease in http packets and duration
2. A decrease in streaming usage duration
3. An increase in ftp packets and duration

It is interesting to note the inverse relationship between the change in depression with http usage and streaming activity. One may hypothesize that a decrease in streaming activity and http usage could mean an increase in social isolation, a condition commonly encountered amongst depressed people. In any case, more experimental evidence is needed before any conclusions can be drawn. Nevertheless, the study opens new avenues of explorations for future research.

6. CONCLUSION AND FUTURE WORK

The study began with the following premise:

> "*Since depression and Internet usage exhibit extensive correlations, Is it possible to use Internet activity as a predictor of depression?*".

As evidenced by the results provided in the study, not only can Internet data be used as a predictor of depression, it can also be used in a transparent and privacy preserving manner. By considering Cisco Netflow™, a proprietary but open protocol for collecting IP traffic information, as the characterization of Internet usage, the proposed classification models achieved an accuracy of ∼74%, demonstrating the feasibility of the approach. The confidence in the proposed framework is further reinforced by the considering deviation patterns in netflow data, resulting in another ∼10% improvement in classification accuracy.

Being the first work to use "real" Internet data, the study uncovered several new statistical results correlating Internet activity with depression. A concise summary of statistical results is available in section 3.4 and 5.4. As far as the statistical results are concerned, an expert interpretation of Internet traffic features and their relation to user activity is needed. At this point, the results raise more questions than the answers. Nevertheless, the use of netflow for studying depression opens several new avenues for future research.

In addition to depression classification, a new metric was proposed for computing the distance between any given netflow samples. In today's ever increasingly cyber-connected world, the author foresees several interesting applications with the proposed distance metric. For example, the content aware netflow distance metric provides an easy and transparent method of identifying users with *similar* online behavior. This information can be used to personalize advertisements, capture user preferences, and can even have applications in marketing.

In conclusion, the study demonstrated the feasibility of predicting depression using Cisco Netflow data. The strength of the proposed framework lies in its practicality, extensibility, transparency, and privacy preserving nature.

6.1. FUTURE WORK

The work presented in this thesis is barely the tip of an iceberg. This section describes a few approaches that can potentially be used to improve the classification accuracy. It is intended as a guide to future researchers who wish to extend and build on this thesis. There are several avenues for improvement, classified into three types of categories.

6.1.1. Experimental Improvements.

- **Larger Data Samples:** The first and most obvious improvement of all is to collect more amount of data. In this thesis, a total of 165 participants' data was utilized. One of the immediate tasks is to increase the number of samples and reassess the reliability of the framework.

- **Additional Control Factors:** Data can be collected from a variety of majors and universities across the country. With additional attributes such as ethnicity, major etc., more fine grained statistical insights can be gained.

- **Resolving the Internet-Depression Conjecture:** There has been a long lasting debate as to whether or not the use of Internet *causes* depression. By using structural equation modeling (SEM), it is possible to detect both the causality and its direction. As real Internet data will be involved, the results will hardly be debated.

6.1.2. Improving Accuracy.

- **Automated Feature Extraction:** As feature extraction is crucial for supervised learning models, it might be beneficial to investigate the possibility of automating the feature extraction process using Genetic programming (GP). Unlike the computational complexity of GP, once the feature extraction program trees are determined, they can be used efficiently in real-time operations.

- **Consider Tsallis Entropy:** In chapter 3, it was shown that netflow traffic features exhibit a non-normal distribution. This means that Tsallis entropy, a generalization of shannon entropy, might be better suited. Moreover, studies have shown that Tsallis entropy may be better suited to characterize Internet traffic [67, 80, 51]. One can also derive a larger number of entropy based features by considering ip-address and port-address as discrete events in addition to flows.

- **Consider Entropy Time-Series:** Instead of using absolute entropy features, a normalized entropy time-series, capturing the change in entropy with respect to time, might be more beneficial as baseline deviation method showed greater promise.

6.1.3. Multi-Instance ARTMAP. In the probabilistic classifier described in section 4.3.2, data was clustered before using it for classification. This approach is very similar to how ARTMAPs operate. Therefore, it might be possible to adapt existing ARTMAP architectures to work with multi-instance data format. If successful, the MI-ARTMAP can be used to deploy a *robust* depression classifier.

6.2. NOTE TO THE READERS

The author encourages interested researchers to extend and evaluate the proposed framework. All the necessary source code and scripts can be obtained from the author upon request.

APPENDIX A

QUESTIONNAIRES AND CONSENT FORMS

Consent Form: Attitudes and Experiences of College Students Spring 2011

30 minutes of Experiment Credit for Psych 50

Thank you for agreeing to participate in this study, "Attitudes and Experiences of College Students." Your participation is completely voluntary and you may discontinue your participation at any time. For this study you will fill out several questionnaires assessing various attitudes and experiences of undergraduate college students. You must be 18 or older to participate. If you are not 18, you should leave the room.

All information related to this study will be confidential and the results will only be reported in terms of group data. If you have any questions regarding this study, please feel free to contact Dr. Frances (Dee) Haemmerlie Montgomery at 573 341-4810. For additional information regarding human participation in research, feel free to contact the Missouri S&T Institutional Review Board Office (573 341-4305).

Please indicate that you have read this statement of informed consent, that you are at least 18 years of age, and that you have willingly agreed to be a participant by signing your name in the space provided below.

I have read the above, understand my rights as a participant in research, and wish to participate in this study. I have been given the opportunity to ask questions about this study and receive a copy of this consent form for my records.

________________________________ January 19, 2011
Print Participant's Name

____
Participant's Signature

General Background Information
Spring 2011

Sex: M ___ F ____ **Age:** ________ **Major:**__________________________

Year in school at S&T: ____ 1st semester freshman ____ 2nd or > semester freshman
____ sophomore
____ junior
____ senior
____ other

GPA on a 4.0 scale: High School _____ Missouri S&T: _____ or ____not have one yet

Racial Background:
White/Caucasian ____ African American____ South Asian____ East Asian _____
Native/Aboriginal______ Hispanic: _______ Other: (specify): __________________________

Life Experiences:
Briefly describe what is the most positive and inspiring school or career related experience that you have had in your life to date:

Briefly describe what is the most negative or discouraging school or career related experience that you have had in your life to date:

Centers for Disease Control and Prevention (CDC)
Health-Related Quality-of-Life Questionnaire

Please answer the following 14 questions developed by the CDC asking about your general health status:

Part A: Healthy Days

1. Would you say in general that your health is:
   ___ Excellent
   ___ Very Good
   ___ Good
   ___ Fair
   ___ Poor

2. Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good?
   ___ Number of days
   ___ None

3. Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good?
   ___ Number of days
   ___ None

4. During the past 30 days, for about how many days did poor physical or mental health keep you from doing your usual activities such as self-care, schoolwork, work, or recreation?
   ___ Number of days
   ___ None

Part B: Activity Limitations

1. Are you LIMITED in any way in any activities because of any impairment or health problem?
   ___ Yes
   ___ No (if no, go to Part C)

2. What is the MAJOR impairment or health problem that limits your activities (note: select only one category):
   ___ Arthritis/rheumatism
   ___ Back or neck problem
   ___ Fractures, bone/joint injury
   ___ Walking problem
   ___ Lung/breathing problem
   ___ Hearing problem
   ___ Eye/vision problem
   ___ Heart problem
   ___ Stroke problem
   ___ Hypertension/high blood pressure
   ___ Diabetes
   ___ Cancer
   ___ Depression/anxiety/emotional problem
   ___ Other impairment problem = ______________________________

3. For HOW LONG have your activities been limited because of your major impairment or health problem? ___ Days ___ Weeks ___Months ___ Years

CDC-2

4. Because of any impairment or health problem, do you need the help of other persons with your PERSONAL CARE needs such as eating, bathing, dressing, or getting around the house?
   ___ Yes ___ No

5. Because of any impairment or health problem, do you need the help of other persons in handling your ROUTINE needs such as everyday household chores, doing necessary schoolwork or business, shopping, or getting around for other purposes?
   ___ Yes ___ No

Part C: Healthy Days Symptoms

1. During the past 30 days, for about how many days did PAIN make it hard for you to do your usual activities such as self-care, schoolwork, work, or recreation?
   ___ Number of Days
   ___ None

2. During the past 30 days, for about how many days have you felt SAD, BLUE, or DEPRESSED?
   ___ Number of Days
   ___ None

3. During the past 30 days, for about how many days have you felt WORRIED, TENSE, or anxious?
   ___ Number of Days
   ___ None

4. During the past 30 days, for about how many days have you felt you did NOT get ENOUGH REST or SLEEP?
   ___ Number of Days
   ___ None

5. During the past 30 days, for about how many days have you felt VERY HEALTHY AND FULL OF ENERGY?
   ___ Number of Days
   ___ None

Recent Affective Experiences Questionnaire

The 20 items below refer to how you have felt and behaved during the **last week.** For each item, read the statement and then answer how often you have felt or behaved that way by checking the blank that best describes the frequency of your feelings or behaviors using the following rules:

Rare or none of the time (<1 day)
Some or a little of the time (1-2 days)
Occasionally or a moderate amount of the time (3-4 days)
Most or all of the time (5-7 days)

1. I was bothered by things that don't usually bother me.
___ Rare ___ Some ___ Occasionally ___Most of the Time

2. I did not feel like eating; my appetite was poor.
___ Rare ___ Some ___ Occasionally ___Most of the Time

3. I felt that I could not shake off the blues even with the help of my family or friends.
___ Rare ___ Some ___ Occasionally ___Most of the Time

4. I felt that I was just as good as other people.
___ Rare ___ Some ___ Occasionally ___Most of the Time

5. I had trouble keeping my mind on what I was doing.
___ Rare ___ Some ___ Occasionally ___Most of the Time

6. I felt depressed.
___ Rare ___ Some ___ Occasionally ___Most of the Time

7. I felt everything I did was an effort.
___ Rare ___ Some ___ Occasionally ___Most of the Time

8. I felt hopeful about the future.
___ Rare ___ Some ___ Occasionally ___Most of the Time

9. I thought my life had been a failure.
___ Rare ___ Some ___ Occasionally ___Most of the Time

10. I felt fearful.
___ Rare ___ Some ___ Occasionally ___Most of the Time

11. My sleep was restless.
___ Rare ___ Some ___ Occasionally ___Most of the Time

12. I was happy.
___ Rare ___ Some ___ Occasionally ___Most of the Time

13. I talked less than usual.
___ Rare ___ Some ___ Occasionally ___Most of the Time

14. I felt lonely.
___ Rare ___ Some ___ Occasionally ___Most of the Time

15. People were unfriendly.

___ Rare ___ Some ___ Occasionally ___Most of the Time

16. I enjoyed life.

___ Rare ___ Some ___ Occasionally ___Most of the Time

17. I had crying spells.

___ Rare ___ Some ___ Occasionally ___Most of the Time

18. I felt sad.

___ Rare ___ Some ___ Occasionally ___Most of the Time

19. I felt that people disliked me.

___ Rare ___ Some ___ Occasionally ___Most of the Time

20. I could not get "going".

___ Rare ___ Some ___ Occasionally ___Most of the Time

Schedule of Life Events

Please think carefully about your life as you answer the questions below. For each question, read the question and then answer what your life has been like based on your race and/or gender (i.e., being female or male). Circle the number that best describes these events using the following rules:

Circle 1 = If the event has NEVER happened to you
Circle 2 = If the event happened ONCE IN A WHILE (less than 10% of the time)
Circle 3 = If the event happened SOMETIMES (10-25% of the time)
Circle 4 = If the event happened A LOT (26-49% of the time)
Circle 5 = If the event happened MOST OF THE TIME (50-70% of the time)
Circle 6 = If the event happened ALMOST ALL OF THE TIME (more than 70% of the time)

1. How many times have you been treated unfairly by teachers or professors because of your race and/or gender?

Race:	1	2	3	4	5	6
Gender:	1	2	3	4	5	6

2. How many times have you been treated unfairly by your employer, boss or supervisors because of your race and/or gender?

Race:	1	2	3	4	5	6
Gender:	1	2	3	4	5	6

3. How many times have you been treated unfairly by your co-workers, fellow students or colleagues because of your race and/or gender?

Race:	1	2	3	4	5	6
Gender:	1	2	3	4	5	6

4. How many times have you been treated unfairly by people in service jobs (by store clerks, waiters, bartenders, waitresses, bank tellers, mechanics and others) because of your race and/or gender?

Race:	1	2	3	4	5	6
Gender:	1	2	3	4	5	6

5. How many times have you been treated unfairly by strangers because of your race and/ or gender?

Race:	1	2	3	4	5	6
Gender:	1	2	3	4	5	6

6. How many times have you been treated unfairly by people in helping jobs (by school counselors, school principals, doctors, nurses, psychiatrists, case workers, dentists, therapists, pediatricians and others) because of your race and/or gender?

Race:	1	2	3	4	5	6
Gender:	1	2	3	4	5	6

7. How many times have you been treated unfairly by neighbors because of your race and/or gender?

Race:	1	2	3	4	5	6
Gender:	1	2	3	4	5	6

8. How many times have you been treated unfairly by your boy/girl-friend, spouse, or other significant person in your life because of your race and/or gender?

Race:	1	2	3	4	5	6
Gender:	1	2	3	4	5	6

9. How many times have you been denied a raise, a promotion, a good assignment, a job, or other such thing at work because of your race and/or gender?

Race:	1	2	3	4	5	6
Gender:	1	2	3	4	5	6

SLE-2

10. How many times have you been treated unfairly by your family because of your race and/or gender?

Race:	1	2	3	4	5	6
Gender:	1	2	3	4	5	6

11. How many times have people made inappropriate or unwanted advances of a sexual nature to you because of your race and/or gender?

Race:	1	2	3	4	5	6
Gender:	1	2	3	4	5	6

12. How many times have people failed to show you the respect that you deserve because of your race and/or gender?

Race:	1	2	3	4	5	6
Gender:	1	2	3	4	5	6

13. How many times you wanted to tell someone off for being biased or prejudiced against you because of your race and/or gender?

Race:	1	2	3	4	5	6
Gender:	1	2	3	4	5	6

14. How many times have you been really angry about something that was done to you because of your race and/or gender?

Race:	1	2	3	4	5	6
Gender:	1	2	3	4	5	6

15. How many times were you forced to take drastic steps (moving to a new school, filing a complaint or grievance, quitting your job, filing a lawsuit and other actions) to deal with the way you have been treated because of your race and/or gender?

Race:	1	2	3	4	5	6
Gender:	1	2	3	4	5	6

16. How many times have you been called a negative name negative because of your race and/or gender (e.g., jerk, bastard, bitch, cunt or other name)?

Race:	1	2	3	4	5	6
Gender:	1	2	3	4	5	6

17. How many times have you gotten into an argument or a fight about something negative that was said or done to you or done to somebody else because of their race and/or gender?

Race:	1	2	3	4	5	6
Gender:	1	2	3	4	5	6

18. How many times have you been made fun of, picked on, pushed, shoved, hit or threatened with harm because of your race and/or gender?

Race:	1	2	3	4	5	6
Gender:	1	2	3	4	5	6

19. How many times have you heard people making racial and/or gender biased or degrading jokes?

Race:	1	2	3	4	5	6
Gender:	1	2	3	4	5	6

APPENDIX B

IRB APPROVAL

Campus Institutional Review Board Approval Form
Missouri University of Science and Technology

This is to certify that the research proposal entitled:

A Computationally Intelligent Framework for Identifying Depression in College Students via Mining Internet Usage Patterns

Submitted by: **Sriram Chellappan**
Department: **Computer Science**

has been reviewed by the Campus IRB and approved with respect to the study of human subjects as appropriately protecting the rights and welfare of the individuals involved.

Type of Approval: _______Exempt **X** Expedited ________Full

Approval Date: 8/26/10 Expiration Date: 10/1/14

Note that approval of this research is contingent upon the following agreement by the researcher(s):

1) To report potentially serious events to the Campus IRB by the most expeditious means within five days of occurrence. The IRB may require an additional written report.
2) To submit a **Change in IRB Approval Form UMRIRB-2***, if the project changes in any way that affects human subjects.
3) To maintain copies of all pertinent information, including copies of informed consent agreements, for a period of three years from the date of completion of the research.
4) To adhere to all UMR Policies and Procedures relating to human subjects, as written in accordance with 45 Code of Federal Regulations 46.
5) To be aware that Federal and University Regulations require continuing review of research projects involving human subjects. Therefore, **this approval will expire one year from date of approval.** To meet this requirement, **Continuing Review Report UMRIRB-4* should be filed within one year of the original approval date.** However, projects receiving Exempt Approval and lasting less than one year do not need to provide this report. The campus IRB reserves the right, at any point, to inspect project records to ensure compliance with federal regulations.

*See http://www.umr.edu/~irb/forms.html for the necessary forms.

Approved By: Richard Hall Title: Professor, Information Science and Technology

Date: 8/26/10

APPENDIX C

APPLICATION-WISE CATEGORIZATION

A total of 61 application categories were identified by filtering flows based on set combinations of port, protocol and sometimes, the destination ip-address as allocated by IANA [6]. Due to complex conditional dependencies involved in application-wise classification, a brief pseudocode is provided.

Given the fact that the amount of data involved is fairly large, the speed of processing was an important consideration. To meet this demand, the author developed a full-fledged C code to process all the raw data in a single pass. The program takes netflow logs as input and outputs a file in which each line is of form “$\langle label \rangle : \langle flows \rangle \langle packets \rangle \langle octets \rangle$”. The label describes the kind of traffic (Ex - `http`, `nntp` etc.) summarized. As the complete code is rather large, the author refrained from including it in the appendix. Nevertheless, the source code may be obtained by contacting the author.

The following pseudocode describes the application types, their port ranges and protocols.

Pseudocode for Application-wise Aggregation

```
if(PORT_IS(1214)&&IS_TCP)
APP_IS(FASTTRACK_APP);
else if(PORT_IN_RANGE(6881,6889)&&IS_TCP)
APP_IS(BITTORRENT_APP);
else if(PORT_IN_RANGE(6346,6350)&&IS_TCP)
APP_IS(GNUTELLA_APP);
else if(PORT_IN_RANGE(20,21)&&IS_TCP)
APP_IS(FTP_APP);
else if((PORT_IN_RANGE(80,81)||PORT_IS(8080))&&IS_TCP)
APP_IS(HTTP_APP);
else if((PORT_IS(119)||PORT_IS(563))&&IS_TCP)
APP_IS(NNTP_APP);
else if(PORT_IS(443)&&IS_TCP)
APP_IS(HTTPS_APP);
else if(PORT_IS(2049)||PORT_IS(1110))
APP_IS(NFS_APP);
else if((PORT_IS(25)||PORT_IN_RANGE(109,110)||PORT_IS(143)
||PORT_IS(220)||PORT_IS(465)||PORT_IS(585)
||PORT_IS(587)||PORT_IS(993))&&IS_TCP)
APP_IS(MAIL_APP);
else if(PORT_IS(22)&&IS_TCP)
APP_IS(SSH_APP);
else if(PORT_IS(23)&&IS_TCP)
APP_IS(TELNET_APP);
```

Pseudocode for Application-wise Aggregation (Continued)

```
else if(PORT_IS(123)&&IS_UDP)
APP_IS(NTP_APP);
else if(PORT_IS(53))
APP_IS(DNS_APP);
else if(PORT_IN_RANGE(4661,4665))
APP_IS(EDONKEY_APP);
else if((PORT_IS(6690)||PORT_IS(19114))&&IS_TCP)
APP_IS(FREENET_APP);
else if(PORT_IS(500)&&IS_UDP)
APP_IS(IPSEC_IKE_APP);
else if(prot==50)
APP_IS(IPSEC_ESP_APP);
else if(prot==51)
APP_IS(IPSEC_AH_APP);
else if(dst_ip>>24==232)
APP_IS(SSM_APP);
else if(dst_ip>>28==14)
APP_IS(MULTICAST_APP);
else if(PORT_IS(2811)&&IS_TCP)
APP_IS(GSIFTP_APP);
else if(PORT_IN_RANGE(5020,5022)&&IS_TCP)
APP_IS(BBFTP_APP);
else if(PORT_IN_RANGE(5031,5033)&&IS_TCP)
APP_IS(BBCP_APP);
else if(PORT_IN_RANGE(5000,5009))
APP_IS(IPERF_APP);
else if((PORT_IS(194)||PORT_IN_RANGE(6666,6670))&&IS_TCP)
APP_IS(IRC_APP);
else if(PORT_IS(135)||PORT_IN_RANGE(137,139)||PORT_IS(445)
||PORT_IN_RANGE(568,569)||PORT_IS(1512))
APP_IS(MSWINDOWS_APP);
else if(PORT_IN_RANGE(6970,6973)||PORT_IN_RANGE(7070,7071)
||PORT_IS(554))
APP_IS(REALPLAYER_APP);
else if(PORT_IN_RANGE(7000,7006))
APP_IS(AFS_APP);
else if((PORT_IN_RANGE(6000,6005)||PORT_IS(7100)
||PORT_IS(1024)||PORT_IS(6016))&&IS_TCP)
APP_IS(X11_APP);
else if(PORT_IS(1755)||PORT_IS(7007)||PORT_IS(135))
APP_IS(WINMEDIA_APP);
else if((PORT_IS(1720)||PORT_IS(1503))&&IS_TCP)
APP_IS(H323_APP);
else if(PORT_IS(1558))
APP_IS(STREAMWORKS_APP);
else if(PORT_IS(41170)&&IS_UDP)
APP_IS(BLUBSTER_APP);
else if(PORT_IS(6699)||PORT_IS(6257))
APP_IS(WINMX_APP);
```

Pseudocode for Application-wise Aggregation (Continued)

```
else if(PORT_IN_RANGE(8000,8005))
APP_IS(SHOUTCAST_APP);
else if(PORT_IN_RANGE(5500,5503))
APP_IS(HOTLINE_APP);
else if(PORT_IS(161))
APP_IS(SNMP_APP);
else if(PORT_IS(113)&&IS_TCP)
APP_IS(IDENT_APP);
else if(PORT_IS(3128))
APP_IS(SQUID_APP);
else if(PORT_IS(1080)&&IS_TCP)
APP_IS(SOCKS_APP);
else if(PORT_IS(4000)||PORT_IN_RANGE(6112,6119))
APP_IS(BATTLENET_APP);
else if(PORT_IS(26000)||PORT_IN_RANGE(27910,27961))
APP_IS(QUAKE_APP);
else if(PORT_IN_RANGE(28000,28008))
APP_IS(STARSIEGE_APP);
else if(PORT_IN_RANGE(6700,6702))
APP_IS(CARRACHO_APP);
else if(PORT_IS(0)&&(IS_TCP||IS_UDP))
APP_IS(PORTZERO_APP);
else if(PORT_IS(27005)||PORT_IS(27015))
APP_IS(HALFLIFE_APP);
else if(PORT_IS(5190))
APP_IS(AIM_APP);
else if(PORT_IS(873)&&IS_TCP)
APP_IS(RSYNC_APP);
else if(PORT_IS(388))
APP_IS(UNIDATA_LDM_APP);
else if(PORT_IN_RANGE(412,413))
APP_IS(NEOMODUS_APP);
else if((PORT_IS(2048)&&IS_TCP)||(PORT_IS(2050)&&IS_UDP))
APP_IS(BACKBONE_RADIO_APP);
else if((PORT_IN_RANGE(2047,2048)||PORT_IS(1972))&&IS_UDP)
APP_IS(CAMARADES_APP);
else if(PORT_IS(27900)||PORT_IS(28900)
||PORT_IN_RANGE(29900,29901)
||((PORT_IS(13193)||PORT_IS(6515))&&IS_UDP))
APP_IS(GAMESPYARCADE_APP);
else if(PORT_IS(771))
APP_IS(RTIP_APP);
else if(PORT_IS(111))
APP_IS(PORTMAPPER_APP);
else if(PORT_IN_RANGE(49606,49609)&&IS_UDP)
APP_IS(VOIP_APP);
else if(PORT_IS(6714)&&IS_TCP)
APP_IS(IBP_APP);
```

Pseudocode for Application-wise Aggregation (Continued)

```
else if(PORT_IS(112))
APP_IS(MCIDAS_APP);
else if(PORT_IS(9000)&&IS_UDP)
APP_IS(ASHERON_APP);
else if((PORT_IS(47624)&&IS_TCP)||(PORT_IS(6073)&&IS_UDP)
||PORT_IN_RANGE(2300,2400))
APP_IS(DIRECTX_APP);
else if(PORT_IN_RANGE(41000,42000)&&IS_TCP)
APP_IS(AUDIOGALAXY_APP);
else APP_IS(UNIDENTIFIED)
```

BIBLIOGRAPHY

[1] Caligare s.r.o. NetFlow Portal [online]. 2006. Available from: `http://netflow.caligare.com/index.htm` [cited 2011-07-07] .

[2] Cisco System, Netflow Services Solutions Guide [online]. January 2007. Available from: `http://www.cisco.com/univercd/cc/td/doc/cisintwk/intsolns/netflsol/nfwhite.pdf` [cited 2011-07-07] .

[3] Cisco System, October 2007, Introduction to Cisco IOS Netfow - A Technical Overview.

[4] College kids are Digital Demo [online]. Available from: `http://www.emarketer.com/Article.aspx?R=1007329` [cited 2011-02-30].

[5] College Students embrace the web [online]. Available from: `http://www.imediaconnection.com/content/8237.asp` [cited 2011-04-03].

[6] IANA, Internet Assigned Numbers Authority [online]. Available from: `http://www.iana.org/assignments/port-numbers` [cited 2010-05-02].

[7] Study to explain sleep and depression in teen [online]. Available from: `http://www.universityofcalifornia.edu/news/article/22290` [cited 2011-04-23].

[8] Teenage depression [online] available from `http://www.helpguide.org/mental/depression_teen_teenagers.htm` 2011-02-03].

[9] The NSDUH report [online] Available from: `http://www.oas.samhsa.gov/2k9/149/MDEamongAdults.pdf` [cited 2010-04-03].

[10] Web pages in 1992 [online]. Available from: `http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html` [cited 2011-08-16].

[11] Web pages in 1992 [online]. Available from: `http://royal.pingdom.com/2008/04/04/how-we-got-from-1-to-162-million-websites-on-the-internet/` [cited 2011-08-16].

[12] C. Abad, Y. Li, K. Lakkaraju, X. Yin, and W. Yurcik. Correlation between netflow system and network views for intrusion detection. In *Workshop on Information Assurance (WIA04). April*, pages 14–17. Citeseer, 2004.

[13] S. Ali and M. Shah. Human action recognition in videos using kinematic features and multiple instance learning. *IEEE transactions on pattern analysis and machine intelligence*, pages 288–303, 2008.

[14] H. Alt and M. Godau. Measuring the resemblance of polygonal curves. In *Proceedings of the eighth annual symposium on Computational geometry*, pages 102–109. ACM, 1992.

[15] G. Anagnostopoulos and M. Georgiopoulos. Ellipsoid art and artmap for incremental unsupervised and supervised learning. *Proceedings of the IEEE–INNS–ENNS*, 2:1221–1226, 2001.

[16] K. Anderson. Internet use among college students: An exploratory study. *Journal of American College Health*, 50(1):21–26, 2001.

[17] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. *Advances in neural information processing systems*, pages 577–584, 2003.

[18] G. Attardi, A. Gullì, and F. Sebastiani. Automatic web page categorization by link and context analysis. In *Proceedings of THAI*, volume 99, pages 105–119. Citeseer, 1999.

[19] G. Attardi, A. Gullì, and F. Sebastiani. Theseus: categorization by context. In *Proceedings of the 8th International World Wide Web Conference*. Citeseer, 1999.

[20] P. Auer and R. Ortner. A boosting approach to multiple instance learning. *Machine Learning: ECML 2004*, pages 63–74, 2004.

[21] K. Bessière. Effects of Internet Use on Health and Depression: A Longitudinal Study. *J Med Internet Res*, 12(1):e6, 2010.

[22] C. Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998.

[23] A. Campbell, S. Cumming, and I. Hughes. Internet use by the socially fearful: Addiction or therapy? *CyberPsychology & Behavior*, 9(1):69–81, 2006.

[24] G. Carpenter. Default artmap. *CAS/CNS Technical Report Series*, (008), 2010.

[25] G. Carpenter and S. Grossberg. A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer vision, graphics, and image processing*, 37(1):54–115, 1987.

[26] G. Carpenter, S. Grossberg, N. Markuzon, J. Reynolds, and D. Rosen. Fuzzy artmap: A neural network architecture for incremental supervised learning of analog multidimensional maps. *IEEE Transactions on Neural Networks*, 3(5):698–713, 1992.

[27] G. A. Carpenter, S. Grossberg, and J. H. Reynolds. Artmap: Supervised real-time learning and classification of nonstationary data by a self-organizing neural network. *Neural Netw.*, 4:565–588, September 1991.

[28] K. Chen and L. Hong. User identification based on game-play activity patterns. In *Proceedings of the 6th ACM SIGCOMM workshop on Network and system support for games*, pages 7–12. ACM, 2007.

[29] X. Chen, C. Zhang, and W. Chen. A multiple instance learning framework for incident retrieval in transportation surveillance video databases. In *IEEE 23rd International Conference on Data Engineering Workshop*, pages 75–84. IEEE, 2007.

[30] H. Chung and M. Klein. Improving identification and treatment of depression in college health. *SPECTRUM*, 2007.

[31] M. C.M. and G. H. The Relationship between Excessive Internet Use and Depression: A Questionnaire-Based Study of 1319 Young People and Adults. *Psychopathology*, 2010.

[32] N. Cristianini and J. Shawe-Taylor. *An introduction to support Vector Machines: and other kernel-based learning methods.* Cambridge Univ Pr, 2000.

[33] W. de Vries, G. Moura, and A. Pras. Fighting spam on the sender side: A lightweight approach. *Networked Services and Applications-Engineering, Control and Management*, pages 188–197, 2010.

[34] T. Dietterich, R. Lathrop, and T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1-2):31–71, 1997.

[35] K. Douglas, J. Collins, C. Warren, L. Kann, R. Gold, S. Clayton, J. Ross, and L. Kolbe. Results from the 1995 national college health risk behavior survey. *Journal of American College Health*, 46(2):55–67, 1997.

[36] L. Ertoz, E. Eilertson, A. Lazarevic, P. Tan, V. Kumar, J. Srivastava, and P. Dokas. Minds-minnesota intrusion detection system. *Next Generation Data Mining*, 2004.

[37] J. Gangwisch, D. Malaspina, K. Posner, L. Babiss, S. Heymsfield, J. Turner, G. Zammit, and T. Pickering. Insomnia and Sleep Duration as Mediators of the Relationship Between Depression and Hypertension Incidence. *American journal of hypertension*, 23(1):62–69, 2009.

[38] M. Genton. Classes of kernels for machine learning: a statistics perspective. *The Journal of Machine Learning Research*, 2:299–312, 2002.

[39] M. Griffiths and R. Wood. Risk factors in adolescence: The case of gambling, videogame playing, and the Internet. *Journal of Gambling Studies*, 16(2):199–225, 2000.

[40] S. Grossberg. Adaptive pattern recognition and universal encoding ii: Feedback, expectation, olfaction, and illusions. *Biological Cybernetics*, 23:187–202, 1976.

[41] T. Grtner, P. Flach, A. Kowalczyk, and A. Smola. Multi-instance kernels. In *Proceedings of the 19th International Conference on Machine Learning*, pages 179–186. Citeseer, 2002.

[42] D. Hann, K. Winter, and P. Jacobsen. Measurement of depressive symptoms in cancer patients:: Evaluation of the center for epidemiological studies depression scale (ces-d). *Journal of Psychosomatic Research*, 46(5):437–443, 1999.

[43] S. Haykin. *Neural networks: a comprehensive foundation*. Prentice hall, 1999.

[44] N. Hensler-McGinnis. Cyberstalking Victimization: Impact and Coping Responses in a National University Sample. 2008.

[45] M. Hofmann. Support vector machineskernels and the kernel trick. 2006.

[46] M. Hospitals and C. Clinics. Pro-Anorexia/Pro-Bulimia Websites: A Dangerous Influence.

[47] V. Jacobson, C. Leres, S. McCanne, et al. TCPDUMP, 1989.

[48] M. Kan and H. Thi. Fast webpage classification using url features. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 325–326. ACM, 2005.

[49] M. Kearns and D. Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Computation*, 11(6):1427–1453, 1999.

[50] J. Kisch, E. Leino, and M. Silverman. Aspects of suicidal behavior, depression, and treatment in college students: Results from the spring 2000 national college health assessment survey. *Suicide and Life-Threatening Behavior*, 35(1):3–13, 2005.

[51] E. Kohler, J. Li, V. Paxson, and S. Shenker. Observed structure of addresses in ip traffic. In *Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurment*, IMW '02, pages 253–266, New York, NY, USA, 2002. ACM.

[52] R. Kubey, M. Lavin, and J. Barrows. Internet use and collegiate academic performance decrements: Early findings. *Journal of Communication*, 51(2):366–382, 2001.

[53] R. LaBrie, H. Shaffer, D. LaPlante, and H. Wechsler. Correlates of college student gambling in the United States. *Journal of American College Health*, 52(2):53–62, 2003.

[54] L. Lam and Z. Peng. Effect of Pathological Use of the Internet on Adolescent Mental Health. *Archives of Pediatrics & Adolescent Medicine*, 2010.

[55] O. Maron and T. Lozano-Pérez. A framework for multiple-instance learning. In *Advances in neural information processing systems*, pages 570–576. Citeseer, 1998.

[56] O. Maron and A. Ratan. Multiple-instance learning for natural scene classification. In *In The Fifteenth International Conference on Machine Learning*. Citeseer, 1998.

[57] N. Melnikov and J. Schönwälder. Cybermetrics: user identification through network flow analysis. In *Proceedings of the Mechanisms for autonomous management of networks and services, and 4th international conference on Autonomous infrastructure, management and security*, AIMS'10, pages 167–170, Berlin, Heidelberg, 2010. Springer-Verlag.

[58] M. Mitchell and J. Jolley. *Research design explained.* Wadsworth Pub Co, 2010.

[59] J. Morahan-Martin and P. Schumacher. Loneliness and social uses of the Internet. *Computers in Human Behavior*, 19(6):659–671, 2003.

[60] C. Morgan and S. Cotten. The relationship between internet activities and depressive symptoms in a sample of college freshmen. *Cyberpsychology & behavior: the impact of the Internet, multimedia and virtual reality on behavior and society*, 6(2):133, 2003.

[61] F. Mrchen. *Time series feature extraction for data mining using DWT and DFT.* Citeseer, 2003.

[62] J. Murre. *Learning and categorization in modular neural networks.* Lawrence Erlbaum, 1992.

[63] N. C. on Addiction, S. A. at Columbia University (CASA), and U. States. Depression, Substance Abuse and College Student Engagement: A Review of the Literature. 2003.

[64] T. Pao and P. Wang. Netflow based intrusion detection system. In *IEEE International Conference on Networking, Sensing and Control*, volume 2, pages 731–736. IEEE, 2004.

[65] I. H. E. C. F. A. O. D. . V. Prevention. Festering Beneath The Surface: Gambling And College Students.

[66] L. Radloff. The ces-d scale: A self report depression scale for research in the general. *Applied psychological measurement*, 1(3):385–401, 1977.

[67] A. Scherrer, N. Larrieu, P. Owezarski, P. Borgnat, and P. Abry. Non-gaussian and long memory statistical characterizations for internet traffic with anomalies. *IEEE Trans. Dependable Secur. Comput.*, 4:56–70, January 2007.

[68] B. Schlkopf, C. Burges, and A. Smola. *Advances in kernel methods: support vector learning*, volume 2. MIT press Cambridge, MA, 1999.

[69] M. Schneider, G. Dunton, and D. Cooper. Media Use and Obesity in Adolescent Females&ast. *Obesity*, 15(9):2328–2335, 2007.

[70] C. Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

[71] A. Sperotto, G. Vliek, R. Sadre, and A. Pras. Detecting spam at the network level. *The Internet of the Future*, pages 208–216, 2009.

[72] W. Stallings. *SNMP, SNMPv2, SNMPv3, and RMON 1 and 2*. Addison-Wesley, 1999.

[73] S. Stevens and T. Morris. College dating and social anxiety: using the Internet as a means of connecting to others. *CyberPsychology & Behavior*, 10(5):680–688, 2007.

[74] L. Sweeney et al. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty Fuzziness and Knowledge Based Systems*, 10(5):557–570, 2002.

[75] B. Van Voorhees, J. Fogel, T. Houston, L. Cooper, N. Wang, and D. Ford. Beliefs and attitudes associated with the intention to not accept the diagnosis of depression among young adults. *The Annals of Family Medicine*, 3(1):38, 2005.

[76] K. Veropoulos, C. Campbell, and N. Cristianini. Controlling the sensitivity of support vector machines. In *Proceedings of the international joint conference on AI*. Citeseer, 1999.

[77] S. Villani. Impact of media on children and adolescents: a 10-year review of the research. *Journal of Amer Academy of Child & Adolescent Psychiatry*, 40(4):392, 2001.

[78] J. Wang and J. Zucker. Solving the multiple-instance problem: A lazy learning approach. In *MACHINE LEARNING-INTERNATIONAL WORKSHOP*, pages 1119–1126. Citeseer, 2000.

[79] N. Weidmann, E. Frank, and B. Pfahringer. A two-level learning method for generalized multi-instance problems. *Machine Learning: ECML 2003*, pages 468–479, 2003.

[80] W. Willinger, V. Paxson, and M. S. Taqqu. *Self-similarity and heavy tails: structural modeling of network traffic*, pages 27–53. Birkhauser Boston Inc., Cambridge, MA, USA, 1998.

[81] I. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

[82] Y. Wu, D. Agrawal, and A. El Abbadi. A comparison of dft and dwt based similarity search in time-series databases. In *Proceedings of the ninth international conference on Information and knowledge management*, pages 488–495. ACM, 2000.

[83] J. Yang, R. Yan, and A. Hauptmann. Multiple instance learning for labeling faces in broadcasting news video. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 31–40. ACM, 2005.

[84] M. Ybarra. Linkages between depressive symptomatology and Internet harassment among young regular Internet users. *CyberPsychology & Behavior*, 7(2):247–257, 2004.

[85] C. Yen, J. Yen, and C. Ko. Internet addiction: ongoing research in Asia. 2010.

[86] K. Young and R. Rogers. The relationship between depression and Internet addiction. *CyberPsychology & Behavior*, 1(1):25–28, 1998.

[87] L. Zadeh. Fuzzy sets. *Information and control*, 8(3):338–353, 1965.

[88] Q. Zhang and S. Goldman. Em-dd: An improved multiple-instance learning technique. *Advances in neural information processing systems*, 2:1073–1080, 2002.

[89] W. Zhenqi and W. Xinyu. Netflow based intrusion detection system. In *2008 International Conference on MultiMedia and Information Technology*, pages 825–828. IEEE, 2008.

[90] Z. Zhou. Multi-instance learning: A survey. *AI Lab, Department of Computer Science and Technology, Nanjing University, Tech. Rep*, 2004.

[91] Z. Zhou and M. Zhang. Multi-instance multi-label learning with application to scene classification. *Advances in Neural Information Processing Systems*, 19:1609, 2007.

VITA

Raghavendra Kotikalapudi was born in Secunderabad, India, on February 18, 1988. He received his Bachelor's degree in Computer Science from Shri Mata Vaishno Devi University, India, in August 2009, and specialized in the field of computational intelligence (Neural Networks in particular). He subsequently joined Missouri University of Science and Technology (formerly University of Missouri – Rolla) in Fall 2009 and received his Master's degree in Computer Science in December 2011. During the course of his study, he worked with the Department of Defence over a period of nine months, where he helped with the development of a commercial *Virtual Landmine Detection Training Simulator*. He also worked in the capacity of Graduate Research Assistant at Intelligent Systems Center (ISC) and Applied Computational Intelligence Laboratory (ACIL). His areas of interest include machine learning, data mining, evolutionary computation, virtual reality and software engineering.