Journal of Health Informatics Research manuscript No.
(will be inserted by the editor)

Harnessing Psycho-lingual and Crowd-Sourced Dictionaries for Predicting Taboos in Written Emotional Disclosure in Anonymous Confession Boards

Arindam Paul
Wei-keng Liao
Alok Choudhary
Ankit Agrawal

Received: date / Accepted: date

Abstract There have been many efforts in the last decade in the health informatics community to develop systems that can automatically recognize and predict disclosures on social media. However, a majority of such efforts have focused on simple topic prediction or sentiment classification. However, taboo disclosures on social media that people are not comfortable to talk with their friends represent an abstract theme dependent on context and background. Recent research has demonstrated the efficacy of injecting concept into the learning model to improve prediction. We present a vectorization scheme that combines corpus and lexicon-based approaches for predicting taboo topics from anonymous social media datasets. The proposed vectorization scheme exploits two context-rich lexicons LIWC and Urban Dictionary. Our methodology achieves cross-validation accuracies of up to 78.1% for the supervised learning task on Facebook Confessions dataset, and 70.5% for the transfer learning task on the YikYak dataset. For both the tasks, supervised algorithms trained with features generated by the proposed vectorizer perform better than vanilla $tf - idf$ representation. This work presents a novel methodology for predicting taboos from anonymous emotional disclosures on confession boards.

Keywords E-mental health, text mining, human computer interaction, machine learning, social media informatics, anonymity

1 Introduction

Social media websites have become popular for discussing uncomfortable topics and support seeking [1]. However, identifiable communication systems suffer from inhibited behavior because of privacy and reputation concerns [2]. Although, anonymous forums provide a safe space for discussing mental health [3]

Arindam Paul, Northwestern University, Evanston, IL 60201, US
E-mail: arindam.paul@eecs.northwestern.edu

and uncomfortable issues [4], anonymity has been associated with disinhibition because of freedom from accountability and self-presentation concerns [5]. [6] suggests the importance of shared writing as a medium of emotional disclosure. Specifically, users have shown inhibition in discussing health concerns with their named identities on the internet [7,8]. Such spaces have been characterized as hotbeds of negativity like flaming [9] and cyberbullying [10]. Some student newspapers across different colleges have complained about the presence of micro-aggressions [11] on Yik Yak [12,13].

However, we find uncomfortable topics being discussed on anonymous forums. De Choudhury's work [14] reveals disinhibition in the discussion of mental health topics in Reddit, and anonymous users have taken part in more emotionally engaging communication than users with pseudonymous or named identities and urge effective private interventions for people vulnerable to different types of mental illnesses. Our past work had revealed students were engaging in asking queries about taboo and stigma topics in a partially anonymous environment of Facebook Confession Boards (FCBs) [15] with negligible negative responses. The majority of the posts sought information from a local community "Does anyone know if you can get checked for STDs at X Health Center? and is it expensive?" or offered an observation or remark about the community "I wish gay girls at LGBT parties were more approachable".

The proposed work aims to create a novel supervised machine learning based methodology that can learn and predict taboo topics from a highly contextual anonymous dataset harnessing context via context-rich lexicons. This work describes a methodology of combining a psycho-social [16] and crowdsourced lexicon-based approach with a corpus-based approach from anonymous self-disclosure forums. As the aim of this work is to present a data-driven methodology of ascertaining written emotional disclosure in students by predicting taboos in confessions, this methodology demonstrates a synthesis of a lexicon-based approach from crowdsourced and psycho-lingual dictionaries with a corpus-based approach for social text classification.

Multiple classification algorithms are evaluated on the proposed vectorization scheme, along with a comparison against the cross-validation accuracy results for other vectorization schemes. The system is evaluated in two ways: a) comparative analysis with machine learning algorithms on feature matrices from our proposed vectorization approach to other approaches on the FCB dataset, and b) transfer learning experiment on YikYak dataset, another anonymous social media platform. Our proposed methodology achieves cross-validation accuracies of up to 78.1% for the supervised learning task on the FCB dataset, and 70.5% for the transfer learning task on the YikYak dataset.

2 Background

The study of taboos in FCBs presents a unique combination of anonymity and locality in social media disclosures. In this section, background literature about studies about the impact of anonymity, locality, and taboos on social

media are presented. Furthermore, a background study on the two lexicons used in our system is discussed.

2.1 Anonymity and Self-Disclosure

Discussing mental health is a stigma topic [17–20], and the user might find a downvote and particularly, a removal to be a very negative response. We have seen repetitive negative feedback can actively discourage new users from staying in an online community (Everything2) [21]. In Everything2, we see some users do not participate actively but prefer being observers [22] but still form an essential part of the user-base. Wohn [23] and Lampe [24]'s work demonstrates that negative feedback discourages new users from returning to these respective online communities (Everything2 and Slashdot). Both of these forums allow users to have pseudonymous identities. The user reputations on these forums are public, i.e. other users are aware of this. However, Birnholtz [15] in his 2015 work found that a combination of anonymous and named identities led to a prosocial interaction. Furthermore, an emerging body of works has attempted to understand the nuances of context in different forms of text-based disclosures. DErrico et al. introduced the concept of acid communication [25] where they explored negative social emotions such as irritation, disappointment, guilt, envy, contempt, and awe. It was distinct from emotion analysis across five primary emotions anger, happiness, disgust, sadness, and fear, as they were not the most common emotions present in social communication. In their 2016 work, Ofek et al. [26] exploited concept information for developing an unsupervised knowledge enrichment system for sentiment analysis. Such works have demonstrated the success of techniques that configure affective computing systems by harnessing concept. Domain-specific lexicons perform better in comparison to domain-independent lexicons [27,28] for sentiment analysis. Feldman et al [29] determines which is the most appropriate set of questions to ask for health interventions.These works demonstrate how self-disclosure on online forums are connected to mental and emotional health. However, most of these approaches are either limited to qualitative studies or unsupervised text mining tasks or sentiment prediction.

Anonymity has been seen to have a positive impact on self-disclosure, and the SIDE [30] model in social psychology describes that members of a group form a group identity and conform to norms. Thus deindividuation in an anonymous environment can lead to a more collective identity. Postmes T. et al. [31] found that anonymity in a group can promote normative behavior, and normative processes can shape behavior in anonymous groups although members in the community do not know each other. Sassenberg and Postmes [32] found that strategic and cognitive processes interact to produce social influence within the group based on the perception of society and self within it, and those due to the positioning of self vis-a-vis a group. Researchers have studied the impact of anonymity for many decades. Wildman [33] investigated the influence of anonymity on survey responses. Choudhury et al.'s works [34,7,14,

35] hinted that dissociative anonymity creates an atmosphere of disinhibition in sharing about mental health concerns and smoking and drinking abstinence on Reddit. Andalibi et al. [36] investigated social media disclosures of sexual abuse in their 2016 paper. In their 2004 work, Eysenbach et al. demonstrated that people connect with others in similar circumstances [37].

2.2 Locality

Locality has an impact on both named and anonymous social media. In particular, the condition of anonymity in a geographically local setting can be violated if specific individuals are identified [38]. Personal information can be accidentally revealed on locally anonymous apps such as YikYak [39] or specific individuals can be identified that can result in cyber-bullying attacks [40, 10].

From studies of location-based dating applications, it is known that location can affect the type of content users are willing to share online [41,38]. Past studies about online interaction with nearby people have shown that people seek information about local topics [15], coordinate social encounters [42], or reach out for and provide help in crises [43].

In the recent past, resources for sharing information with, and asking questions to members of local communities are becoming popular. Some anonymous communities such as Cyclopath [44] and EveryBlock [45] allocate persistent pseudo-anonymous identities. Another application, YikYak, allows members of offline communities, such as colleges or other such campuses, to anonymously share with their colleagues or friends [15].

2.3 Taboos

Baxter et al. [46] defines taboo topics as those that are "off limits" to one party or another in a social relationship, anticipating a negative outcome from such a discussion. Goodwin et al. [47] formulated catalogs about potential taboo topics in different cultures. Their work indicated that taboo could vary contextually, and they found common taboo themes for a Western audience include family matters/details, hygiene, prejudice, and sexual topics. An elaborate labeling scheme for taboo topics based on social science literature [46,47] was developed as part of our previous work [15]. There were nine categories of taboos originating in the dataset: 1) death, 2) bodily functions, 3) sex, 4) illegal substances (e.g. drugs and other controlled substances), 5) protected social categories (such as gender, race, and sexual orientation), 6) finances, 7) physiological health, 8) mental health and 9) academic performance.

2.4 Lexicons

In this paper, we harness two dictionaries : LIWC and Urban Dictionary.

LIWC is a well-recognized psycholingual lexicon based tool that counts words (unigrams) in psychologically meaningful categories that analyze text files on a word-by-word basis using an internal dictionary of frequent words and word stems. During the 2008 U.S. elections, LIWC was used [48] to analyze and distinguish the usage frequency of different words/categories by political candidates. The current English LIWC dictionary contains more than 4,500 words. It classifies words into many linguistic and psychological categories that harness social, cognitive, and affective processes. 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 tallied with a word in the dictionary, and the associated term characteristics are extracted.

Urban Dictionary [49] (UD)is the largest source for slang and Internet terms with over six million crowd-sourced definitions. In comparison, Oxford English Dictionary has just over 250,000 entries [50]. Internet Linguistics [51, 52] is a relatively new field of research but already has shown signs of changing mainstream discourse. Urban Dictionary allows any user to submit a definition or description for a given word. It has outgrown its initial intent of a repository of slangs and modern cultural references into a full-grown dictionary. 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 those that are not.

Both dictionaries provide useful context but are distinct from each other. LIWC was developed by psycholinguists who studied how people tended to use different words based on their emotional state. In that context, it can be used as a vectorizer by creating numerical features from a body of text with each category serving as each dimension. As UD can provide a huge lexicon of words derived from popular culture unlike other dictionaries such as Dictionary.com [53] and Merriam-Webster.com [54], it can be used to find related colloquial words for most used words in each taboo category. This helps in synthetically creating a more richer corpus with a relatively smaller training data.

2.5 Text Mining Algorithms

We compare our proposed work with other popular and successful text mining approaches. Naive Bayes [55, 56] is a Bayesian classification algorithm and has demonstrated success for text classification using Bag of Words or $tf - idf$ representations. LinearSVM [57] is another algorithm that is popular for text categorization as it is relatively agnostic of the sparsity of the feature matrix. Random Forests [58] and Randomized Decision Trees [59] (also known as ExtraTrees) use an ensemble of decision trees to make a decision and are one

of the most successful traditional machine learning algorithms. LSA [60] is a technique in natural language processing for analyzing and comparing concepts across a set of documents. It is also used for dimensionality reduction to generate a dense matrix by Singular Value Decomposition on a sparse Bag of Words or $tf - idf$ representation. Embedding schemes such as GloVe [61] and Word2Vec [62] that consider co-occurrence of different words, have demonstrated state-of-the-art performance for most machine learning tasks in the recent past. GloVe is a pre-trained unsupervised learning algorithm based for obtaining co-occurrence vector representations for generating word embeddings from a corpus containing Wikipedia, Twitter and a collection of webpages. Word2Vec is another embedding scheme that utilizes a shallow two-layered neural network to construct a co-occurrence matrix from an unlabeled corpus. Word2Vec has two flavors: Continuous Bag of Words [62] and Skipgrams [63]. LSTMs are supervised recurrent neural networks that incorporate long-term word dependencies.

Wikarsa et al. [64] developed a system using naive bayes algorithm to predict six primary emotions: happiness, sadness, anger, disgust, fear, and surprise. Lupan et al. [65] developed an emotional state monitoring system using Latent Semantic Analysis called Emo2 to quantify emotions induced by news articles. Herzig et al. [66] used a word embedding approach on five datasets for emotion detection across different domains, and saw significant improvements over traditional methods. LSTMs are the current state-of-the-art for many emotional text mining problems. Schoene et al. [67] used a type of LSTM to classify suicide notes. Su et al. [68] used an LSTM network to predict across seven emotional classes: anger, boredom, disgust, anxiety, happiness, sadness, and surprise, and found large improvements over other predictive methods. Chancellor et al. [69] provides a detailed critical review of the predictive techniques for mental health status on social media.

Further, transfer learning is a form of machine learning that focus on storing knowledge gained while solving one problem and applying it to a different but related problem [70]. Although most machine learning systems are designed to address single tasks, transfer learning can accelerate learning across different but related problems. For instance, knowledge gained while learning to recognize automobiles could apply when trying to recognize trucks. Furthermore, as most real-world social media mining applications involve a data stream and not a static data source, the distribution of the data is not known a priori. Hence, evaluation of a proposed model using transfer learning on a similar but different dataset enhances confidence about the generalizability of the model.

3 Dataset

We describe the data collection, metadata information and annotation process for the two datasets in this section.

3.1 Data Collection

FCBs are facebook groups targeted at offline communities such as universities [71], high schools, and workplaces. FCBs allow posting via an external web form such as SurveyMonkey that anybody can anonymously submit content to and is later re-posted to the corresponding FCB by the moderator. However, commenters on the FCBs are identified by their Facebook profiles. For our study, we use FCBs from top universities and liberal arts colleges (based on US News & World Report [72, 73]). The student population of the schools for which FCBs were chosen ranged from 1000 to 45,000 students with the volume of posts varying between 100 to 20,000 posts. There was no correlation found between post volume and college size. Timeline Scraper API was developed that harnessed the Python-based Facebook Graph API [74] for downloading timeline information for the confession boards.

Table 1 Description of the FCB and YikYak Dataset. YikYak posts have a character limit of 200 characters.

Details	**FCB**	**YikYak**
API	Timeline Scraper [75]	Pyak [76]
No. of universities	50	50
No. of US States	22 and D.C.	22 and D.C.
No. of posts	90,329	100,000
No. of labeled posts	4000	1000
Average length (in no. of characters)	231.02	187.64

YikYak is an anonymous mobile-based social media app that combines GPS with instant messaging allowing users to post a YikYak message called "yak" anonymously. A yak can have a maximum size of 200 characters and visible to other nearby users within a variable radius of 1.5-5 miles (depending on user density), that makes it well suited for college campuses [39, 77]. Anyone can post, vote or comment on content within the limits of this zone, but users outside the radius have only view privilege. With the features of geo-locality, anonymity, and ephemerality of the posts, YikYak provides a reciprocal data source worthy of future investigation. An open source GitHub code [76] written in python was used for collecting yaks. For consistency and to avoid lexical differences due to location, the same set of universities were used for Facebook Confessions.

Although both FCBs and YikYak are confession forums, they are different in many ways. The visibility of FCBs are global. One can view FCBs in any part of the world. YikYak is only visible locally (the dataset is collected by synthetically updating location to be proximal to university campuses). This distinction leads to hyper-local nature of yaks compared to FCB confessions. FCBs are moderated. In fact, not only are posts dropped by moderators in some cases, but also campus moderators can inform the school authorities for posts with threats [78]. YikYak is not moderated or in a sense auto-moderated as posts automatically get downvoted. However, it is possible that controver-

sial posts can get popular on YikYak which might have been taken down by moderators for FCBs. Also, FCB posts are permanent unless the moderator pro-actively takes down old posts or the page is taken down. Yaks are ephemeral and vanish after a while. Most importantly, the limited length of the posts in yaks lead to more abbreviations and hashtags compared to facebook confessions where there is no character limit. These differences make it interesting to study both confession forums.

3.2 Metadata

The text, date, and number of likes and comments were extracted for each confession post. As the posts were anonymous, any other demographic data could not be collected. There was no difference between labeled and unlabeled posts in post length and comment volume. However, there was a small difference in the number of likes, but it was statistically not significant $p<0.05$. The comments were not annotated as the number of comments per posts was not very high.

Similarly, in the YikYak dataset ($p<0.5$), there was no statistically significant difference in the metadata information between the labeled and unlabeled yak data. 1000 yaks were randomly chosen for labeling ensuring all the universities were represented. Table 1 gives a description of metadata information of the datasets.

Individual or university identifiers were removed, and any examples with identifying details are avoided. as it is critical for researchers to consider user privacy and the possibility of inadvertent identification even when the dataset is public.

3.3 Annotation Process

The annotation process for labeling taboos was non-trivial and time-consuming as it required an in-depth understanding of taboo literature. It was hence important to focus on quality and do in-house training rather than use Amazon Mechanical Turk [79, 80]. The annotators were undergraduate students in social sciences. An annotation scheme used in our past work [15] was implemented which in turn was based on past literature on taboos. There are nine taboo categories - protected categories, death/dying, academics, illegal substances, physiological health, mental health, personal financial situation, bodily function, and sex with each post assigned to no more than one taboo category, denoted by class labels from 1 through 9. In case a post does not contain a taboo, it is labeled as 0. For the purpose of understanding the dataset, a group of 3 annotators labeled 700 posts, and an agreement of more than 80% across all the taboo categories was achieved. The goal of this phase was to attune the annotators with the labeling scheme and ensure consistency. The 700 posts used from this initial phase was discarded, and a new set of 4000 posts

Table 2 30% of all the posts were taboo-related. The following tables describes each taboo with an example. The class label precedes each taboo in the dataset.

Class Label	Taboo category	Description	Percentage %	Example
1	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.
2	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..
3	Academics	Discussion of poor performance at school, poor grades, worries about academic success, and achievement.	5.3	I am on the verge of failing 2 classes...
4	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
5	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...
6	Mental health	Discussion of mental illness/eating disorders	5.2	Is there anyone who was depressed but somehow got out of it ?....
7	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
8	Bodily Functions	Mention of bodily excretions, physical 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 smartphones
9	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!

were labeled, of which 1000 were labeled by all the annotators (agreement >93%) and the remaining 3000 were split between the annotators.

In the event of contention between two or more categories, the category that is most pertinent was selected. It is to be noted that the topic of the post content can be different from the taboo topic mentioned in the post. In Table 2, description of each taboo category is presented with their relative percentage with respect to taboo posts and an example. Table 3 delineates example posts in which the general topic of the post was distinctly different from the taboo. A table cataloging examples of taboos, labeled by the annotators for the YikYak dataset is provided in the Appendix.

The annotators found about 30% of all the posts were taboo-related i.e. belonged to the 9 taboo categories, and the remaining 70% belonged to non-taboo categories. This is not expected as we don't expect majority of disclosures to be taboo. Following are some examples which we did not consider taboos although a simple semantic topic analysis of the posts might tag them with that taboo. This is because the disclosure is not discussing something 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. Academics: I cant stand this school, and I'm sick of trying to. - i) mention of "this school" is not related to performance and hence not tagged as academics, and ii) as the student seems to leave the school and not suicidal - hence not tagged as mental health

Table 3 Examples from each category where the taboo topic was distinct from the general topic of the post.

Taboo category	**Example**	**Reason**
Protected Groups	I hate Asian food.	It does not refer to a hatred of a community but a cuisine.
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 annotated as mental health and not death.
Academics	I can't stand this school, and I'm sick of trying to	Mention of 'this school' is not related to performance or grades
Illegal Substances	I wish more people on campus smoked. Not anything illegal, just cigarettes. I actually dont even smoke but second-hand smoke is the absolute best. Its such a good smell but its a pretty rare occurrence that someone is walking across campus while smoking.	Text not related to illegal drugs, just cigarettes.
Physiological Health	Yesterday night was the first time I had thrown up since I got to Brown. No it wasn't because of alcohol... it was because of that goddamned sushi from The Gate! Don't eat that shit! You will vomit for three straight hours	Refers to food poisoning/temporary illness that is not a stigmatized/chronic medical category.
Mental Health	I think the university should be embarrassed by the state of the CAPS program. Students seriously need that help	Complaining about a lack of adequate mental health facilities different from revealing own status as mentally ill.
Financial	To anyone who works as a waiter or a waitress: I like to go out to dinner every now and then with my girlfriend.I don't have much money so I can't give out larger tips, but like to show my appreciated with what I have.	Asking about tipping etiquette. Admitting class status but not explicitly and not as a taboo.
Bodily Functions	I've tried to have sex multiple times but it never really works out, the whole fitting the <explicit>into the <explicit>. ... How do we learn AGHH someone tell me bc I am horny as <explicit>	Mention of private parts not in context of bodily function but sex.
Sex	I've made it a goal to hook up with (almost) every girl from a certain sorority	Hooking up not necessarily synonymous with sex

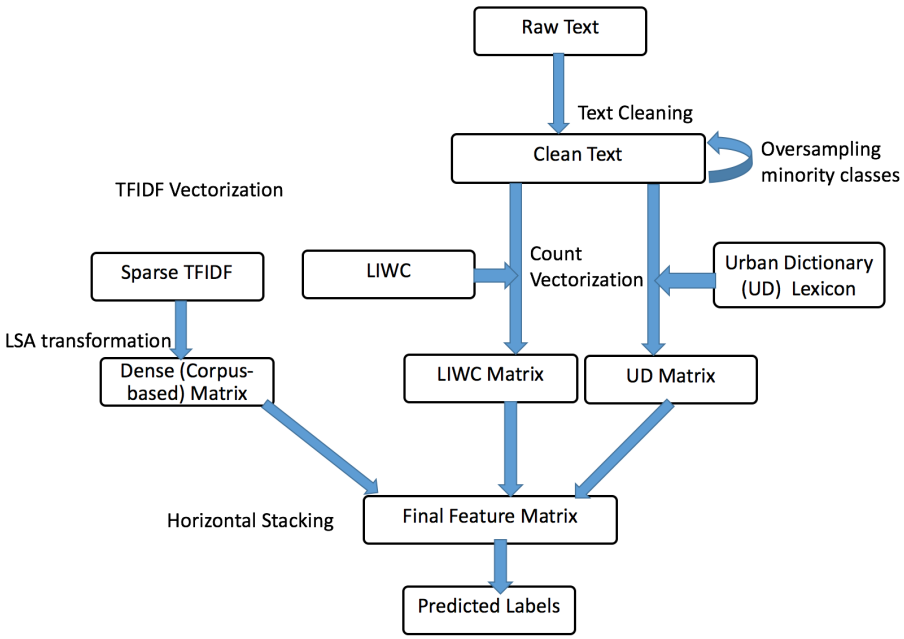

Fig. 1 Flow-diagram of the entire text mining system.

4 Method

In this section, the various steps involved in the proposed taboo categorization system - text cleaning, oversampling of minority classes, vectorization and classification are described. Table 1 depicts the flow diagram of the system.

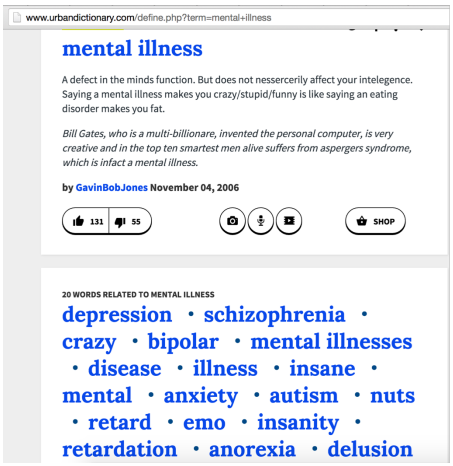

Fig. 2 Example from Urban Dictionary (courtesy: urbandictionary [49])

4.1 Data Cleaning

Data cleaning [81] is a preliminary and integral step in social text mining as text data from social media is highly unstructured and noisy in nature. Furthermore, the annotators observed that FCB data contained more noise in the form of bad grammar or typos compared to generic Facebook pages. This is expected as previous research have indicated that users of anonymous environments are less concerned with self-presentation as compared to identified spaces [82].

Furthermore, posts that only included URLs were removed. A preliminary text analysis illustrated that posts that commenced with urls contained spam or some generic information. It is to be noted that the removal of slangs was avoided during the text cleaning phase as the context from slang words are harnessed in the proposed approach.

The TextBlob API [83] was used for grammar and typo correction.

4.2 Oversampling of minority classes

The taboo posts formed a small percentage (30%) of the entire dataset, and many of the taboo categories formed less than 1% of the labeled corpus. To make sure the taboo posts were a representative sample of the whole set of FCB posts, posts containing taboo were oversampled. Oversampling is a common procedure in spam detection algorithms as spam emails are a small subset of the universal set of all emails. The imbalance in the labeled dataset is compensated by applying an oversampling technique called Synthetic Minority Oversampling Technique (SMOTE) [84]. In this approach, k nearest neighbors of a training sample belonging to the minority class are generated. Thus, the minority class(or classes) is oversampled exploiting the artificial training

samples. Random oversampling techniques were also investigated but SMOTE delivered better performance. Different degrees of oversampling, the number of times a minority class sample is oversampled, were investigated. The best trade-off between performance and over-fitting was determined at an oversampling of 100 % - on average, each taboo post is repeated once.

4.3 Vectorization and Classification

For the problem of text categorization of a document, the usual $tf - idf$ based representation of a document is a feature-vector representation of a given document as a set of term sequences, including term t and term weight w. The document is made up of pairs of $<t, w>$with the term and weight representing the features which express the post content and value relevant to the coordinate respectively. Thus, every document (d) is mapped to the target space as a feature vector. In the case of the term frequency, the simplest choice is to use the raw frequency of a term in a document, the number of times that term t occurs in the document d. The inverse document frequency is a metric for determining how much information the document can provide, that is, whether the term is commonly or rarely present across all the documents. Mathematically, it is the logarithmically scaled inverse fraction of the documents that contain the word. We obtain it by dividing the total number of documents by the number of documents containing the term and then taking the logarithm of the quotient. The $tf - idf$ matrix representation is obtained by taking a product of the term frequency with the inverse document frequency.

The LIWC [85] and Urban Dictionary lexicons were used to enrich the vector space and introduce more context into the model. For Urban Dictionary, a python based scraper was designed, that could extract related words for the top 20 words based on the results of the $tf - idf$ vectorized classification model for each taboo category. Fig 2 presents a snapshot of the related words returned by the Urban dictionary website for the search term 'mental illness'. It is to be noted that the entire Urban Dictionary is not used as the lexicon. The primary motivation of using Urban Dictionary stems from the reasoning that the confessions corpus for the disclosures contained slangs and modern cultural references that can be harnessed and incorporated into the model. The LIWC text analysis tool provides 64 semantic categories. For both the Urban Dictionary and LIWC lexicons, a vector of token counts based on a number of occurrences is constructed. The Urban Dictionary-based matrix is composed of count vectors for each of the nine taboo categories, and the LIWC-based matrix is composed of count vectors for each of the 64 LIWC categories. A catalog of words extracted from Urban Dictionary and words for selected LIWC categories are presented in the appendix. The BeautifulSoup [86] and requests [87] library was employed for extracting the related words from the Urban Dictionary website.

The vectorizer is constructed by first creating a sparse $tf - idf$ representation of the corpus. LSA transformation is performed to transform the matrix to a dense representation using dimensionality reduction via singular value decomposition. The resultant dense matrix is stacked with the Urban Dictionary (UD) and LIWC-based feature matrix. Different combinations of stacking the vectorizer matrices were explored. Table 5 provides a comparison of the cross-validation accuracies for each of the combinations. It must be emphasized that in this work, stacking refers to feature stacking that combines distinct sets of features from multiple sources. Feature stacking is distinct from model stacking that involves stacking multiple models for performing supervised classification.

The Scikit-Learn [88] library was used for feature engineering, dimensionality reduction and supervised machine learning tasks. The Gensim [89] library was employed for generating word embeddings. Keras [90] wrapper with Tensorflow [91] backend was availed for the benchmarking experiments using LSTM.

4.4 Transfer Learning

Transfer learning can help us harness the learned context of learning from a source dataset for a task on a destination dataset. This is critical in the context of anonymous social media in particular as it one anonymous confession platform can get shut down or lose popularity. We observed churn of users from FCBs to YikYak and then after YikYak closed down [92], there was a churn to Whisper [93] and Reddit confession forums, and now FCBs are regaining popularity. Due to the ephemerality of these platforms, it would ordinarily require regenerating training data for each new forum, and getting high quality annotated data can be logistically expensive. One of the motivations of this work was to demonstrate that a dictionary based approach from the corpus of one technology medium can work on another medium.

Typically in transfer learning, the source dataset might provide more overall thematic context which the destination dataset may not be able to provide. The destination dataset on the other hand provides more specific context. Given that FCB does not have character or word limits, we believe that we can gain more contextual information from the FCB dataset that cannot be achieved as effectively in the shorter yak posts - restricted to 200 characters. The transfer learning experiments from FCB (source) on the YikYak (target) dataset instead of training a combined model would help us validate and evaluate the efficacy and generalizability of the dictionary-based approach. There is a second reason for choosing FCBs as the source dataset - FCBs are still active forums while YikYak has been retired. FCBs has been around for almost a decade and while some individual university pages have stopped generating content or have been closed, new FCB pages have started.

5 Results and Discussion

In this section, we would present the experimental results on the FCB and YikYak datasets using proposed approach, and compare them with other approaches including state-of-the-art techniques such as LSTM and Embeddings. Further, we discuss some rationale behind the superior performance of our proposed algorithm with other techniques for this problem.

5.1 Experimental Results

Table 4 presents the comparison of cross-validation accuracy for the proposed stacked vectorizer across different machine learning algorithms about other text vectorization schemes that have proved to be successful for various text mining tasks. Extensive grid search across hyper-parameters and different combinations of stopword lists and n-gram range were performed for all the machine learning algorithms until the best cross-validation performance was achieved. For the LSTM classifier, various combinations of loss functions, batch sizes, and dropout were explored.

Table 4 Evaluation of cross-validation accuracy across different models for FCB and YikYak are presented(* indicates models that used other vectorizers instead of the proposed vectorizer in this work).

Model	**FCB%**	**YikYak%**
Bag of Words [94]*+LinearSVM	0.68	0.63
$tf - idf$ [95]*+LinearSVM	0.73	0.65
Multinomial Naive Bayes [56]	0.72	0.64
Bernoulli Naive Bayes [56]	0.72	0.63
Linear SVM [57]	0.75	0.68
Random Forest [58]	**0.76**	**0.69**
Extra Trees [59]	**0.78**	**0.71**
LSA (Unigram) [60]*	0.54	0.42
LSA (Bigram) [60]*	0.56	0.40
GloVe Embedding (Unigram) [61]*	0.60	0.45
Glove Embedding (Bigram) [61]*	0.63	0.48
CBOW Embedding (Unigram) [62]*	0.51	0.49
CBOW Embedding (Bigram) [62])*	0.53	0.49
Skipgram Embedding (Unigram) [63]*	0.58	0.46
Skipgram Embedding (Bigram) [63]*	0.59	0.45
LSTM [96]	0.69	0.62
LSTM (with CBOW) [96]*	0.72	0.65

The prediction accuracy using RandomForests and ExtraTrees algorithms and the proposed vectorization scheme on the FCB dataset surpasses the accuracy using LinearSVM on a vanilla $tf - idf$ representation (statistically significant $p<0.01$). Figures 3, 4 present the confusion matrices for classification using vanilla $tf - idf$ representation and our proposed vectorization

scheme respectively. Figure 5 illustrate the confusion matrix for the predicted labels after cross-validation.

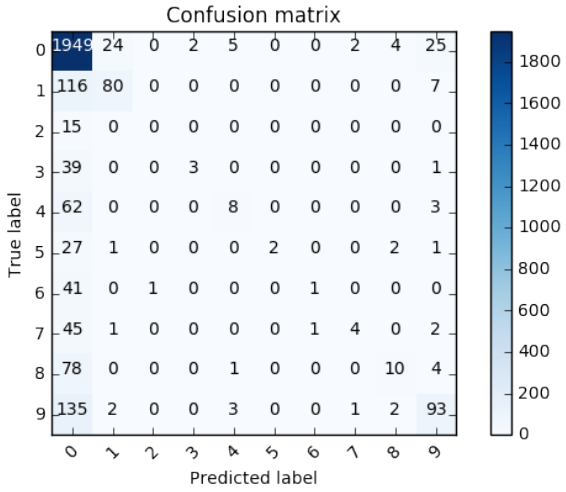

Fig. 3 Confusion Matrix using $tf - idf$. Labels 1 to 9 are the class labels for the taboo categories that use the same scheme depicted in Table 2. Label 0 is the label attributed to a post with no taboo.

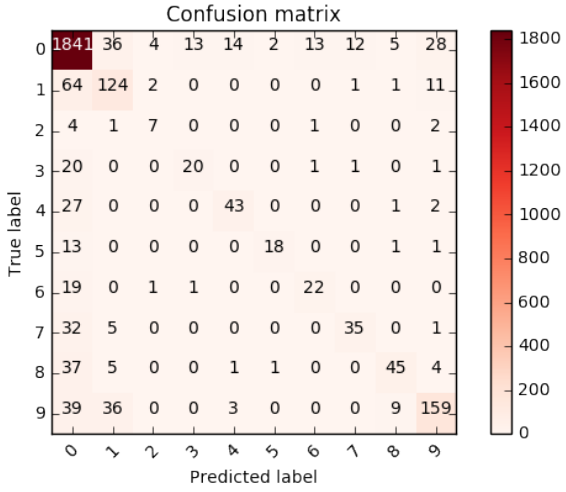

Fig. 4 Confusion Matrix using proposed model. Labels 1 to 9 are the class labels for the taboo categories that use the same scheme depicted in Table 2. Label 0 is the label attributed to a post with no taboo.

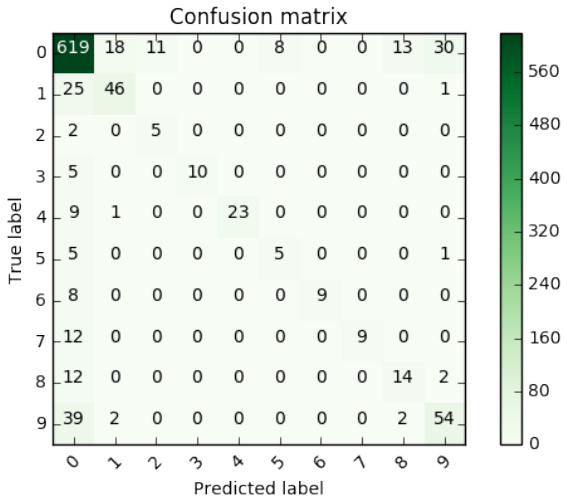

Fig. 5 Confusion matrix for the YikYak dataset. Labels 0 to 9 are the class labels for the taboo categories that use the same scheme depicted in Table 2. Label 0 is the label attributed to a post with no taboo.

Different feature stacking combinations for the proposed vectorizer are explored, and the results of the comparison are presented in Table 5. The prediction accuracy using RandomForests and ExtraTrees algorithms on the proposed vectorization scheme on the YikYak dataset surpasses the accuracy using LinearSVM and a vanilla $tf-idf$ representation (statistically significant $p<0.05$). The accuracy for the transfer learning task is lower across all the algorithms compared to the supervised task. We gather there are two primary rationales for this. The YikYak dataset has a different distribution of taboo categories compared to FCBs. Furthermore, the annotation schema used for labeling was primarily developed for categorizing taboos in FCBs.

Table 5 Comparison of cross-validation accuracy across different combinations of the stacked vectorizer approach (upto 3 significant digits). For each of the combinations, the best model has been presented. ExtraTrees classifier performed best for all the stacked combinations.

Model	**FCB%**	**YikYak%**
tf-idf(vanilla)	0.571	0.562
tf-idf + LIWC	0.700	0.638
tf-idf + UD	0.701	0.659
LIWC + UD	0.732	0.702
***tf-idf*+LIWC + UD**	**0.781**	**0.705**

5.2 Discussion

In this work, we attempt to build a supervised learning approach to predict taboo topics by harnessing psycho-lingual and crowd-sourced dictionaries. The proposed vectorization approach was compared against other vectorization schemes namely Bag of Words, $tf-idf$, LSA, GloVe and Word2Vec. Although accuracy using vanilla $tf-idf$ was lower than the proposed stacked vectorizer, it performed much better than other vectorization approaches. This was not unexpected as Word2Vec models perform well on much larger datasets, and word embeddings in GloVe - trained on a corpus of wikipedia and twitter data has different distribution of content and semantic information compared to the FCB dataset.

For the vanilla $tf-idf$ based model, the best performance was achieved using LinearSVM [57]. This can be attributed to SVMs [97] being universal learners as support vectors can be considered independent of the dimensionality of the feature space. Hence, SVM can learn from sparse feature matrices originating from Bag of Words or $tf-idf$ representations. For the proposed stacked vectorizer, the best performance was achieved using Extra Trees and Random Forests classifiers. Both algorithms utilize an ensemble of decision trees that allow them to reduce the classification bias.

Although LSTMs do not perform at par with RandomForest or ExtraTrees for the FCB dataset, they perform better than other algorithms. It can be anticipated that training on a larger labeled corpus would lead to better cross-validation accuracy. The lower accuracy on transfer learning task on the YikYak dataset is understandable due to an even smaller dataset.

One interesting observation from this study was the superior performance of RandomForest and ExtraTrees compared to LinearSVMs, which usually perform best amongst traditional machine learning algorithms for text categorization tasks. This can be attributed to the reduced dimension of the feature matrix when using the proposed vectorizer compared to vanilla $tf-idf$.

A comparison of the confusion matrices for the vanilla $tf-idf$ representation (Figure 3) with the proposed stacked vectorizer (Figure 4) demonstrates the success of introducing context via use of lexicons. $tf-idf$ representation is better at categorizing texts that do not contain any taboo and this may be due to bias in the classifier towards the majority class which denotes no taboo. All the taboo categories are minority classes. However, as a result of both dimensionality reduction of the $tf-idf$ matrix as well as combining it with feature representation from the lexicons, the bias is reduced using our proposed vectorizer.

The novel vectorization scheme propounded in our study illustrates the scope of concept-driven supervised learning models to predict abstract topics such as taboos from a social media corpus. The importance of understanding context is even more important for supervised learning from a small dataset. Application of deep neural networks on text categorization tasks has suggested reduced need for feature engineering and reduction. However, the caveat with deep neural network-based models such as LSTMs or convolutional neural net-

works is that it usually necessitates a large labeled dataset. Thus, for smaller datasets, an explicit understanding of the dataset domain, and subsequent feature engineering can produce better prediction accuracy. Table 5 depicts that inclusion of both corpus and lexicon-based information help in enriching prediction models and supersede accuracy compared to only corpus or lexicon based feature representations.

6 Conclusions & Future Work

A methodology for prediction of taboo topics from social media disclosures using the synthesis of a corpus-based approach with crowd-sourced and psycholingual lexicons is propounded in this work. Psychological text analysis tool LIWC and crowd-sourced dictionary Urban Dictionary are combined with $tf - idf$ vectorization for supervised learning of taboos from anonymous social media datasets. The proposed approach that stacks feature matrices extracted from corpus and lexicon-based approaches deliver higher prediction accuracy than learning from corpus-based or lexicon-based approaches alone. The proposed methodology achieves cross-validation accuracies of up to 78.1% on the supervised learning task on FCB dataset and 70.5% on the transfer learning task on the YikYak dataset. With this ensemble methodology, abstract concepts or themes (in this case taboo) can be identified. The relative success of transfer learning on the YikYak dataset hints at the success of generalizing the approach for supervised learning from self-disclosure texts to learn abstract themes.

An effective active learning system can lower the expense of annotation by selecting samples that would be essential for improving classification accuracy. Furthermore, we plan to release this work in the future as a web-based application and API where a client can submit a social media post or an unlabeled corpus respectively as a request and obtain a prediction with the confidence score for each taboo category. The success of ensemble decision tree based algorithms in reducing bias in the classification results urges the exploration of combining multiple learning models using boosting and bagging [98]. Although word2vec did not yield satisfactory results on the FCB and YikYak datasets, future exploration of paragraph vector [99] can overcome the loss of semantic information while learning from a dataset of varying lengths. We would urge researchers to investigate other combinations of combining corpus and lexicon-based approaches, including combining embedding-based approaches with lexicons.

Acknowledgements This work is supported in part by the following grants: NSF award CCF-1409601; DOE awards DE-SC0014330, DE-SC0019358.

Conflict of Interest Statement

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Ilaria Montagni, Isabelle Parizot, Aine Horgan, Juan-Luis Gonzalez-Caballero, José Almenara-Barrios, Carolina Lagares-Franco, Juan-Luis Peralta-Sáez, Pierre Chauvin, and Francesco Amaddeo. Spanish students use of the internet for mental health information and support seeking. *Health informatics journal*, 22(2):333–354, 2016.
2. Meredith Ringel Morris, Jaime Teevan, and Katrina Panovich. What do people ask their social networks, and why?: a survey study of status message q&a behavior. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 1739–1748. ACM, 2010.
3. Alexis Quesada-Arencibia, Enrique Pérez-Brito, Carmelo R García-Rodríguez, and Ana Pérez-Brito. An ehealth information technology platform to help the treatment of mental disorders. *Health informatics journal*, 24(4):337–355, 2018.
4. Ray B Jones and Emily J Ashurst. Online anonymous discussion between service users and health professionals to ascertain stakeholder concerns in using e-health services in mental health. *Health informatics journal*, 19(4):281–299, 2013.
5. John Suler. The online disinhibition effect. *Cyberpsychology & behavior*, 7(3):321–326, 2004.
6. Alison M Radcliffe, Mark A Lumley, Jessica Kendall, Jennifer K Stevenson, and Joyce Beltran. Written emotional disclosure: Testing whether social disclosure matters. *Journal of social and clinical psychology*, 26(3):362–384, 2007.
7. Munmun De Choudhury, Meredith Ringel Morris, and Ryen 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.
8. Mark W Newman, Debra Lauterbach, Sean A Munson, Paul Resnick, and Margaret E Morris. It's not that i don't have problems, i'm just not putting them on facebook: challenges and opportunities in using online social networks for health. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, pages 341–350. ACM, 2011.
9. Patrick B Osullivan and Andrew J Flanagin. Reconceptualizing flamingand other problematic messages. *New Media & Society*, 5(1):69–94, 2003.
10. Elizabeth Whittaker and Robin M Kowalski. Cyberbullying via social media. *Journal of School Violence*, 14(1):11–29, 2015.
11. Derald Wing Sue, Christina M Capodilupo, Gina C Torino, Jennifer M Bucceri, Aisha Holder, Kevin L Nadal, and Marta Esquilin. Racial microaggressions in everyday life: implications for clinical practice. *American psychologist*, 62(4):271, 2007.
12. Yik yak perpetuates culture of intolerance — the emory wheel. `http://emorywheel.com/yik-yak-perpetuates-culture-of-intolerance/`. (Accessed on 04/15/2017).
13. The daily northwestern : Hayes: Yik yak unveils social problems. `http://dailynorthwestern.com/2014/05/14/opinion/hayes-yik-yak-unveils-social-problems/`. (Accessed on 04/15/2017).
14. Munmun De Choudhury and Sushovan De. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *ICWSM*. Citeseer, 2014.
15. Jeremy Birnholtz, Nicholas Aaron Ross Merola, and Arindam 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.
16. Erik H Erikson. Major stages in psychosocial development. *The life cycle completed: A review*, pages 55–82, 1982.

17. Patrick Corrigan. How stigma interferes with mental health care. *American psychologist*, 59(7):614, 2004.
18. Siobhan ONeill, Raymond R Bond, Alexander Grigorash, Colette Ramsey, Cherie Armour, and Maurice D Mulvenna. Data analytics of call log data to identify caller behaviour patterns from a mental health and well-being helpline. *Health informatics journal*, page 1460458218792668, 2018.
19. Karen Clarke, John Rooksby, and Mark Rouncefield. You've got to take them seriously': meeting information needs in mental healthcare. *Health informatics journal*, 13(1):37–45, 2007.
20. Ljilja Ruzic and Jon A Sanford. Needs assessmentmhealth applications for people aging with multiple sclerosis. *Journal of Healthcare Informatics Research*, 2(1-2):71–98, 2018.
21. Chandan Sarkar, Donghee Yvette Wohn, and Cliff Lampe. Predicting length of membership in online community everything2 using feedback. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work Companion*, pages 207–210. ACM, 2012.
22. Alcides Velasquez, Rick Wash, Cliff Lampe, and Tor Bjornrud. Latent users in an online user-generated content community. *Computer Supported Cooperative Work (CSCW)*, 23(1):21–50, 2014.
23. Donghee Wohn, Alcides Velasquez, Tor Bjornrud, and Cliff Lampe. Habit as an explanation of participation in an online peer-production community. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2905–2914. ACM, 2012.
24. Cliff Lampe and Erik Johnston. Follow the (slash) dot: effects of feedback on new members in an online community. In *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work*, pages 11–20. ACM, 2005.
25. Francesca DErrico and Isabella Poggi. Acidity. the hidden face of conflictual and stressful situations. *Cognitive Computation*, 6(4):661–676, 2014.
26. Nir Ofek, Soujanya Poria, Lior Rokach, Erik Cambria, Amir Hussain, and Asaf Shabtai. Unsupervised Commonsense Knowledge Enrichment for Domain- Specific Sentiment Analysis. *Cognitive Computation*, 8(3):467–477, February 2016.
27. Farhan Hassan Khan, Usman Qamar, and Saba Bashir. Multi-Objective Model Selection (MOMS)-based Semi-Supervised Framework for Sentiment Analysis. *Cognitive Computation*, 8(4):614–628, February 2016.
28. Farhan Hassan Khan, Saba Bashir, and Usman Qamar. Tom: Twitter opinion mining framework using hybrid classification scheme. *Decision Support Systems*, 57:245–257, 2014.
29. Keith Feldman, Spyros Kotoulas, and Nitesh V Chawla. Tiqs: Targeted iterative question selection for health interventions. *Journal of Healthcare Informatics Research*, pages 1–23, 2018.
30. Stephen D Reicher, Russell Spears, and Tom Postmes. A social identity model of deindividuation phenomena. *European review of social psychology*, 6(1):161–198, 1995.
31. Tom Postmes, Russell Spears, Khaled Sakhel, and Daphne 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.
32. Kai Sassenberg and Tom Postmes. Cognitive and strategic processes in small groups: Effects of anonymity of the self and anonymity of the group on social influence. *British Journal of Social Psychology*, 41(3):463–480, 2002.
33. Richard C Wildman. Effects of anonymity and social setting on survey responses. *Public Opinion Quarterly*, 41(1):74–79, 1977.
34. Munmun 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.
35. Acar Tamersoy, Munmun De Choudhury, and Duen Horng 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.
36. Nazanin Andalibi, Oliver L Haimson, Munmun De Choudhury, and Andrea 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.

37. Gunther Eysenbach, John Powell, Marina Englesakis, Carlos Rizo, and Anita 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.
38. Courtney Blackwell, Jeremy Birnholtz, and Charles Abbott. Seeing and being seen: Co-situation and impression formation using grindr, a location-aware gay dating app. *new media & society*, page 1461444814521595, 2014.
39. Yik yak - find your herd. https://www.yikyak.com.
40. Amy Binns. Facebooks ugly sisters: Anonymity and abuse on formspring and ask. fm. *Media Education Research Journal*, 2013.
41. Jeremy Birnholtz, Colin Fitzpatrick, Mark Handel, and Jed R Brubaker. Identity, identification and identifiability: The language of self-presentation on a location-based mobile dating app. In *Proceedings of the 16th international conference on Human-computer interaction with mobile devices & services*, pages 3–12. ACM, 2014.
42. Daniel M Sutko and Adriana de Souza e Silva. Location-aware mobile media and urban sociability. *New Media & Society*, page 1461444810385202, 2011.
43. Sarah Vieweg, Amanda L Hughes, Kate Starbird, and Leysia Palen. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1079–1088. ACM, 2010.
44. Cyclopath. http://cyclopath.org. (Accessed on 09/27/2016).
45. Everyblock. http://www.everyblock.com. (Accessed on 09/27/2016).
46. Leslie A Baxter and William W Wilmot. Taboo topics in close relationships. *Journal of Social and Personal Relationships*, 2(3):253–269, 1985.
47. Robin Goodwin and Iona Lee. Taboo topics among chinese and english friends a cross-cultural comparison. *Journal of Cross-Cultural Psychology*, 25(3):325–338, 1994.
48. Kevin Lanning and Geoffrey Maruyama. The social psychology of the 2008 us presidential election. *Analyses of Social Issues and Public Policy*, 10(1):171–181, 2010.
49. Urban Dictionary. Urban dictionary, llc. *San Francisco, available at www. urbandictionary. com/define. php*, 2013.
50. Nora McLeese. How selfie got into the dictionary: an examination of internet linguistics and language change online. 2015.
51. David Crystal. *Internet linguistics: A student guide*. Routledge, 2011.
52. Andreas H Jucker and Christa Dürscheid. The linguistics of keyboard-to-screen communication: A new terminological framework. *Linguistik online*, 56(6), 2013.
53. Dictionary.com — meanings and definitions of words at dictionary.com. http://www.dictionary.com. (Accessed on 09/23/2016).
54. Dictionary and thesaurus — merriam-webster. http://www.merriam-webster.com. (Accessed on 09/23/2016).
55. Haiyi Zhang and Di Li. Naive bayes text classifier. In *Granular Computing, 2007. GRC 2007. IEEE International Conference on*, pages 708–708. IEEE, 2007.
56. Andrew McCallum, Kamal Nigam, et al. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Madison, WI, 1998.
57. Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pages 137–142. Springer, 1998.
58. Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.
59. Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006.
60. Scott C Deerwester, Susan T Dumais, George W Furnas, Richard A Harshman, Thomas K Landauer, Karen E Lochbaum, and Lynn A Streeter. Computer information retrieval using latent semantic structure, June 13 1989. US Patent 4,839,853.
61. Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543, 2014.
62. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

63. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
64. Liza Wikarsa and Sherly Novianti Thahir. A text mining application of emotion classifications of twitter's users using naive bayes method. In *2015 1st International Conference on Wireless and Telematics (ICWT)*, pages 1–6. IEEE, 2015.
65. Diana Lupan, Mihai Dascălu, tefan Trăuan-Matu, and Philippe Dessus. Analyzing emotional states induced by news articles with latent semantic analysis. In *International Conference on Artificial Intelligence: Methodology, Systems, and Applications*, pages 59–68. Springer, 2012.
66. Jonathan Herzig, Michal Shmueli-Scheuer, and David Konopnicki. Emotion detection from text via ensemble classification using word embeddings. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, pages 269–272, 2017.
67. Annika M Schoene, George Lacey, Alexander P Turner, and Nina Dethlefs. Dilated lstm with attention for classification of suicide notes. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 136–145, 2019.
68. Ming-Hsiang Su, Chung-Hsien Wu, Kun-Yi Huang, and Qian-Bei Hong. Lstm-based text emotion recognition using semantic and emotional word vectors. In *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*, pages 1–6. IEEE, 2018.
69. Stevie Chancellor and Munmun De Choudhury. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine*, 3(1):1–11, 2020.
70. Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.
71. Mit confessions. https://www.facebook.com/beaverconfessions.
72. National university rankings — top national universities — us news best colleges. http://colleges.usnews.rankingsandreviews.com/best-colleges/rankings/national-universities. (Accessed on 09/27/2016).
73. National liberal arts college rankings — top liberal arts colleges — us news best colleges. http://colleges.usnews.rankingsandreviews.com/best-colleges/rankings/national-liberal-arts-colleges. (Accessed on 09/27/2016).
74. Jesse Weaver and Paul Tarjan. Facebook linked data via the graph api. *Semantic Web*, 4(3):245–250, 2013.
75. Timeline scraper - dashboard - facebook for developers. `https://developers.facebook.com/apps/463500207102372/dashboard/`. (Accessed on 06/23/2017).
76. Braden Groom. Pyak. https://github.com/bradengroom/pyak, 2015.
77. Carson L Nemelka, Cameron L Ballard, Kelvin Liu, Minhui Xue, and Keith W Ross. You can yak but you can't hide. In *Proceedings of the 2015 ACM on Conference on Online Social Networks*, pages 99–99. ACM, 2015.
78. Elena Kadvany. Anonymous confessions pages are surging in popularity on high school and college campuses. why?, Mar 2020.
79. Amazon mechanical turk - welcome. https://www.mturk.com/mturk/welcome. (Accessed on 10/12/2016).
80. Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. Amazon's mechanical turk a new source of inexpensive, yet high-quality, data? *Perspectives on psychological science*, 6(1):3–5, 2011.
81. Erhard Rahm and Hong Hai Do. Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, 23(4):3–13, 2000.
82. Nicole C Krämer and Stephan Winter. Impression management 2.0: The relationship of self-esteem, extraversion, self-efficacy, and self-presentation within social networking sites. *Journal of media psychology*, 20(3):106–116, 2008.
83. Steven Loria. Textblob: simplified text processing. *Secondary TextBlob: Simplified Text Processing*, 2014.
84. Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, pages 321–357, 2002.

85. James W Pennebaker, Roger J Booth, and Martha E Francis. Liwc2007: Linguistic inquiry and word count. *Austin, Texas: liwc. net*, 2007.
86. Leonard Richardson. Beautiful soup documentation, 2007.
87. Requests: Http for humans requests 2.18.1 documentation. http://docs.python-requests.org/en/master/.
88. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011.
89. R Rehurek and P Sojka. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 2011.
90. François Chollet et al. Keras, 2015.
91. TensorFlow Team. Tensorflow: Large-scale machine learning on heterogeneous systems.(2015). *Software available from tensorflow.org*, 2015.
92. Grant Williams and Anas Mahmoud. Modeling user concerns in the app store: A case study on the rise and fall of yik yak. In *2018 IEEE 26th International Requirements Engineering Conference (RE)*, pages 64–75. IEEE, 2018.
93. Rachel A Grunkemeyer. 10. whisper–an effective use of anonymous persuasion? 2016.
94. Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.
95. Hans Peter Luhn. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of research and development*, 1(4):309–317, 1957.
96. Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
97. Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992.
98. J Ross Quinlan et al. Bagging, boosting, and c4. 5. In *AAAI/IAAI, Vol. 1*, pages 725–730, 1996.
99. Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1188–1196, 2014.