Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 144 (2018) 96–104

Procedia Computer Science

www.elsevier.com/locate/procedia

INNS Conference on Big Data and Deep Learning 2018

Demographics Analysis of Twitter Users who Tweeted on Psychological Articles and Tweets Analysis

Sajeev Udayakumar[a,*], Damith Chamalke Senadeera[a], Selvaraj Yamunarani[a], Na Jin Cheon[a]

[a]*Nanyang Technological University, 50 Nanyang Avenue, 639798, Singapore*

Abstract

This project is focused on exploring the appealing trends in the profiles of different users, with respect to different aspects of the user demographics which includes gender, whether the user is an individual or an organization, psychological and their academic background, who had discussed psychological articles on Twitter. To perform this task we retrieved the details of psychological articles, related tweets and twitter user details by web scraping using python. Then to assign suitable labels, we have used the Rapid Miner to create a model using training data. After the labeling process, we have analyzed the patterns in user profiles. Further, the tweet contents were analyzed focusing on the psychological topics more discussed by the users. This analysis will give the insight to psychological researchers, what topics are being discussed by the Twitter users and what are trending topics among Academic communities and Nonacademic Communities.

Keywords: Psychology; Twitter; Python; Web scraping; Rapid Miner; Text mining; Support Vector Machine

1. Introduction

The introduction of social media is one of the significant events in this modern era to create a huge impact on our lives. These social networks are very widespread in this decade since this internet enabled services such as smart phones, tablets, and personal computers. Nowadays users use social media to share information and thoughts with each other. This paves way for a large amount of data to be shared on social media. In short communication content such as tweets was important to study as it was believed to have an effect on the society. According to Riff, Lacy, and Fico, social affiliations such as family and community involvement were important predictors of peoples attitudes and behaviors, and a network of personal influence was identified as key factors influencing their decisions.[1, 2]

In the academic community, some individuals use these tweets as a medium to announce their publications of new articles and to cite those articles in the tweets. Due to the big data associated with these social media, its not

[*] Corresponding author. Tel.: +6581155139.
E-mail address: uksajeev@gmail.com

10.1016/j.procs.2018.10.509

possible for us to apply similar approaches used in traditional ways available in research evaluation to analyze about the impacts of research articles in the society. Hence, tools such as Altmetrics play an important role in analyzing research impact based on social media activities.[3] Altmetrics measures the impact of the social and academic use of research in different fields. It includes academic comments, citation and reuse and social media components such as tweets, shares, as well as ratings. Further, this is defined broadly and offers a variety of avenues to get a complete idea of measuring the research impact. The metrics of research aim to quantify the significance of research such as how many highly cited papers an author has produced, how many publications of articles in heavily cited journals and how many times the article has been cited. Metrics are an attempt to provide researchers a direct response to the value of work. Psychology field is selected for this project as this field was having a higher citation rate and one of the top Altmetrics coverage when compared with other disciplines.[4]

The collected information from Altmetric website was utilized to collect data from Twitter. These data, in turn, was again used to recognize the users demographic information such as gender, whether the user is an individual or an organization and to determine the characteristics whether the user is from the psychological background or not and from the academic or non-academic background. The afterward further critical analysis was performed to identify the interesting terms of tweets based on different demographic groups. The previous work has been focused on finding the motivations for tweeting research articles and they have analyzed the trends about different demographic groups and how actively they are engaging in discussions on academic articles on tweets.[5]

Our work emphasizes on finding an overview of psychology related problems or topics which are discussed by Twitter users, using analysis of Twitter user profiles and frequent terms encountered in tweets. This will be beneficial for understanding the insight of social media users (specifically Twitter users) towards psychology and related issues by different demographic users around the globe. In this paper we have discussed the previous work in the literature review section, then we have described how the data was scraped from the web and then how the models were created using Rapid Miner tool. Next section describes the analysis of the of twitter use profile findings and finally, we have included the conclusion of the paper which also describes the limitations and future works.

2. Literature Review

Mainly machine learning algorithms are used to detect the hidden knowledge on the content. It is nothing but extracting the interesting features from the text. Gender can be predicted based on persons profiles, patterns, messages, and communication channel. The investigator examined the gender prediction based on text, forums, searching terms, emails, reviews, and comments using machine learning algorithms.[5] The user profile, tweeting behavior, linguistic style of user message and users social network had been used to detect the gender.[6] The gender had been detected using profile names and the first and last names. Also, they could detect based on the messages shared among them. The search queries and terms also used to predict the gender.

Researchers could detect the terms, which is used by female and male to classify. As well as forum discussions and blog discussions had been used to detect the gender. Males had unique separation from females vice-versa in terms of view and communication. Reviews and comments also had been used to predict the gender. Females thinking are differed from males thinking. They had some distinctive view from the female vice-versa. This twitter description and social medias short description is very well applied for predicting the gender and ethnicity. The web links connected with other sites also used as gender prediction.[6] That linked blogs and websites description provided the details about the author. From this demographic details could be determined easily. Users linguistic style and lexical usage were also used to predict the gender because males and females linguistic style and lexical usage differed from each other.[5] In this project, we have used the approach of detecting the gender using the classifiers available in RapidMiner.

In previous studies, a Twitter user is categorized as psychology field user if there are keywords such as “psych”, “autism”, “autistic”, “studying minds” or “mental” appears in his/her description. Also, Universities, Research Institutes, Publishers, Researcher Networking Groups, Research Teams/Projects and Libraries are the six labels given to the academic organizations.Similarly, there are eleven categories of non-academics organization. Namely commercial organizations, interest and reference forums, Non-Governmental Organization (NGO) and Non-Profit Organization (NPO), networking groups, social campaign forums, hospitals and clinics, product pages,public figure pages, church and religious organizations, event pages and museums. When it comes to academic individuals they have used researchers, faculty members, postgraduate students, undergraduate students and librarians as six categories. This classification was based on the terms that appear in their descriptions. Non-academic Individuals had only 2 categories under this label, which are individuals and psychologists. This classification is also based on the description of individuals.[7, 8]

3. Data Collection Through Web Scraping

For the purpose of this study, articles in the English language from 587 psychological journals were collected from Web of Science (WoS) citation indexing service. First, a query was formulated using a python script to be used in advanced search option of the web of science website in order to retrieve English language articles published in the given set of psychology related journals for years 2014 and 2015. As a result of this query, we retrieved 67064 such articles.Afterwards, we filtered out the articles out of this set which incorporated a Digital Object Identifier (DOI), which is a unique alphanumeric string assigned by a registration agency (the International DOI Foundation) to identify content and provide a persistent link to its location on the Internet. The publisher assigns a DOI when your article is published and made available electronically) along with the subcategories for the psychology discipline in the Thomson Reuter Social Science Citation Index using a RapidMiner process. After this filtering, we obtained 64934 papers with DOI.[9]

Then the DOI of each paper in the filtered data set was used to retrieve the Altmetric id for each paper in order to collect tweets related to each paper using the Altmetric API embedded in a python script. Then we again filtered out the data set, to obtain papers which are listed in Altmetrics.com (that is which have an Altmetric id) using a RapidMiner process and this resulted in us with 42891 papers. After that using the Altmetric ids retrieved, almetric.com's online data explorer (i.e. Altmetrics Explorer) was scrapped using a web scraping script in python to obtain the tweets related to each of these papers represented by an Altmetric id. Mainly the fields 'datetime', 'twitter user page', 'twitter user name', 'tweet content' and 'tweet post url' were retrieved from Altmetrics Explorer in order to use for our analysis. Using this script, 252310 tweets related to these papers were scraped from altmetrics.com. After collecting this data set using a RapidMiner process we were able to find out that out of the 42891 papers with Altmetric ids only 37627 papers had tweets related to them.

3.1. User Profile Manually Labeling of Training Data set

For training of the models, we selected 3000 random records of the English Twitter user profiles and manually labeled them by identifying the labels Academic/Non-Academic, Individual/Organization and Psychology/Non-psychology. These labels were based on the description of the user. We followed some guidelines to label the data, which are described below,

Individual : If the tweets represent personality as well as with profile name and are identified by keywords like I or me are classified as the individual.

Organization : A Twitter user is classified into Organization if the description represents association, institute, colleges, communities and are identified by keywords like Institute, website, Academy, We, blog,forum,centre, Follow us and so on.

Academic : A Twitter user is classified as Academic if the user has an academic background like professors, faculty members, student, researchers, teachers, librarians, professional networking groups, and scientists. The keywords used for identification are professor, postdoc, scientist and so on.

Non-Academic : A Twitter user is classified as Non-Academic if there is no academic related description or background.

Psychology : A Twitter user is classified into Psychology if the tweet content has autism, autistic, mental, psych, studying minds, neuroscience and some other jargon related to the field of psychology.

Non-Psychology : A tweeter can be classified into Non-Psychology if the description has no psychology background.

3.2. Inter-Coder Reliability

Two coders initially performed the coding for the whole training data set and the inter coder reliability measures such as percentage agreement rate along with Cohens Kappa values were calculated for each label. Afterwards for the

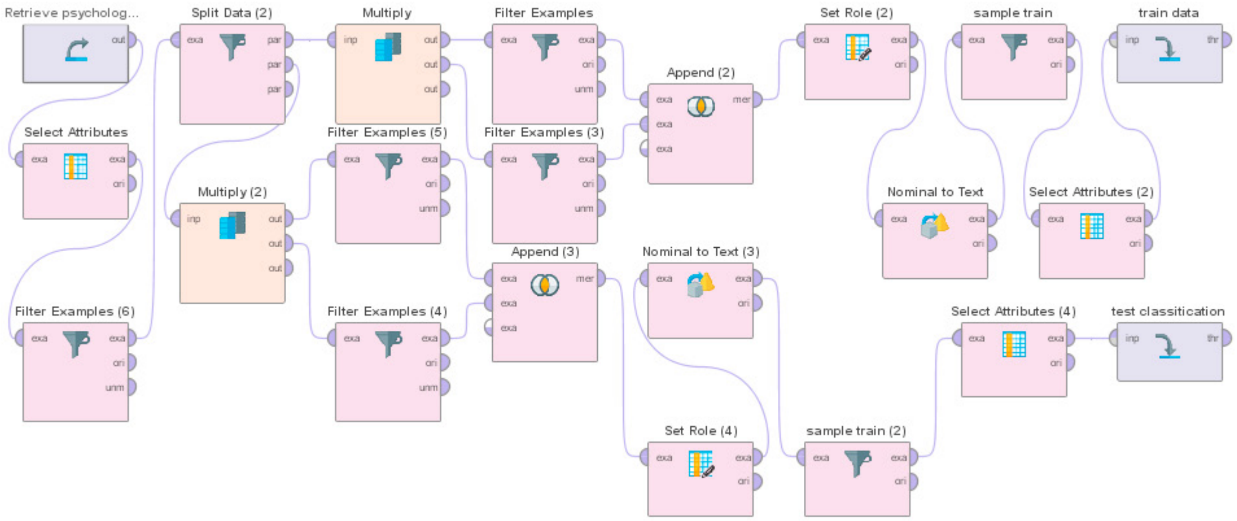

Fig. 1. Training and Test Data set Preparation

cases where the labeling was not matched between the two coders, after discussions, the final label was assigned. The percentage agreement rate along with Cohens Kappa values obtained for each label for this set of 3000 records are given below. In general, for all the categories percentage agreement rate was above 89% and Cohens Kappa value was above 76%.

Table 1. Summary for Analysis of Inter-Coder Reliability

Category	Percentage Agreement	Kappa
Academic/Non-Academic	91.50%	77.90%
Psychology/Non-Psychology	92.60%	76.40%
Individual/Organization	89.90%	78.50%

4. Model Creation for Labeling

4.1. Creating Training and Test Data sets

For the purpose to separate training and test data sets we have created a Model using RapidMiner to randomly select the data points for Training and Testing where 80% of Data taken for Training and rest of 20% taken for Testing. When selecting data points for the training process, we tried to keep it balanced with both labels as much as possible. This will enhance the overall performance of the model created as shown in Figure 1. We created another program to do the processing and preparation of the remaining full data set compromising of the full data set. There we will have an additional column to include the Academic or Non-Academic column. These data sets will be used in the model creation process. This whole process was replicated for the other processes on Psychology or Non-Psychology field and Individual and Organization field.

4.2. Creation of Model using Training Data

Figure 2 shows the generalized our main process model that will process data sets from stores created after separating them to the training set and test set. In support of our training data set named with Retrieve train store, we used diverse operators applied in the sub-process of the operator *Process Training Documents* such as *Tokenize*, *Transform*

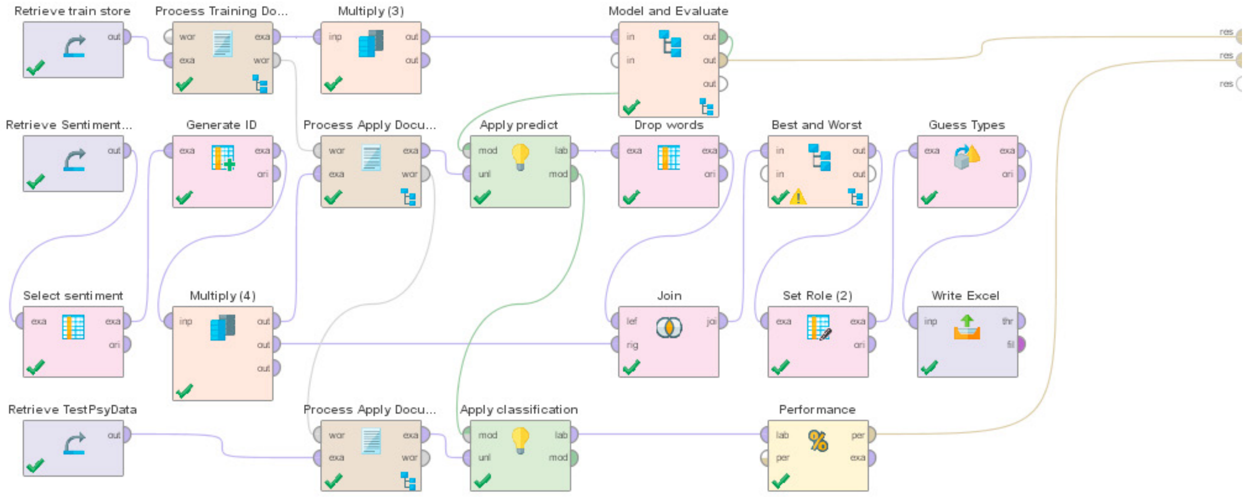

Fig. 2. Model Creation and Data set Labeling

cases, *Filter Stopwords* and *Generate n-Grams* to improve the accuracy of our process. Classification had to remove it as it gave low accuracy when using that operator. This may be due to word "I". Because this is one of the strong words helps to identify the individuals. When Filter Stopwords used this particular word may be removed. Then we proceed to use the Multiply operator, thus we can discover the weight of the words used to recognize Academic and Non-Academic cases and measuring the training set overall accuracy when trying out dissimilar classifiers such as *Support Vector Machine (SVM)*, *Random Forest* and *Naive-Bayes* using the same input data.

4.3. Model Creation for Gender Labeling

In this process, we used the Extract Gender Operator and in the preprocessing of the data set, we need to prepare user details with first name and last name. It is optional to use the country option available in this operator. So we had to prepare the data according to this process. In this preparation, we had to remove some special symbols added by the users in the middle of the name.

4.4. Results of the Training and Test Data sets

During the model creation process of Academic/Nonacademic, we tried to incorporate different classifiers like Na ve-Bayes, Random Forest, k-Nearest Neighbour(k-NN), and SVM. The accuracy values we got for both Training data set cross-validation and the test data set is tabulated in the Table 2.

Table 2. Summary for Accuracy of Different Classifiers

Classifier	Training Dataset Accuracy	Test Dataset Accuracy
Naive-Bayes	71.65% ± 3.70%	72.59%
Random Forest	72.05% ± 0.15%	64.24%
k-NN (k = 1000)	71.05% ± 0.25%	65.29%
Support Vector Machine (SVM)	87.50% ± 1.64%	86.80%

From the observations, we can see that SVM gives more accurate prediction when compared to other classifiers. So we finalized to continue with that operator. As our classifications are binary classifications, either Academic or Non-Academic SVM gave better results compared to other classifiers. Given the set of input and label, SVM is aiming to divide and mark each training example (e.g. record) into either of the classes so that each of the training examples is divided by a clear gap that is as wide as possible. The following figure depicts the high level of SVM classifier that

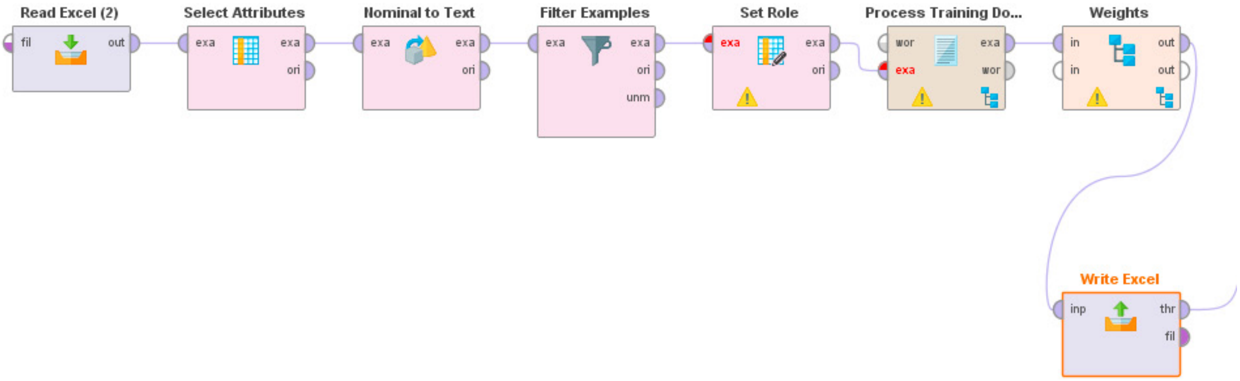

Fig. 3. Model of Overall Process of Tweet Content Analysis

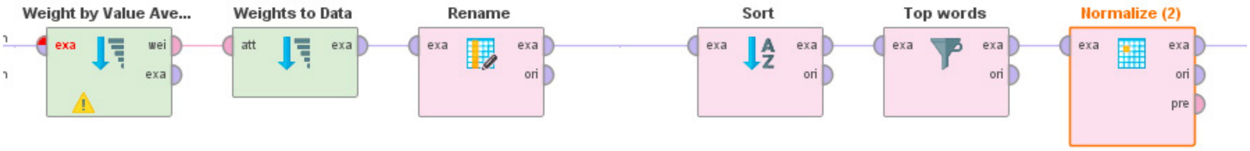

Fig. 4. Part of Process related to Manipulation of Weights

divides the examples into two groups.[10] The following table describes the accuracy and evaluation measure values for the different labeling we did in the project.[11]

Table 3. Summary of Evaluation Measures of three Models

Labeling	Training Data set Accuracy	Test Data set Accuracy	Precision	Recall	F- Measure
Academic/Non-Academic	87.50% ± 1.64%	86.90%	86.58%	90.08%	88.27%
Psychology/Non-Psychology	90.85% ± 1.29%	88.73%	91.19%	98.20%	94.56%
Individual/Organization	84.12% ± 1.31%	78.50%	86.47%	92.37%	89.32%

4.5. Model creation for Tweet Content Analysis

Tweet content corresponding to the filtered-out twitter users was analyzed in order to find the most important words based on the given labels Academic/Non-Academic, Psychology/Non-Psychology, Individual/Organization and Male/Female (for individuals). First, the tweet contents are filtered out according to the attribute label (that is either Academic/Non-Academic, Psychology/Non-Psychology, etc.) and then the preprocessing of tweet content is done. Then all the words are transformed into lower case and the stop-words are removed from the word list as stop-words will be more frequent in any text data set. Afterwards, all the words (tokens) produced which occur more than 0.01% of total records are sent to assign weights in order to rank the words according to their importance. This overall RapidMiner process is shown in Figure 3.

Then weight assignment for each word is done using the operator Weight by Value Average, where it will calculate feature weights to characterize the given class using the input document corpus. That is characteristic features will receive a higher weight than fewer characteristic features. The weight of a feature is determined by calculating the average value of this feature for all examples of the target class. Ultimately weights are ordered according to the descending order and the top 50 words are being extracted for analysis. The RapidMiner process for this weight assignment is depicted in Figure 4.

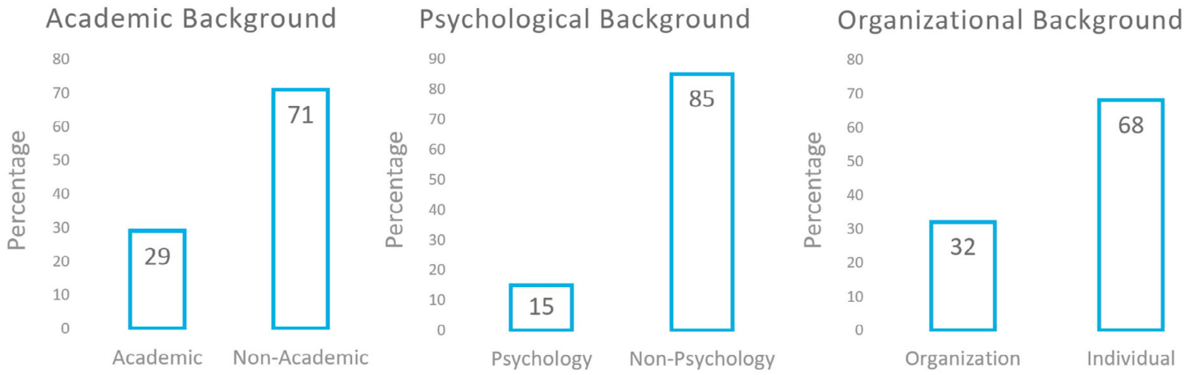

Fig. 5. Percentage graphs on Demographics of all users

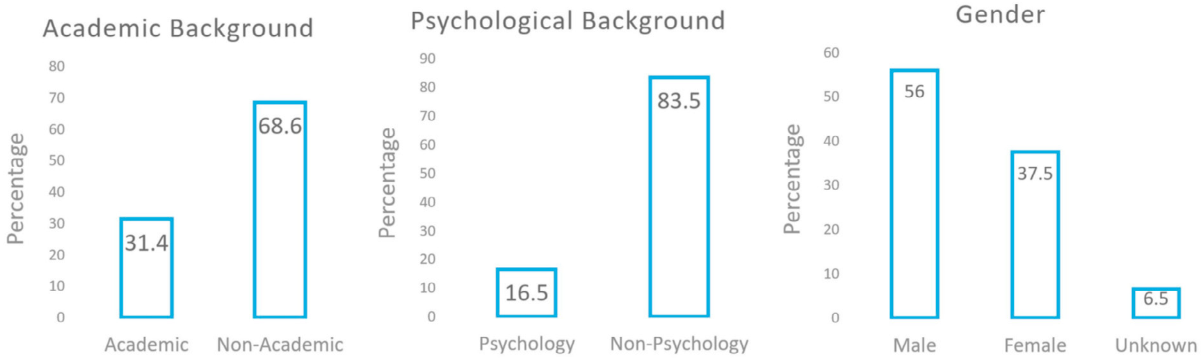

Fig. 6. Percentage graphs on Demographics of Individual users

5. Analysis of User Profiles and Tweets

5.1. Analysis of Labeled Data

Based on the labeling performed using the models created using RapidMiner,we are presenting the following analysis on the different demographics of Twitter users. The charts in the following figure explains the percentage values of users based on three main demographics. Furthermore, the next figure describes how the demographics differ among the individual users.

5.2. Analysis of Tweet Contents

During this tweet analysis, we have included 95088 numbers of tweets. In the Table 4, we have summarized the few of the topmost psychological topics in the decreasing order of their weight values tweeted by users on psychological articles collected between 2014 and 2015. We can see that most of the topics are commonly discussed by all categories. For examples topics like autism, children, health, depression, anxiety and cognitive are discussed by users belong to all category. But significantly we can see that IJED which is one of the journals based on Eating Disorder was more tweeted by Academic users. And a word like Spectrum is commonly used by users having the Psychological background. Even though most of the words are related to psychology, there were words like risk and HIV which are not directly connected to psychology too mentioned in the tweets.

By analyzing the topics discussed by male and female users we could find more common topics, but the weights computed showed some interesting findings. For example, the terms social, children, autism, women, anxiety, depression, workers were more used by females than males. Even though first fifty words had nearly similar words shuffled between two lists, there were some distinct terms more commonly used by female users as well as male users. For examples words like emotional, Facebook, relationship, positive, adolescents and life are more used by females, meanwhile, words like control, science, review, personality, and group are used by males. These observations tell us that more female users are influenced by social networks and male users are more concern about their personalities.

Table 4. Frequent Terms on Tweet Contents of Different Demographic Groups

Demographic Category	Topmost mentioned terms and topics
Academic Usersc	Social, Children, Autism, Health, Eating, Women, Behavior, Cognitive, Work, IJED (International Journal of Eating Disorders), Anxiety, Memory, Depression, Adolescents, Sexual
Nonacademic Users	Social, Autism, Children, Health, Sexual, Cognitive, Risk, Women, People, Behavior, Anxiety, Disorder, Men, Depression, Young, Treatment, HIV
Psychology Users	Autism, Social, Cognitive, Children, Anxiety, Disorder, Affective, Neuroscience, Memory, Spectrum, Depression, Risk, Adults, Behavior, Brain, Treatment
Non-Psychological Users	Social, Children, Sexual, Health, Women, Risk, Behavior, Autism, Men, Eating, Cognitive, HIV, Disorder, Young, Anxiety, Adolescents, Mental, Adults, Relationship, Depression
Organizational Users	Autism, Social, Children, Health, Sexual, Cognitive, Risk, Disorder, Behavior, Anxiety, Eating, Women, HIV, Adolescent, Adults, Young, Treatment, Depression, Parenting
Individual Users	Children, People, Autism, Health, Women, Cognitive, Behavior, Risk, Sexual, Men, Memory, Learning, Anxiety, Performance, Brain, Depression, Work, Mental, Disorder, Personality, Relationship

Table 5. Frequent Terms of Tweet Contents of Individual Males and Females

Demographic Category	Topmost mentioned terms and topics
Individual Male Users	Social, People, Children, Health, Autism, Women, Cognitive, Behavior, Risk, Sexual, Men, Memory, Learning, Performance, Meta, Brain, Anxiety, Mental, Work, Depression, Control, Science, Review, Personality, Group
Individual Female Users	Social, Children, Autism, Health, People, Women, Risk, Behavior, Cognitive, Anxiety, Men, Adults, Disorder, Sexual, Memory, Depression, Work, Stress, Language, Brain, Time, Mental, Learning, Relationship, Positive, Facebook, Emotional, Adolescents, Life

6. Conclusion

During the process of web scraping for Almetric ids, tweet contents, and Twitter user information one of the main hurdles encountered is that collecting bulk data from web consumes a lot of time. So we had to collect data in different batches. Another difficulty we encountered is that for some APIs in collecting web data there are rate limits. Therefore we had to interact with those APIs in such a way that we do not exceed the given rate limits by introducing time lapses between our API calls. In the overall web, scrapping process was somewhat complex due to these constraints.

During the labeling of the Training Data set, we faced lots of trouble in identifying some of the jargon which is related to psychology where we had no previous knowledge about. So we had to check for many definitions in web or dictionaries to complete the labeling. While assigning gender labels we had to do preprocessing of names of all the users to split them to first name and last name. Further, there were lots of names had special symbols like stars, and then we had to remove them. In the process modeling, we had to incorporate different preprocessing components in RapidMiner. In some instances, we had to communicate with RapidMiner community to find ways to improve our systems. Also through our analysis, we could clearly see that to perform better text classifications SVM is the best model that gives more accuracy when comparing with other classifiers. Also, we found that preprocessing operators work may cause a reduction in accuracy values, for example in the classification of Individual vs Organization, when using the StopWords Operator we got lower accuracy, we removed it and continued the analysis.

In the next stage after assigning labels, we did a quantitative analysis of each category and those analyses showed interesting patterns like Males are more interested than Female Twitter users on psychological articles and research. When it comes to tweet analysis we found some significant patterns like more females tweeting about Facebook and relationship than males related to psychology. These analyses would help researchers on psychology to focus on their future results as well as how to get to more population of users in continents like Asia to spread the knowledge on Psychology.

There are a number of limitations that we encountered during this project. One of the main limitations is the lack of knowledge in psychology field when manually labeling data. If we could have obtained the help of an expert in psychology field during our labeling, we could have improved the accuracy of our classification. During manual labelling for certain words we did web search to check whether that word is related to psychology. Another limitation of this project is that due to technical limitations in data collection and processing, data corresponding only for 2014 and 2015 was analyzed. Further content analysis of the relevant tweets can be performed based on other labels as well in order to find interesting patterns. In addition, the analysis was carried out according to the data extracted from the Twitter user profiles only, which may not depict the real-life profiles of the Twitter users in a completely accurate manner.

For future works, we can collect more data not limiting two years 2014 and 2015 and analyzed to obtain further insights. Also, for the predicting the labels different other classifying algorithms can be tested and more feature engineering can be performed in order to improve accuracy. Further results obtained from analysis of these psychology related articles can be compared with results obtained by analysis of articles from various other disciplines.

References

1. Abernethy, A.M.. Daniel riffe, stephen lacy, frederick g. fico, analyzing media messages: Using quantitative content analysis in research, lawrence erlbaum associates: Mahwah, new jersey, 1998. 2000.
2. Culotta, A., Ravi, N.K., Cutler, J.. Predicting the demographics of twitter users from website traffic data. In: *29th AAAI Conference on Artificial Intelligence, AAAI 2015 and the 27th Innovative Applications of Artificial Intelligence Conference, IAAI 2015*. AI Access Foundation; 2015, .
3. Roemer, R.C., Borchardt, R.. From bibliometrics to altmetrics: A changing scholarly landscape. *College & Research Libraries News* 2012; **73**(10):596–600.
4. Sadah, S.A., Shahbazi, M., Wiley, M.T., Hristidis, V.. Demographic-based content analysis of web-based health-related social media. *Journal of medical Internet research* 2016;**18**(6).
5. Pennacchiotti, M., Popescu, A.M.. A machine learning approach to twitter user classification. *Icwsm* 2011;**11**(1):281–288.
6. Deitrick, W., Miller, Z., Valyou, B., Dickinson, B., Munson, T., Hu, W.. Gender identification on twitter using the modified balanced winnow. *Communications and network* 2012;**4**(3):189–195.
7. Mislove, A., Lehmann, S., Ahn, Y.Y., Onnela, J.P., Rosenquist, J.N.. Understanding the demographics of twitter users. *ICWSM* 2011; **11**(5th):25.
8. Na, J.C.. User motivations for tweeting research articles: A content analysis approach. In: *International Conference on Asian Digital Libraries*. Springer; 2015, p. 197–208.
9. Lui, M., Baldwin, T.. langid. py: An off-the-shelf language identification tool. In: *Proceedings of the ACL 2012 system demonstrations*. Association for Computational Linguistics; 2012, p. 25–30.
10. Suykens, J.A., Vandewalle, J.. Least squares support vector machine classifiers. *Neural processing letters* 1999;**9**(3):293–300.
11. Joachims, T.. Text categorization with support vector machines: Learning with many relevant features. In: *European conference on machine learning*. Springer; 1998, p. 137–142.