AnonyMine: Mining anonymous social media posts using psycho-lingual and crowd-sourced dictionaries

Arindam Paul
Electrical Engineering and Computer Science
Northwestern University
Evanston, IL 60208
apaul@u.northwestern.edu

Ankit Agrawal
Electrical Engineering and Computer Science
Northwestern University
Evanston, IL 60208
ankitag@northwestern.edu

Wei-keng Liao
Electrical Engineering and Computer Science
Northwestern University
Evanston, IL 60208
wkliao@eecs.northwestern.edu

Alok Choudhary
Electrical Engineering and Computer Science
Northwestern University
Evanston, IL 60208
choudhar@northwestern.edu

ABSTRACT

There is lot of research activity in the area of opinion mining and sentiment analysis, which deals with the computational treatment of opinion, sentiment, and subjectivity in text. Social media websites have become increasingly popular for discussing uncomfortable topics. However, there are limited resources for mining and automatically labeling posts discussing self-disclosure. There is great incentive for a system which can be useful for monitoring emotional state of users, both for the research community as well as for mental health and business purposes.

This paper presents a case where we leverage information from psycho-lingual and crowd-sourced dictionaries to create a system which can automatically predict anonymous posts about taboo topics on a social media site (Facebook Confessions). We achieve more than 80% accuracy for the most popular taboo topics, and an overall accuracy of 61.25 % across all taboo categories. We evaluate our system in two ways: a) comparing against human-annotated posts on another anonymous social media platform YikYak b) an evaluation against existing state-of-the-art models.

CCS Concepts

•**Information retrieval** → Sentiment analysis; •**Computing methodologies** → Natural language processing; Machine learning;

Keywords

Emotion Mining; Sentiment Analysis; Social Media Mining

DOI: 10.1145/1235

1. INTRODUCTION

Sentiment analysis has become a prominent research area in the fields of machine learning and Natural Language Processing(NLP) [20]. Automatically understanding the theme and intent of social media posts is an important aspect for companies' customer services. For social media companies, it is critical for capturing and understanding emotions and positive and negative sentiments of their users. At the same time, the internet is changing the lexicon of users and it is lexically highly active and rapidly changing [7]. There has been a lot of work on text summarization [30], opinion mining [20, 19] etc.

Social media websites have become increasingly popular for discussing uncomfortable topics. Our past work had revealed students engaging in asking queries about taboo and stigma topics in a partially anonymous environment of Facebook Confession Boards(FCBs) [5] with negligible negative responses. Taboo topics are topics which people are not comfortable discussing in an identified or face-to-face scenario, as they expect negative outcomes in that case [27]. Common taboo topics for a Western audience include family matters/details, hygiene, prejudice, sexuality, finances, and feelings of attraction between friends [15]. It can be seeking information 'Does anyone know if you can get checked for STDs at [NAME OF HEALTH CENTER]? and is it expensive?'or an observation or remark 'I wish gay girls at LGBT parties were more approachable.' FCBs are Facebook pages targeted at offline communities like universities, high schools, workplaces etc e.g. [1]. They allow anonymous posting via an external web form (e.g., SurveyMonkey) where anybody may anonymously submit content that is then reposted to the FCB by the moderator. They represent a unique combination of locality, anonymity and identifiability.

Mining self-disclosure posts is fraught with subjectivity. However, anonymity and the short length of the posts make the task of understanding the context very challenging. The same expression can indicate different sentiment in different domains.Further, we are interested in identifying posts which discuss taboos or uncomfortable topics and categorizing them, and not just in summarizing or understanding the

general topic of the post. We do not have personal or other social information about the posters except the name of the university in which they post. Knowing the characteristics and context of the language used, including slangs, is essential for NLP and sentiment analysis [26]. Researchers are mostly dependent on human labeling for such tasks which is an expensive and time-consuming endeavor. The motivation of our work was to create a machine learning based model which can maximize the learning from a highly contextual anonymized data-set. In our work, we elucidate a methodology of using a psycho-social [12] and crowd-sourced lexicon based approach to understand and automatically classify taboo topics from an anonymous self-disclosure forum. This approach would allow researchers to train their models on a small dataset given that the lexicon they use are contextual rather than just based on a simple Term frequency inverse document frequency(tf-idf) based model.

The classifier for the posts are trained by extracting feature vectors using tf-idf, and the two lexicons we used for this study (LIWC and Urban Dictionary). Linguistic Inquiry Word Count(LIWC) [22], [29] is a well recognized psycho linguist lexicon based tool which counts words (unigrams) in psychologically meaningful categories. The motivation to use Urban Dictionary came from the abundance of popular cultural and slang terms in the posts. We use around 4000 posts with a 75% and 25% split between the training and unseen holdout validation set. The primary contribution of our paper is proving how a crowd-sourced and psycho-lingual lexicon system can enrich our understanding of highly subjective social media posts(anonymous self-disclosure or otherwise). The rest of the paper proceeds as follows. In Section 2, we provide a literature review of related work and description about the dictionaries used in this study. Section 3 presents the dataset and related statistics used in our study. The infrastructure and methodology is described in Section 4. Section 5 gives an detailed discussion on the performance of our system in an incremental manner i.e. as we improved our system and presents the evaluation of our system. Section 6 discusses limitations of our work, and the future work. Last but not least, in Section 7, we recapitulate our contribution in the conclusion.

2. BACKGROUND AND RELATED WORK

2.1 Anonymity and Self-Disclosure

Anonymity have been seen to have positive impact towards self-disclosure[27],[28]. SIDE [25] is a model in social psychology which describes how members in a group can form a group identity (and conform to norms) and deindividuation can lead to giving more voice to their collective identity. Postmes T. et al. [23] found that anonymity in a group can promote normative behavior, and normative processes can shape behavior in anonymous groups despite the less direct contact of group members with each other. We have found people seeking online communities to seek support about health [32] and Choudhury's work hinted that dissociative anonymity creates an atmosphere of disinhibition in sharing about mental health concerns [8, 10, 9] and smoking and drinking abstinence [28]on reddit. Andalibi et al. [3] studies social media disclosures of sexual abuse. People happen to connect with people in similar circumstances [13].

2.2 Taboos

Taboo topics are defined as topics which one is hesitant to discuss with their respective friends [24]. We elaborated a detailed labeling scheme for taboo topics based on social science literature [4, 15] and based on a similar schema in our previous work [5]. There were nine categories of taboos originating in our data: 1) death, 2)bodily functions, 3)sex, 4) illegal substances (e.g. drugs and other controlled substances), 5) protected social categories (e.g, gender, race, sexual orientation), 6) finances, 7) physiological health, 8) mental health and 9)academic performance.

2.3 Internet Socio-Linguistics and Crowd-sourced Lexicography

Internet Linguistics is a relatively new field of research but already has shown signs of changing mainstream discourse. Language change occurs under the influence of a combination of social processes, socio-cultural factors and also the geographical area. In the field of Internet linguistics, the corpus, and to what extent it can be constructed, is determined by what is available on the eponymous Internet. Lexicography is seen as complementary [14] to socio-linguistics in reaction to its pervasive neglect of socially conditioned variation in language and because of its focus on the dictionary. However, internet lexicons have created a marriage between these two different linguistic branches. For e.g., Heston et. al [17] in their investigation of local anonymous app YikYak in university campuses found usage of location specific language, in particular usage of implicit language to draw on shared knowledge of the location.

2.3.1 Psycho-Social Dictionary: LIWC

LIWC [21] analyzes text files on a word-by-word basis using an internal dictionary of more than 2,300 of the most common words and word stems. LIWC classifies the words into dozens of linguistic and psychological categories that tap social, cognitive, and affective processes. The 2007 English LIWC dictionary contains 4,500 words [21]. Each word has been classified or rated by experts on 64 word categories: 22 standard linguistic categories (e.g., pronouns, verb, tenses), 32 psychological categories (e.g., affect, cognition, social, biological processes), 7 personal categories (e.g., work, home, leisure), and 3 paralinguistic dimensions (assents, fillers, nonfluencies). Each word in a text is matched to a word in the dictionary and the associated word characteristics are extracted. In our work, we use vectors based on the association of each of unigrams(words) and bigrams in various LIWC categories similar to tf-idf vectors.

2.3.2 Crowd-sourced dictionary: Urban Dictionary

UrbanDictionary [11] is the largest source for slang and Internet terms with over six million crowd-sourced definitions of predominantly slang terms. In comparison, Oxford English Dictionary has just over 250,000 entries [18]. Urban Dictionary was designed as a satirical crowd-sourced dictionary. Anyone is able to submit a definition for any word on Urban Dictionary, which can also be a description rather than a strict definition. It has outgrown its initial intention into a dictionary containing but not limited to popular cultural references and slangs. Its lexicon has also broadened to include words or phrases of any usage, rather than just slang. Quality control is imposed through up and down voting by users to float up popular and accepted definitions and

reject and bury those that are not.

3. DATASET

3.1 Geographic Information

To gather data from a range of FCBs [5], we searched using keywords "confession" or "confessional" and college names for US News & World Report's 2013 top-ranked 100 universities and 100 liberal arts colleges in the United States. Of the 200 universities and colleges, 90 had FCBs. We eliminated those that were empty or inactive, or had slight variations in their configuration (e.g., anonymous comments). This left 52 active FCBs from 50 colleges (some had multiple FCBs) in 22 US states (plus Washington, DC), ranging in size from 1000 to 45,000 students (per their web sites). FCBs ranged in post volume from around 10 to 20,000 posts. There was no correlation between post volume and college size.

3.2 Metadata

For each post, we collected the content, date, and the numbers of likes and comments. As the posts were anonymous, we could not gather any other demographic data. This process, completed in April 2014, yielding 90,329 posts of which we randomly choose 3000 for training and 1000 for an unseen validation set. Compared to non-coded posts, there were no significant differences in post length or comment volume. The first 3000 were taken from 35 universities which was used in our previous study [5] and the other 1000 were annotated later. The actual number of posts used for train and test dropped to 2803 and 928 respectively because we removed the ones which had just urls in them.

3.3 Taboos

It is important to note that the general topic of the post is different from the taboo topic mentioned in the post. It is possible that the general topic of a post is different from the taboo topic. This is because, the taboo topic is based on a different premise. As mentioned in Section 2, a taboo topic is a motif where one is not comfortable in discussing about it.

For e.g. the example for taboo topic Illegal substances, "Am I an evil, vicious person because I am so weak that drugs have become more vital than water to me". A semantic topic analysis would give "personality" and "negative emotions" as topics and sentiments respectively. However, the taboo in question is drugs (illegal substances).

In the table 1, we present a description of each taboo category with their relative percentage among taboo posts and an example.

We also describe examples which we did not consider taboos in that category although a simple semantic topic analysis of the posts might tag them with that taboo. This is because either it was not mentioned in an uncomfortable

1. **Sexual**: I've made it a goal to hook up with (almost) every girl from a certain sorority - **Hooking up not necessarily synonymous with sex**

2. **Death**: Reiterating a point I read on this page. I tried to kill myself last year, for reasons that boiled down to the fact that while I was sitting alone in my room I could not figure out why life was worth living. You don't fight thoughts like these with more thoughts; you fight them with living. You'd be amazed how much it helps just to be around other people, it's the main thing that has turned my life around - **suicidal thoughts so marked as mental health and not death**

3. **Academics**: I can't stand this school, and I'm sick of trying to. - **mention of 'this school' is not related to performance, i.e. grades**

4. METHOD

4.1 Taboo categories

Suppose we have set of $\mathcal{C}$ Facebook confession posts, and the set of $\mathcal{T}$ taboo posts is a proper subset of $\mathcal{C}$ i.e. $\mathcal{T} \subset \mathcal{C}$. Again, there are 9 categories of taboos $\mathcal{S}$, $\mathcal{R}$, $\mathcal{I}$, $\mathcal{F}$, $\mathcal{P}$,$\mathcal{M}$, $\mathcal{D}$, $\mathcal{B}$ and $\mathcal{A}$. They represent sex, protected categories (race, religion etc), illegal substances, finances, physiological health, mental health, death/dying, bodily function (excretion etc) and academics respectively. Each of these categories does not have any posts common to each other i.e. $\mathcal{X} \cap \mathcal{Y} = 0$ for all combinations of the taboos where $\mathcal{X}$,$\mathcal{Y}$ represent different taboo categories.

4.2 Cleaning and Normalization

We normalize each post before we classify it. This involves fragmenting the sentences into individual words, removal of all non-alphanumeric characters, conversion to lowercase of all characters, removal of stop words and porter stemming. So, a post like "Feeling lazy to get up from bed to go to school" is normalized to feel, lazi, bed, school. We filter out yaks containing urls(e.g. beginning with http, www) as they tend to contain spam or contain generic information.

4.3 Oversampling

The taboo posts formed a small percentage of the total number of posts. And taboo posts are divided into nine further categories. Hence, the number of posts in each category is small compared to non-taboo posts. Oversampling is a common procedure in spam detection algorithm, when the number of posts are very small compared to the data<cite>. We compensate the imbalance in the category representation w.r.t. taboos in the given training set by applying Synthetic Minority Oversampling Technique (SMOTE) [6]. In this technique, k nearest neighbors of a training sample belonging to the minority class is generated. Therefore, the minority class(or classes) is over-sampled exploiting the artificial training samples. We tried different degrees of oversampling i.e. how many times an imbalanced minority sample is over-sampled. We found an optimal balance around 2 and 3 for most categories. It is also worth mentioning that we also applied random sampling at first, however, SMOTE gives much better results. We believe this is because we had a small training data-set of around 3000 posts, and an even smaller number of posts with taboo.

4.4 Bag of words Vectorization and Classification

Table 1: Description of different taboo categories,with their coding symbol, relative percentage and an example

Code	Taboo category	Description	Percentage	Example
$\mathcal{S}$	Sex	Discussion of sex or sexual desires	30.3	I'm a terrible <religion>. I can't stop thinking about sex.. And having it with every cute guy I see!
$\mathcal{R}$	Protected Categories	Primary focus includes gender/sexual orientation/religion/ethnicity/disability discussions	26.3	As a <race> man from a fairly diverse high school, I had expected <school> to be relatively devoid of prejudice.
$\mathcal{I}$	Illegal substances	Mention of drug (includes underage drinking) use, dependency, inappropriate use, abuse, or otherwise non-normative drug use	8.4	Am I an evil, vicious person because I am so weak that drugs have become more vital than water to me
$\mathcal{F}$	Finances	Discussion of explicit mentions of income, socioeconomic status that would be considered not allowed or otherwise improper in polite discussion.	6.4	I may have to drop out of <school> as my parents cannot afford the tuition
$\mathcal{P}$	Physiological health	Discussion of topics relating to diagnosable physical diseases (including mention of symptoms), illnesses, health statuses	4.4	Are there any other diabetics whose meter I can use. My insurance is not letting me...
$\mathcal{M}$	Mental health	Discussion of mental illness/eating disorders	5.2	Is there anyone who was depressed but somehow got out of it ?....
$\mathcal{D}$	Death	Discussion of death or dying, e.g. coping with death, fear of death	1.9	A girl from my hometown committed suicide three days ago...She hung herself..
$\mathcal{B}$	Bodily Functions	Mention of bodily excretions, bodily processes, private parts when the focus or context of the post is not explicitly sexual in nature	11.8	Anyone remembers how boring pooping was before smart-phones
$\mathcal{A}$	Academics	Discussion of poor performance at school, poor grades, worries about scholastic success, and achievement.	5.3	I am on the verge of failing 2 classes...

4.4.1 tf-idf

For the problem of text categorization of a document, the usual $tf - idf$ based representation is a feature-vector document representation taking one post as a set of term sequences, including term t and term weight w. Then the document will be made up of the pairs of <t, w>. Term Ti and Weight Wi would represent the features which express the post content, and value relevant to the coordinate. So every document (d) is mapped to the target space as a feature vector.

In the case of the term frequency tf(t,d), the simplest choice is to use the raw frequency of a term in a document, i.e. the number of times that term t occurs in document d, which measures how frequently a term occurs in a document. The inverse document frequency is a measure of how much information the word provides, that is, whether the term is common or rare across all documents. It is the logarithmically scaled inverse fraction of the documents that contain the word, obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient.

$$tf - idf(t, d, D) = tf(t, d) * idf(t, D)$$

4.4.2 Lexicons

We use two lexicons (mentioned in Section 2: LIWC and Urban Dictionary to enrich the vector space. We use the simpler frequency based model for these two lexicons as these dictionaries were chosen based on the context of the posts. For LIWC, we used the licensed online LIWC database [21].

For Urban Dictionary, we designed a python based scraper which extracted related words for top 20 words found in the tf-idf vectorized classification model for each Taboo category. The motivation for adding urban dictionary was two-fold. First, the corpus for the confessions used slangs and popular culture references. Hence, we believed that we should build a model based on such a lexicon. Urban Dictionary [11] provides a huge lexicon of words derived from popular culture unlike Dictionary.com, Merriam-Webster.com etc. We enriched the initial corpus of words from tf-idf by adding words from the related words section. We present a snapshot of the urban dictionary website for a search term in Fig 1. Second, we had a fairly small labeled dataset available for training (2083 posts) and the understanding of the classifiers would be augmented by the Urban Dictionary Corpus. A small labeled dataset is often the case as human annotation is costly especially when the labeling task is highly

Table 2: Part of LIWC Dictionary

LIWC Category	example words in that category
Health	abortion*, ache*, aching, acne, addict*, advil, aids, alcohol*, alive, allerg*, amput*, anorexi*, antacid*, antidepressant*, appendic*, arthr*, aspirin*, asthma*, bandage*, bandaid, binge*, binging, bipolar
Ingest	alcohol*, anorexi*, appeti*, ate, bake*, baking, bar, bars, beer*, binge*, binging, boil*, booz*, bread, breakfast*, brunch*, bulimi*, cafeteria*, candie*, candy, chew*, chow*, cigar*, cocktail*, coffee*, coke*
Positive Emotion	freed*, freeing, freely, freeness, freer, frees*, friend*, fun, funn*, genero*, gentle, gentler, gentlest, gently, giggl*, giver*, giving, glad*
Inhibition	blocking, blocks, bound*, brake*, bridle*, careful*, caut*, cease*, ceasing, compulsiv*, confin*, conflict*, conserv*, constrain*, constrict*, contain*, contradic*, control*, curb*, curtail*, defenc*, defens*, delay*, denia*, denie*, deny*, disciplin*, discourag*, disregard*, duti*, duty, enclos*, fenc*
Religious	catholic*, chapel*, chaplain*, christ, christian*, christmas*, church*, clergy, confess*, convent, convents, crucifi*, crusade*, demon*, devil*, divin*, doom*, episcopal*, evangel*, faith*, fundamentalis*, gentile*, god'*, gospel*, heaven*, hell, hellish, hells, hindu*, holie*, holy, hymn*, immoral*

contextual. The human annotators need to understand basics of taboo literature for them to annotate as was the case in our past work [5].

TF uses the score based on the frequency of a word in a post. On the other hand, TFIDF amortizes the score of the words found in the post based on its frequency in the entire corpus of posts. For the urban dictionary and LIWC models, we did not amortize their score but rather used the simple count or frequency.

We refer to these frequencies as lf(t,d) and uf(t,d) for LIWC and UrbanDictionary respectively.

The final vectorizer system uses an incremented lexicon model:

$$tfidf+(t,d,D) = tf(t,d).idf(t,D) + lf(t,d)$$

$$tfidf++(t,d,D) = tfidf+(t,d,D) + uf(t,d)$$

Hence,
$tfidf++(t,d,D) = tf(t,d).idf(t,D) + lf(t,d) + uf(t,d)$

The addition sign refer to the incremental nature of the model rather than the exact sequence or mathematical addition of feature vectors.

We present example words for selected LIWC categories in Table 2, and example words for the Taboo categories from the corpus created from Urban Dictionary in Table 3.

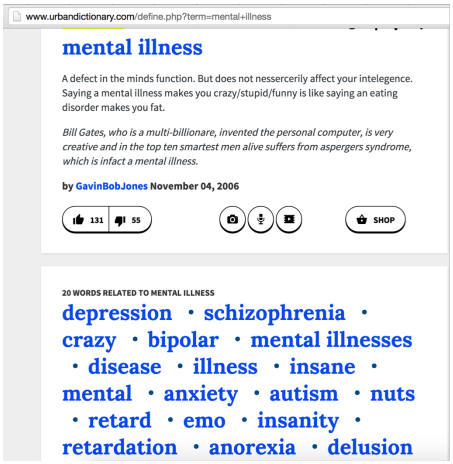

Figure 1: Example from Urban Dictionary (courtesy: urbandictionary)

4.4.3 *Classification*

Once we have converted the posts into a vectorized matrix, we train them over state of the art classification techniques (Naive Bayes, SVM etc).

In the following figure 2, the broad steps of our methodology is mentioned.

5. DISCUSSION AND EVALUATION

5.1 Results

For the final model, the accuracy on the unseen holdout set was 74.1 % when we used unigrams, and 72.4 % when we used both unigrams and bigrams. After we vectorized our corpus using tf-idf, LIWC and Urban Dictionary, we tried several multi-class classifiers to train and predict (mentioned in Section 4).

The most popular and effective models for language based models are Naive Bayes and Support Vector Machines. We used grid search for the above learning algorithms across kernels, loss functions, etc. till we achieved the best performance on the training set. We investigated different stop word, n-gram range across these kernels in our grid search. We would like to report that the highest performance of 74.1 % was achieved by a regularized Linear SVM classifier with stochastic gradient descent learning.

We had reached similar performance (74.4%) using a vectorized model consisting of LIWC and Urban Dictionary lexicons which does not use tf-idf. However, in spite of a slightly higher (0.3%) overall accuracy,our final model gave the best results for the False Positive and False Negative, as well as optimal accuracy over all taboo categories.

One of the most challenging part of our work was reducing false negatives. As our model has seen fewer posts containing taboo (less than one-third of the 2803 posts), initially

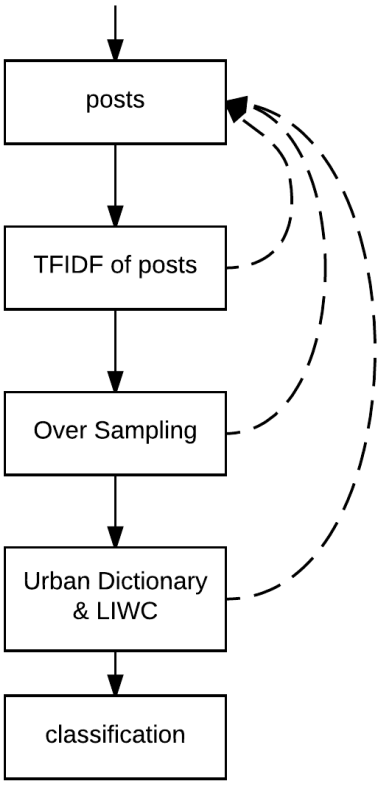

Figure 2: Flow-diagram of the lexical model

it was very defensive in marking a post as a taboo. We overcame part of the problem by using SMOTE, and we performed best (least False Negatives) using a vectorized model containing tf-idf, LIWC and Urban Dictionary.

We present the table comparing the average accuracy of different models in Table 4. Figure 3 compares the accuracy of the models across different taboos at 4 stages (before oversampling, after oversampling, after oversampling using LIWC and Urban Dictionary without tf-idf, our final ensemble method. Figure 4 illustrate the false positive and negative rates across different taboo categories.

5.2 Domain Translation

We evaluated our tool AnonyMine on posts from another anonymous social media application called YikYak. We wanted to understand and illustrate that our technique is not limited to working on Facebook Confession posts, and can generalize to other self-disclosure contexts.

Yik Yak is an anonymous mobile based social media app which combines GPS with instant messaging allowing users to anonymously post to other nearby users within a radius of 5 miles, which makes it well suited for college campuses [2]. A message can have a maximum size of 200 characters. Within the limits of this zone, anyone can post and vote or comment on other people's posts or "yaks". A post gets removed from Yik Yak if it gets down-voted to a score of 5 or it is replaced when it gets "old"(i.e. when it is the oldest among the 200 posts on the Yik Yak board) by a newer post.

Table 3: Example words from urban dictionary for each taboo category(certain words removed for decency)

Taboo category	Urban dictionary words
Sex	penis, vagina,intercourse,gay,love,hot,slut, sexism, sexist, sex
Protected categories	mexican, chinese, bisexual, homosexual, asian, female, queer, girl, heterosexual, funny, white, racism, lesbian, indian, homo, racist, women, homophobia, feminism, woman, gay, slur, homophobes
Illegal substances	stoner, vodka, alcoholic, pot, marijuana, sex, high, chronic, stoned, drinks, alcohol, whiskey, drunk, bowl, hammered, hangover, ganja, cannabis, beer, party, blunt, liquor, drugs, herb, dope, drink, weed, joint, wasted, smoking, drinking, boozing, bud, bar, tobacco, cigarette, smokes, pipe, cigarettes, dank, smoke, toke, grass, bong, wine
Finance	financial, drunk, money, dollar, sex, currency, paper, dough, bitch, homeless, welfare, hood, stocks, wealth, investing, unemployment,accounting, food, black, bank, rich, tax, bread, finances
Physiological health	medication,drugs,medicate,dentist, medicated, doctors, surgery, dr, surgeon, pot, medic, physician, doctor, marijuana, weed, medical, disease, drug, pills, emt, condition, sick, ambulance, medicine, paramedic, nurse, hospital, awesome
Mental health	love, death, single, desperate, suicidal, horny, depression, funny, suicide, sad, pathetic, lame, angry, lone, bad, unhappy, crying, bored, happy, goth, lonely, gay, fat, depressing, cutting, mad, alone, loner, anxiety, depressed, emo, cry, upset, loser, recession, ugly, sadness, emotional
Death	crazy, losses, weight, drunk, win, losing, necro, dead, sad, goth, kill, hell, fail, suicidal, crime, depression, homicide, cutting, dying, memory, life, murder, murderer, murderous, shit, necrophilia, killer, fuck, zombie, failure, suicide, alive, baby, killing, pain, killed, fratricide, lost, maniac, depressed, die, emo, violence, sex, gun, loser, deaded, genocide, boring, defeat
Bodily Functions	shit, crap, poop, dump, toilet, turd, defecate, piss, faeces, fart, feces, sex, urine, wank, eat, expel, fecal+matter, shart, urinate,beautiful, funny, booty, fat, hair, ugly, amazing, butt, girl, face, bodies
Academics	academics, school, college, university, smart, nerd, student, academia, education, study, intellectual, professor, homework, sports, teacher, work, high school, intelligent, research

We had 1000 yaks coded by the author on the similar taboo scheme as used in the Facebook Confessions. The yaks were collected by using a python-based open source github code [16] for a particular location. For consistency and to

Table 4: Evaluation of average accuracy between different models (upto 2 significant digits)

Model	NB%	SGD%
tf-idf (no oversampling with unigram)	0.57	0.56
tf-idf (no oversampling with unigram + bigram)	0.62	0.61
tf-idf (oversampling with unigram)	0.57	0.59
tf-idf (oversampling with unigram + bigram)	0.60	0.62
tf-idf + LIWC (oversampling with unigram)	0.64	0.68
tf-idf + LIWC (oversampling with unigram + bigram)	0.63	0.67
LIWC + UD (oversampling with unigram)	0.71	0.74
LIWC + UD (oversampling with unigram + bigram)	0.68	0.70
tf-idf + LIWC + UD(oversampling with unigram)	0.69	0.74
tf-idf + LIWC + UD(oversampling with unigram + bigram)	0.69	0.72

avoid lexical differences due to location, we used the same set of universities used for the Facebook Confessions. For this evaluation experiment, we used 10-fold cross-validation. We had an overall accuracy of 70.36% and for the major taboo categories (i.e. ones most prevalent in the corpus: sex and protected categories), we had an accuracy of 85.3%.

6. LIMITATIONS AND FUTURE WORK

As with any study, the work has limitations that urge interpretation with caution. One key limitation is that we had a small US university-based dataset comprising less than 4000 posts. Although we believe the data to be representative of an average US student population, we encourage researchers to compare such posts with users in other countries with different cultural norms regarding anonymity. Second, we looked at only one anonymous application Facebook Confession Boards. Although, for evaluation, we used AnonyMine to predict the categories for a mobile-based anonymous social application (YikYak). Last, we restricted ourselves only to the text in the posts and did not investigate the comments, emoticons or urls mentioned in the posts.

As a future work, we are working on releasing AnonyMine as a full-fledged web-based application where a user can enter a social media post and get a response about the taboos and other emotions described in the post. The authors believe recent developments in recurrent neural network based models can be used for learning from a larger anonymous text corpus. We also suggest researchers to explore other combinations of anonymity and social media as we believe we can exploit advances in machine learning to learn more about human emotions.

7. CONCLUSION

In this paper, we describe our methodology of learning from anonymous social posts assisted by a combination of psychological text analysis tool LIWC, and a crowd-sourced pop culture based dictionary Urban Dictionary. With this

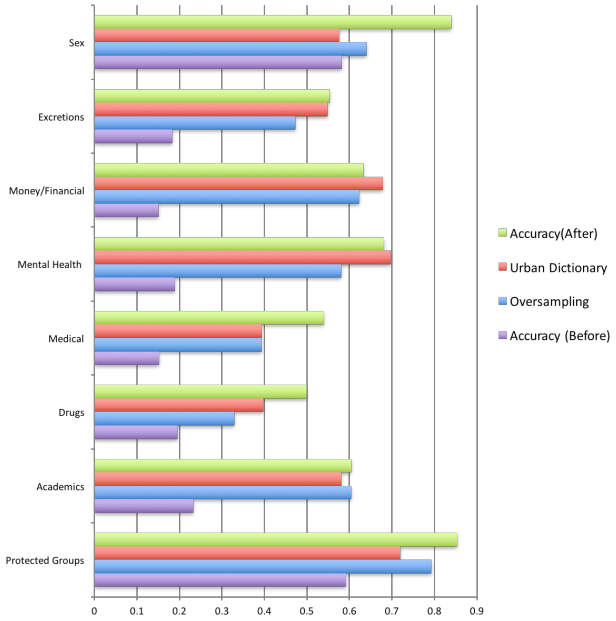

Figure 3: Incremental accuracy across the taboo labels for different models

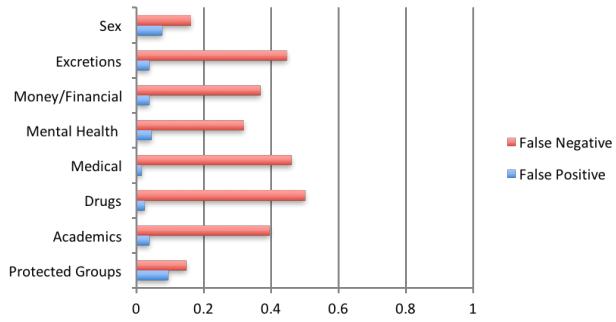

Figure 4: False positive and negative rates of different labels

ensemble methodology, we are able to identify themes (in this case taboo) which may be distinct from the general topic of the text. We believe the learning methodology we outlined, would generalize and help in providing a data-driven understanding to classifying self-disclosure texts. Our approach can be scaled to assist university services in comprehensive monitoring of mental health of students and also extended to other psychological facilities.

8. ACKNOWLEDGMENTS

This work is supported in part by the following grants: NSF awards CCF-1029166, IIS-1343639, CCF-1409601; DOE awards DE-SC0007456, DE-SC0014330; AFOSR award FA9550-12-1-0458; NIST award 70NANB14H012; DARPA award N66001-15-C-4036. The first author expresses his gratitude to the Northwestern Segal Design Institute for their year long design fellowship for the accepted proposal "Taking the college pulse" which funded the author for a design-based interdisciplinary work. The first author would also like to

thank Jeremy Birnholtz from the School of Communication and Doug Downey from the Department of Electrical Engineering and Computer Science at Northwestern University for initial brain-storming. He extends heartfelt thanks to the members of the Segal Design Committee for their valuable suggestions.

9. REFERENCES

[1] Mit confessions. https://www.facebook.com/beaverconfessions.

[2] Yik yak - find your herd. https://www.yikyak.com.

[3] N. Andalibi, O. L. Haimson, M. De Choudhury, and A. Forte. Understanding social media disclosures of sexual abuse through the lenses of support seeking and anonymity. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 3906–3918. ACM, 2016.

[4] L. A. Baxter and W. W. Wilmot. Taboo topics in close relationships. *Journal of Social and Personal Relationships*, 2(3):253–269, 1985.

[5] J. Birnholtz, N. A. R. Merola, and A. Paul. Is it weird to still be a virgin: Anonymous, locally targeted questions on facebook confession boards. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 2613–2622. ACM, 2015.

[6] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, pages 321–357, 2002.

[7] D. Crystal. *Internet linguistics: A student guide*. Routledge, 2011.

[8] M. De Choudhury. Role of social media in tackling challenges in mental health. In *Proceedings of the 2nd international workshop on Socially-aware multimedia*, pages 49–52. ACM, 2013.

[9] M. De Choudhury and S. De. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *ICWSM*. Citeseer, 2014.

[10] M. De Choudhury, M. R. Morris, and R. W. White. Seeking and sharing health information online: Comparing search engines and social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1365–1376. ACM, 2014.

[11] U. Dictionary. Urban dictionary, llc. *San Francisco, available at www. urbandictionary. com/define. php*, 2013.

[12] E. H. Erikson. Major stages in psychosocial development. *The life cycle completed: A review*, pages 55–82, 1982.

[13] G. Eysenbach, J. Powell, M. Englesakis, C. Rizo, and A. Stern. Health related virtual communities and electronic support groups: systematic review of the effects of online peer to peer interactions. *Bmj*, 328(7449):1166, 2004.

[14] N. Fairclough. *Language and power*. Pearson Education, 2001.

[15] R. Goodwin and I. Lee. Taboo topics among chinese and english friends a cross-cultural comparison. *Journal of Cross-Cultural Psychology*, 25(3):325–338, 1994.

[16] B. Groom. Pyak. https://github.com/bradengroom/pyak, 2015.

[17] M. Heston and J. Birnholtz. (in) visible cities: an exploration of social identity, anonymity and location-based filtering on yik yak. *IConference 2016 Proceedings*, 2016.

[18] N. McLeese. How selfie got into the dictionary: an examination of internet linguistics and language change online. 2015.

[19] A. Mogadala and V. Varma. Retrieval approach to extract opinions about people from resource scarce language news articles. In *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining*, page 4. ACM, 2012.

[20] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.

[21] J. W. Pennebaker, R. J. Booth, and M. E. Francis. Liwc2007: Linguistic inquiry and word count. *Austin, Texas: liwc. net*, 2007.

[22] J. W. Pennebaker, M. E. Francis, and R. J. Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71:2001, 2001.

[23] T. Postmes, R. Spears, K. Sakhel, and D. De Groot. Social influence in computer-mediated communication: The effects of anonymity on group behavior. *Personality and Social Psychology Bulletin*, 27(10):1243–1254, 2001.

[24] W. K. Rawlins. Openness as problematic in ongoing friendships: Two conversational dilemmas 1. *Communications Monographs*, 50(1):1–13, 1983.

[25] S. D. Reicher, R. Spears, and T. Postmes. A social identity model of deindividuation phenomena. *European review of social psychology*, 6(1):161–198, 1995.

[26] E. Riloff, J. Wiebe, and W. Phillips. Exploiting subjectivity classification to improve information extraction. In *Proceedings of the National Conference On Artificial Intelligence*, volume 20, page 1106. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2005.

[27] J. Suler. The online disinhibition effect. *Cyberpsychology & behavior*, 7(3):321–326, 2004.

[28] A. Tamersoy, M. De Choudhury, and D. H. Chau. Characterizing smoking and drinking abstinence from social media. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, pages 139–148. ACM, 2015.

[29] Y. R. Tausczik and J. W. Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54, 2010.

[30] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, 10:178–185, 2010.