Age and Gender Prediction on Health Forum Data

Prasha Shrestha[1], Steven Bethard[2], Ted Pedersen[3], Nicolas Rey-Villamizar[1], Farig Sadeque[2], and Thamar Solorio[1]

[1]University of Houston, [2]University of Alabama at Birmingham, [3]University of Minnesota, Duluth
pshrestha3@uh.edu, bethard@uab.edu, tpederse@d.umn.edu, nrey@uh.edu, farigys@uab.edu, solorio@cs.uh.edu

Abstract

Health support forums have become a rich source of data that can be used to improve health care outcomes. A user profile, including information such as age and gender, can support targeted analysis of forum data. But users might not always disclose their age and gender. It is desirable then to be able to automatically extract this information from users' content. However, to the best of our knowledge there is no such resource for author profiling of health forum data. Here we present a large corpus, with close to 85,000 users, for profiling and also outline our approach and benchmark results to automatically detect a user's age and gender from their forum posts. We use a mix of features from a user's text as well as forum specific features to obtain accuracy well above the baseline, thus showing that both our dataset and our method are useful and valid.

Keywords: medical forums, author profiling, gender, age

1. Introduction

Users who are actively engaged on health support forums volunteer rich personal details including disease symptoms, treatment history, medications, family history, and traces of their emotional status. Some active users might even have their entire medical history contained among these posts and replies. Since these forums have a certain level of anonymity, people might be more willing to talk candidly about their conditions. People going through similar situations as the original poster may also open up. Some people might come to the forums directly before going to their doctors, while others might create a post after a visit to their doctor. A survey of users of the DailyStrength health forum found that 160 out of 274 users went to DailyStrength after visiting their doctor (Bell et al., 2011). Moreover, out of the people who gave a perfect trust score to their doctor, more than half turned to the Internet for additional information and support. Although the suggestions provided in these forums as a means of treatment might not be appropriate (Culver et al., 1997), there is no doubt that there is valuable information in them. For example, patient's side effects or symptoms during the course of a treatment could be extracted from the posts and therefore complement the information available to physicians.

Another relevant task is to identify patients for clinical trials since these communities are so widespread across different age groups, ethnicities, socio-economic statuses and regions of the world. To identify potential patients, basic inclusionary criteria have to be met, such as age and gender. But not all users will provide this information publicly (in our data 35% did not include age or gender). User profiling in this data can then solve this problem and also enable aggregated studies of user diseases and symptoms, as well as their online behavior.

In addition to aiding studies performed on health forum data, demographic information of users will also be directly helpful. One clear use of this information is in the study of the relation of conditions to different age groups and/or genders. Most medical studies are performed on a particular group of people who fulfill certain gender and age group criteria. Even when a medical study requires data collected directly from the patients, the demographic information extracted from users in medical forums can be helpful in identifying cohorts that meet the inclusionary criteria for clinical trials.

At the time of this writing, our work is the first author profiling system based entirely upon medical forum data. Our aim is to predict the age group and gender of a user from their original posts or replies on the forum. We have used lexical features in the form of word n-grams of all the posts and replies of a user as our core features.

Apart from the prediction of age and gender, we also performed a thorough analysis of our dataset through the perspective of author profiling. This revealed differences in writing styles and word usages for users from different age groups and genders. We examined our dataset for clues and insights about the profile of the person making the post, such as forum topics where posts were made, length of texts of users, use of familial terms, use of abbreviations and use of age or gender specific words.

There are three main contributions in this paper. First, we present our author profiling dataset for health support forums. Our corpus is the first of its kind for author profiling of medical forum data. Second, we motivate author profile analysis in this type of data. Third, we propose a system that can predict age and gender of health forum users and present benchmarking results for future research in this area. We also discuss interesting findings about the salient topics in the different population groups.

2. Related Work

Medical forum data has been used in a variety of different research works since it is a comprehensive source of data. Jha and Elhadad (2010) have tried to predict the cancer stage of a patient by using the text in their posts and their online behavior. They formulated the problem as a multi-class classification problem with four cancer stages. They used unigrams and bigrams as their text based features and also

explored the use of network features with the hypothesis that patients in similar stages of cancer will interact more with each other. A combination of these three features gave them the best results.

Rolia et al. (2013) tried to make predictions about the condition a person is suffering from based on similarities of their condition to forum posts. Their system displays posts that appeared in medical forums to the patient and the patient gives the posts points according to whether this question is relevant to their condition or not. The system uses the patient's answers and makes an educated guess about the condition they are suffering from. They have used their system in the domain of diabetes but claim that their method is general and can be used for other domains.

Although there has not been any previous work on author profiling of medical forum data, there has been much previous research in other domains. The annual PAN author profiling shared task currently uses social media data such as blogs and tweets. The participants have approached this problem in many different ways (Rangel et al., 2013; Rangel et al., 2014). Most of them use a combination of various character, lexical, stylistic and syntactic features. Across all participants' works, word and character n-grams are the most frequently used features. Since the dataset contains a great deal of spam, some of the approaches also use a spam filter.

Schwartz et al. (2013) predicted age group, gender and personality traits on Facebook posts. They also used n-grams and LDA topics as features to obtain an accuracy of 91.9% for gender and a regression coefficient of 0.84 for age. They also generated word clouds for people of different ages and genders. These word clouds showed that there were distinct differences in word usages between people of different age and genders.

Schler et al. (2006) performed an analysis of the language usage of authors in blogs. They too found marked differences in the text written by authors of varying age and gender. They found that males write mostly on the topics of technology and politics while females write about relationships. They also performed an automatic author profiling. For age prediction, they divided their dataset into three age groups 13-17, 23-27 and 33-42 representing the 10s, 20s and 30s of a person's life. They intentionally used discontinuous groups so that there is no confusion between borderline ages in two groups. They obtained above 80% accuracy for gender and above 75% accuracy for age.

Estival et al. (2007) collected emails in English from both native and non-native speakers of English. Along with age, gender and personality, they also tried to predict native language and country. This work also uses similar character and lexical features. But they also added some email specific extra features such as the category of the email and the html text in the email. They tried to predict a different part of an author's profile: their first language and the country, along with age, gender and personality. They tried experimenting with a variety of classifiers including Support vector machines (SVMs), Random Forests and rule-based learners.While no one classifier was best for all traits, the SVM performed the best for both age and gender.

All of the research on author profiling uses lexical features, as they seem to be the biggest markers of a person's profile attributes. We have also used them here in our work. Some previous works also use domain specific features. We have also explored some of the features specific to medical forums. There is no consensus in the state-of-the-art about the representation of age. Some of the work view it as a continuous variable and perform regression (Schwartz et al., 2013) while most of the other works create their own age groups (Schler et al., 2006; Rangel et al., 2013; Rangel et al., 2014) and perform classification. In our work, we have also tried to see the effect of having different groupings for ages.

Age group	Female	Male	Total
12-17	3,939	715	4,654 (5.51%)
18-29	24,436	3,929	28,365 (33.56%)
30-49	31,496	6,799	38,295 (45.31%)
50-64	8,870	2,534	11,404 (13.49%)
65-	1,282	518	1,800 (2.13%)
Total	70,023 (82.85%)	14,495 (17.15%)	84,518

Table 1: Age and Gender Distribution in our DailyStrength Dataset

3. Health Forum Data

DailyStrength[1] is a medical support group where users can post questions about their ailments, and the community provides support and advice in the form of replies. Out of more than 500 support groups, our dataset consists of posts from the 104 most active groups. These posts were written from July 2006 to June 2015.

For the users who wrote these posts or replies, we also crawled their profile pages. Out of these people, around 65% had information about either their birth date or gender and around 60% had both. We discarded all of the users who did not have both age and gender information, who did not fall in the age range of 12-100 and who had less than 50 words in all of their posts and replies combined. The resulting dataset has 84,518 unique users.[2]

We have two classes for gender: male and female. We did not model age as a continuous variable. It is likely that a person creates posts and replies across several years. For users who post very sparsely, if we only take text from a single year, there will not be sufficient text or sufficient users of that particular age. Instead we divided the dataset into five different age range buckets: 12-17, 18-29, 30-49, 50-64, 65- representing different stages of life. The age group 12-17 represents school-age children, 18-29 represents young adults, 30-49 represents adults, 50-64 represents middle aged people and 65 above represents late adulthood. There are 511 users who have posted when they were at ages that fall into two different age groups. For these users, we separated their posts for the two age buckets and treated them as two different instances.

[1] http://www.dailystrength.org/

[2] Instructions for downloading the dataset can be found at: http://ritual.uh.edu/resources/

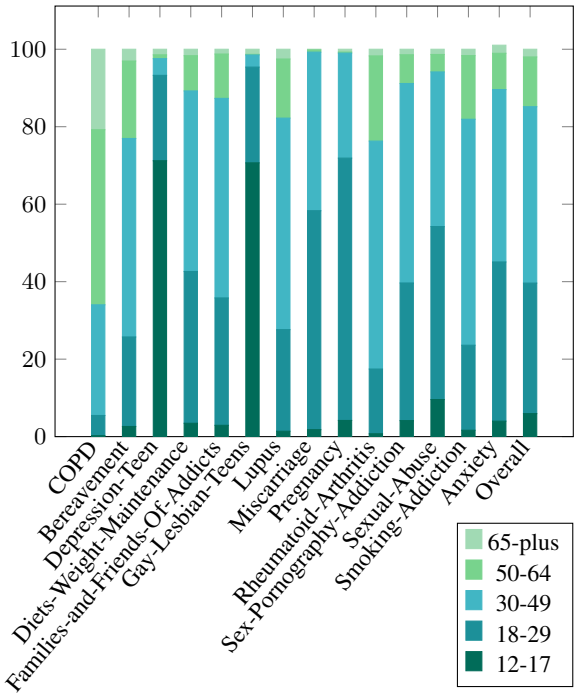


Figure 1: Age group distribution of users posting in support forums

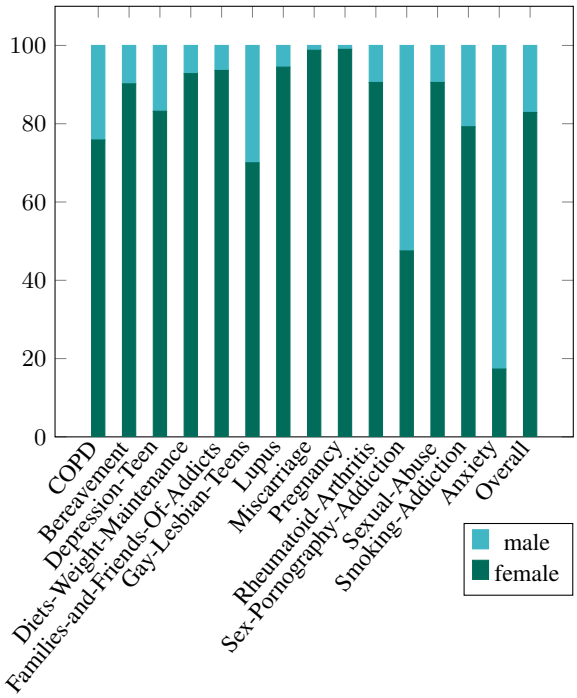


Figure 2: Gender distribution of users posting in support forums

The distribution of our dataset is shown in Table 1. There are nearly five times more female users than male users. Also, the most prolific age group using the forum is 30-49 followed by 18-29. Females of age group 30-49 are the largest demographic in our dataset. We have also tried to use a different age group formulation, which we will talk about in Section 5.

The 104 support groups in our dataset are mostly conditions

Label	Avg # of posts	Avg # of replies	Words per post	Words per reply
12-17	2.42	7.98	132.2	52.93
18-29	2.79	11.14	184.08	83.59
30-49	2.81	18.37	187.12	88.62
50-64	2.79	29.92	195.54	89.62
65-	3.52	34.56	168.73	89.99
female	2.84	17.12	182.65	86.63
male	2.56	17.33	192.28	88.24
Overall	2.79	17.16	184.14	86.9

Table 2: Counts of posts and replies on the training set

that affect both genders and age groups. The age and gender distributions for some of the support groups are shown in Figures 1 and 2 respectively. Only 3 out of the total 104 conditions have more males than females: *Sex-Pornography Addiction, Tinnitus*, and *Atrial Fibrillation*. As can be seen in Figure 2, there are a disproportionate number of females as compared to males. Users who post in conditions like *Miscarriage* and *Pregnancy* are almost all females. For age groups, 35-49 and 25-34 dominate most of the support groups, except for the ones directly aimed at teens such as *Depression-Teens* and *Gay-Lesbian-Teens*, where the age group 12-17 dominates. There is a significantly lower number of users in the age group 65-, with *Chronic Obstructive Pulmonary Disease (COPD)* having the highest percentage of users in this age range.

Table 2 shows the average number of posts/replies and the average words per post/reply for users in the corresponding gender and age group. The users are not very prolific and they are more likely to reply than to create original posts, although the replies are much shorter than the posts. A surprising finding is that the user group 65-and-above has a higher average number of posts than all other age groups, even though they are the smallest group. It is interesting to see that the people who are of age 65 and above are not very likely to use DailyStrength but those who do are very engaged in the forums.

4. Author Profiling

We will now outline our preliminary approach for author profiling in health support groups. As a preprocessing step, we remove the URLs present in the text and the words that were infrequent, occurring less than twice in the posts. For our approach, we tried a number of textual and forum related features. In order to test the usefulness of these features, we separated out 20% of our dataset as the development set. We used 60% as the training set and the remaining 20% as the test set.

We first tried some features specific to forum data. As certain conditions are more likely to affect a particular demographic, we tried to use as features the forums where an author has posted in the past. Liu and Ruths (2013) used first names for gender inference in Twitter data. We also noticed that some users leave clues in their username about gender. For instance, some usernames have *miss, princess, mom*, etc in them. We tried using username character n-grams as our features. But these features did not give us good results on the development set. We did not include them in our final

	Accuracy(%)		
Classification	Age	Gender	Both
Our system	65.59	88.41	58.89
Majority Class Baseline	45.55	82.87	37.38
Gender Balanced			
Our System	61.23	80.96	50.53
Majority Class Baseline	46.74	50.00	24.04

Table 3: Results on the test dataset

	Accuracy(%)		
Experiment	Age	Gender	Both
Our System	66.41	91.00	59.6
Majority Class Baseline	42.25	84.74	36.27
Gender Balanced			
Our System	65.39	82.20	52.80
Majority Class Baseline	43.29	50.00	23.77

Table 4: Results for the test with 3 age groups

system.

We then tried content-based features, namely word unigrams, bigrams and trigrams along with character trigrams. These features worked very well in the development set. Another type of feature that gave us good performance was familial tokens. We search for phrases such as *my husband/hubby/wife/girlfriend* that are highly indicative of a person's gender. We have observed that females use such familial terms 1.5 times more than men do in our dataset. We sort such familial terms into 3 buckets.[3] Phrases like *my husband/hubby/dh* fall into the female bucket (it is very likely that the user is female), phrases like *my wife/gf/girlfriend/dw* come under the male bucket, and the other familial terms used by both genders such as *my mother/father/sister* are in the neutral bucket. We then use these buckets as features. These features are highly predictive: whenever familial terms were present in a document, the phrases from the correct bucket appeared in the posts more than 96% of the time. However, these words do not appear frequently, so we used a cascading method. Whenever this feature is present in the text, we use the prediction from this feature alone. When it is not present, we use our model trained on the word n-grams. Note that these familial term features will only work for texts like those in support groups where the users are mostly talking about themselves.

We view our problem in two ways. First, as a 10-class classification problem where each class is a combination of an age group and a gender. Second, as a 5-class classification problem for age and a separate 2-class classification problem for gender. We used logistic regression as our classifier and explored the cost parameter for logistic regression. We obtained the highest accuracy for the development set in the separate 5-class and 2-class problem setting with c=0.1. Based upon these results on the development set, our final system treats age and gender separately and uses only n-grams for age while using the familial tokens features and n-grams cascaded for gender.

5. Results and Analysis

Our results are shown in Table 3. For comparison, the table also contains the majority class baseline. We obtained accuracies well above the baseline for both age and gender, even though the baseline for gender is already very high. A system that uses only the word and character n-grams gets 88.29% accuracy for gender. After cascading with the familial tokens system we were able to get a slight increase in gender accuracy to 88.41% as shown in the table. Our dataset is heavily imbalanced in the case of gender. The fact that men turn to health care forums much less than women, at least in DailyStrength, is also an interesting finding. We obtain 65.59% accuracy for age group classification and this is a good improvement over the majority class baseline of 45.55%. After combining the results from the two separate age and gender classifiers, we found that we can predict both the age and gender of an author correctly for 58.89% of the authors.

In order to find out how our system would perform if our dataset was not gender-imbalanced, we also created a gender balanced dataset by randomly sampling from the female set. Note that even in this dataset the data is still imbalanced across age groups, and gender is not balanced within an age group. When the data is balanced across gender, our system performs much better than the majority class baseline of 50%. Our system maintains higher than 80% accuracy for gender. This shows that our method works well for gender and the main reason we could not obtain good improvement over the baseline in our original dataset was due to the data being heavily imbalanced. The performance for age prediction decreased in this dataset, even though it is still well above the baseline.

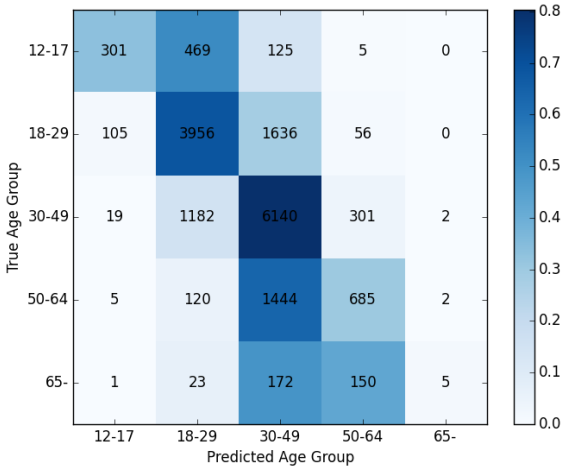


Figure 3: Confusion matrix for age group classification

5.1. Different Age Groups

In our experiments, we have divided the dataset into five age groups. Figure 3 shows the confusion matrix heatmap from the age group classification on our original (non-balanced) dataset. As we can see from the heatmap, most of the incorrect classification is due to confusion with adjacent age

[3]The list of familial tokens can be downloaded from `http://ritual.uh.edu/resources/`

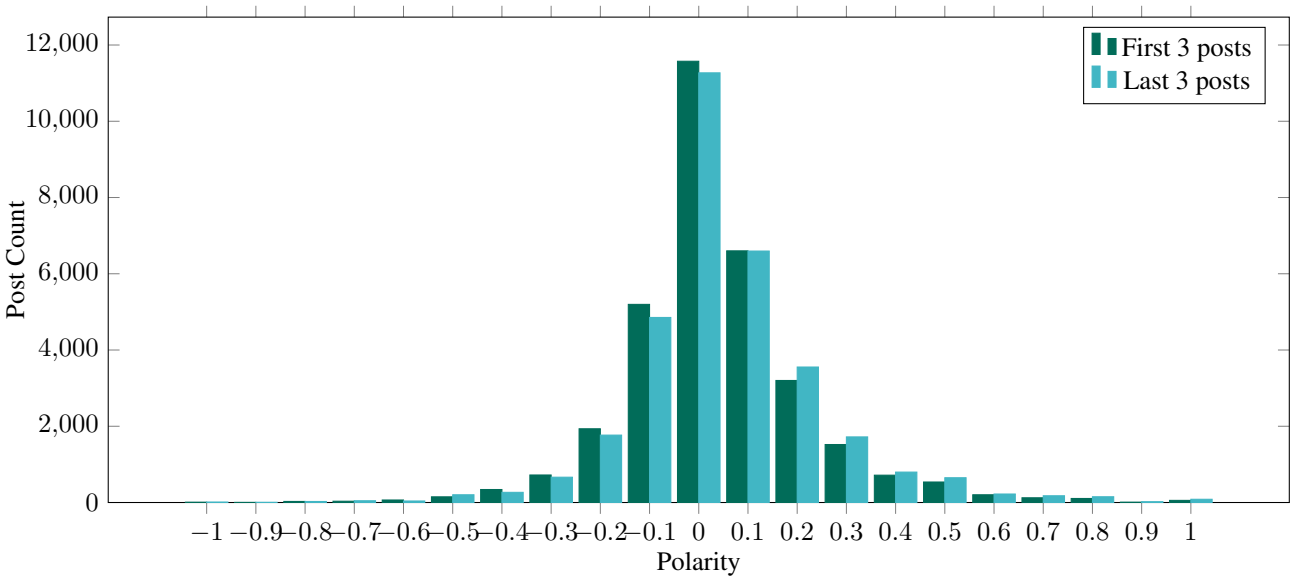


Figure 4: Polarity distribution of the first three and the last three posts of users

groups. In order to see how a change in the age groups affects our system, we divided the data into three disjoint age groups as defined by Schler et al. (2006). The three age groups are: 13-17, 23-27 and 33-42. The results from the three age groups experiment are in Table 4. In this setting as well our system performs well above the baseline. As expected, the results for age group prediction are higher with less number of classes even though the majority class baseline is lower for the three age groups experiment. The majority class for age is 33-42 and this is a subset of the majority class 30-49 of the original five age groups experiment. Since the three age groups are disjoint, there is also less chance of confusion between contiguous ages. This makes the problem inherently easier. The data became even more gender-imbalanced when we divided the data into three age groups. The baseline is slightly higher than the five age groups experiment and so is the result from our system for gender. We also performed a gender balanced experiment as before. The observations for gender were similar to what we observed for the dataset with five age groups. The results for age in the gender balanced dataset are again better than in the five age group experiment.

5.2. Word Usage Variation by Demographics

There were also some interesting observations such as instances of the usage of the word *girlfriend* by female users. When we looked at some of the posts, we saw that many of the females were referring to their friends as "girlfriend". This could have thrown our familial tokens feature off. Our n-gram method alone achieves a high accuracy for age. This shows that the word choices of people are the most characteristic of their age and gender.

We performed odds ratio keyword extraction (Demšar, 2006) from texts of male and female users, which are shown in Figure 5. There is a clear distinction between the type of words used by the two genders. Since the users in the forums related to pregnancy are predominantly female, most of the keywords are also related words such as *ovulation*,

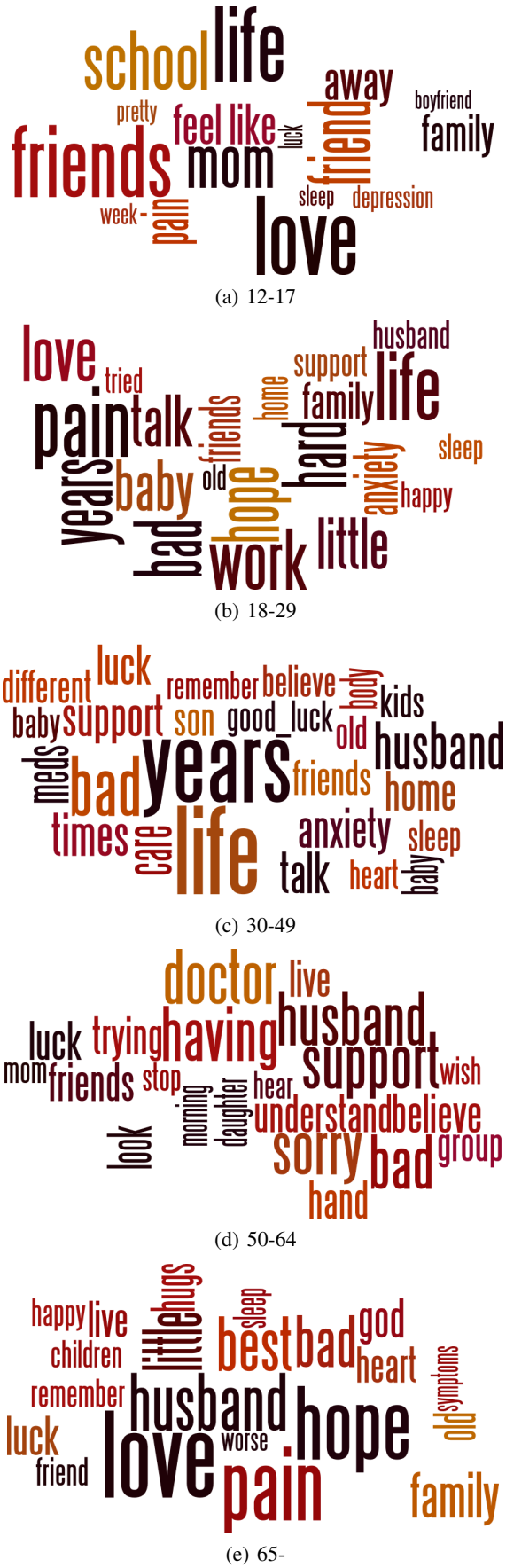
(a) Female

(b) Male

Figure 5: Word clouds of males and females

fertility, and *c-section*. There are also familial tokens such as husband, hubby and baby's that are used by females. The males on the other hand, talk mostly about sex, food, drugs, finance and video games. There is also a familial token, *wife's* that was also found to be a useful feature by our system.

For extracting keywords according to age groups, we treated all of the posts falling under a certain age group as a single document and calculated the tf-idf scores for word unigrams,



(a) 12-17

(b) 18-29

(c) 30-49

(d) 50-64

(e) 65-

Figure 6: Word clouds of various age groups

Condition Itemsets	Support
{Smoking Addiction,Smoking Addiction Recovery}	1704
{Smoking Addiction,Smoking Addiction Recovery,Anxiety}	167
{Smoking Addiction,Smoking Addiction Recovery,Anxiety,Panic Attacks}	32
{Anxiety,Self Injury}	477
{Anxiety,Self Injury,Panic Attacks}	87
{Anxiety,Obsessive Compulsive Disorder}	416
{Anxiety,Migraine,Migraine Headaches}	91
{Depression-Teen,Self Injury}	634
{Depression-Teen,Self Injury,Gay-Lesbian Teens}	39
{Coming Out,Self Injury,Gay-Lesbian Teens}	43
{Sexual Abuse,Self Injury}	615
{Sexual Abuse,Rape,Self Injury}	137
{Sexual Abuse,Post Traumatic Stress Disorder (PTSD),Anxiety}	83
{Physical Emotional Abuse,PTSD,Anxiety}	81
{Sexual Abuse,PTSD,Self Injury}	80
{Sexual Abuse,PTSD,Self Injury,Rape}	33
{Codependency,Healthy Relationships}	185
{Diets-Weight Maintenance,Obesity}	465
{Diets-Weight Maintenance,Anxiety}	196
{Diets-Weight Maintenance,Food Addiction}	177

Table 5: Frequent support group combinations where the same users are active

bigrams and trigrams occurring in the posts after filtering out the stopwords. We used these tf-idf values to generate the word clouds as shown in Figure 6. The word clouds for age groups are not as distinct as those for gender. Nonetheless, there are certain words that give away the age group of a person. People in the age group 12-17 talk about school, friends, love and life. There is overlap between the words used by people in 18-29, 30-49 and 50-64 age groups. People in the age group 18-29 talk about work, baby, husband, support and family. People falling in the age group 30-49 also talk about baby, husband and also about kids, home and life. In the age group 50-64, people talk about familial terms such as mom, husband and daughter. In the age group 65-, people talk about pain, hope, children, husband and luck among other things. The word cloud also contains a keyword old for people in this age group. Husband occurs so frequently in the word clouds because as we have mentioned before, our data set contains much more females than males. A keyword that appears in all of the word clouds is friend/friends, since it remains a constant at every age.

5.3. Frequent Support Group Combinations

In our dataset 27.60% of the users post under more than one support group. It would be interesting to find out which conditions afflict the users at the same time. We performed frequent item set mining of the support groups. An itemset consists of all of the support groups where a user is active

i.e. has a post or a reply. We looked at the frequent itemsets with support ≥ 30 and found 842 such frequent itemsets having 2-4 items. We have presented some of them in Table 5. We ignored the sets that are obvious, such as {Widows-Widowers, Bereavement}, {Miscarriage, Pregnancy after loss}, {Miscarriage, Pregnancy}, {Miscarriage, Pregnancy after loss, Pregnancy, Trying to conceive}, etc. The itemset with the highest support is {Smoking addiction, Smoking addiction recovery} with 1,704 users active in both conditions. This may be an indication that people with adiction acknowledging their condition are actively trying to recover from it. Panic attacks and Anxiety also make up another frequent itemset with the two related to smoking.

Anxiety is present in 16% of the frequent itemsets as it is also the support group having the highest number of unique active users. Anxiety is seen with everything from *Self Injury* and *Panic Attacks* to *Migraine* and *Diets-Weight Maintenance*. Since most of the users of DailyStrength are suffering from some condition, naturally they are likely to be anxious. There was also a lot of users who were going through *Self Injury* along with *Sexual Abuse*. *Self Injury* was also predominant for users who also posted in *Gay-Lesbian Teens*. Similarly, users worried about *Obesity* and *Food Addiction* were likely to also post on *Diets-Weight Maintenance*.

5.4. Health Forum Impact

People join health forums in order to find support to deal with their condition. We performed a small analysis to figure out if users start out pessimistic and then go on to become optimistic or if the opposite happens and they go from optimistic to pessimistic. We took the first three and last three posts of all users having more than 25 posts in total. We obtained the polarities of these six posts by using the Python TextBlob library (Steven Loria, 2016). We show the histogram of the distribution of these polarities for the first three and last three posts in Figure 4. The majority of the posts fall in the neutral range i.e. within [-0.1,0.1). In the negative polarity area, the bars are slightly higher for the counts of the first three posts. Similarly, in the positive polarity region, the bars are slightly higher for the counts of last three posts. This might indicate that the posts written by users when they first join the forum are likely to be more negative than those written later on. In fact, 76.38% of the users whose first three posts had negative polarity have positive polarity in their last three posts. On the other hand, only 17.15% of the users who were positive in their first posts become negative in their last posts. This means that only around 23% of the users who are active in the forum stay negative. This is a very simple analysis and we will need to perform more in-depth investigation but it does seem to suggest that user involvement in support groups has a positive effect in their mood since they write in a more positive tone.

6. Conclusion and Future Work

Here we have presented a large corpus of data collected from the DailyStrength forum containing users' ages, genders, and original posts and replies written on various forums within DailyStrength. To the best of our knowledge, this is the first benchmark dataset for author profiling on health forums. The dataset contains posts from more than 80,000 users covering age ranges from 12 to 100. The dataset is very large and can also be used for author profiling on large datasets and on online datasets. We also present a method to predict the age and gender profile of a forum user given their posts and activities in the forum. We were able to obtain accuracy well above baseline in all cases.

In the future we will explore adding features such as user engagement as a function of length and frequency of original posts and replies. We can also experiment with different divisions of age groups. In our dataset, there are a lot of users who have been active for a lot of years. We have enough data to analyze how an author's outlook and personality changes when they actively participate in the forums for a long time. Another analysis we can perform on the same dataset is to see how a user's writing style changes as they spend more time on the forums. There are many other health forums similar to DailyStrength and in the future we plan to test how our findings will hold on other health forums. Author profiling for health forums is an important task and although our method is a good predictor of users' profiles, with this dataset our hope is that there will be more work on this to follow.

Acknowledgments

This project was partially supported by NSF award No. 1462141.

7. Bibliographical References

Bell, R. A., Hu, X., Orrange, S. E., and Kravitz, R. L. (2011). Lingering questions and doubts: Online information-seeking of support forum members following their medical visits. *Patient Education and Counseling*, 85(3):525 – 528.

Culver, J., Gerr, F., and Frumkin, H. (1997). Medical information on the internet: a study of an electronic bulletin board. *Journal of general internal medicine*, 12(8):466–470.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7:1–30, December.

Estival, D., Gaustad, T., Pham, S. B., Radford, W., and Hutchinson, B. (2007). Author profiling for english emails. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, pages 263–272.

Jha, M. and Elhadad, N. (2010). Cancer stage prediction based on patient online discourse. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, pages 64–71, Uppsala, Sweden, July. Association for Computational Linguistics.

Liu, W. and Ruths, D. (2013). What's in a name? Using first names as features for gender inference in Twitter. In *2013 AAAI Spring Symposium Series*.

Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., and Inches, G. (2013). Overview of the author profiling task at PAN 2013. In *Notebook Papers of CLEF 2013 LABs and Workshops, CLEF-2013, Valencia, Spain, September*, pages 23–26.

Rangel, F., Rosso, P., Chugur, I., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., and Daelemans, W. (2014). Overview of the 2nd author profiling task at PAN 2014. In *Proceedings of the Conference and Labs of the Evaluation Forum (Working Notes)*.

Rolia, J., Yao, W., Basu, S., Lee, W.-N., Singhal, S., Kumar, A., and Sabbella, S. (2013). Tell me what I don't know–making the most of social health forums. In *Healthcare Informatics (ICHI), 2013 IEEE International Conference on*, pages 447–454, Sept.

Schler, J., Koppel, M., Argamon, S., and Pennebaker, J. W. (2006). Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, volume 6, pages 199–205.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E. P., and Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE*, 8(9):e73791, 09.

Steven Loria. (2016). TextBlob: Simplified Text Processing. `https://textblob.readthedocs.org/en/dev/1`. [Online; accessed Mar 07, 2016].