AGU International Journal of Engineering & Technology http://www.aguijet.com

(AGUIJET) 2018, Vol. No. 6, Jan-Jun **e-ISSN: 2455-0442, p-ISSN: 2455-6734**

"INNOVATING A DISTANCE SUPERVISION METHOD TO MODEL A USER DEMOGRAPHICS PREDICTION TO EVALUATE TEXTUAL AND SOCIO CONTEXT OF USERS"

Aayush Goel
Department of Polymer Science Chemical Technology
***Delhi Technological University** (Formerly Delhi College of Engineering)*

ABSTRACT

A great platform has been provided by the penetration and growth of social networks, in the way that it has been made possible to present different services, impacting reach of these services. Therefore, it has become imperative to understand the demographics of social network users in order to ascertain user behavior and devise strategies positively affecting determining business profit.

In this paper, I propose an algorithm that fits a regression model to predict user demographics using both textual content and social network evidence of Twitter clients crosswise over seven unique parameters and applies the model to arrange singular clients based on ethnicity, sexual orientation and political inclination. A remotely named dataset made by gathering group of onlookers estimation information for sites is used in the way that web activity demographic information from Quantcast is brought. Web activity socioeconomics of a website with the devotees of that webpage on Twitter are matched to fit a relapse show between an arrangement of Twitter clients and their normal statistic profile.

The proposed method is evaluated on the basis of performance metrics for optimal inference of demographics basis textual content and network influence of users. The algorithm employs a distantly supervised learning approach and the results obtained are compared to [1]. In the proposed approach, a distantly labeled dataset created by collecting audience measurement data from the website of a popular audience measurement company is utilized. In order to measure the effectiveness of the algorithm, accuracy evaluation on the basis of both linguistic features and social network features was employed.

*The proposed method which utilized a distantly trained regression model provided classification accuracy that was competitive with a fully-supervised approach. I was able to thus establish that **an algorithm based on a distantly trained regression model can provide performance comparable to a fully supervised approach when evaluated on the basis of both linguistic features and social network features of certain users in real-life social context.***

1. INTRODUCTION

There is certain coherence in the nature of interaction among members of social network that is evident intrinsically in the aforementioned network. The uniformity in patterns of self- organization increases with the increase in network size. Social network analysis techniques have been employed over the past to ascertain the structure of relationships between social entities. The varieties of relationships represented by social networks include friendship, kinship and ties among the participants.

Different fields of study utilize the information obtained from social networks by applying certain analytics. Demographics and other social media patterns play a major role in such analytics. User demographic patterns of suitable interest include location, gender, education, relationship status, political bends etc. Understanding such patterns and arriving at a right conclusion is critical to further progress in this area of research. Sampling of social media users permits researchers to overcome the constraint of considerable bias in selection from the uncontrolled data. In addition, public messaging campaigns help in reaching the right set of target demographics.

Social networks have become a great means of revolution in terms of business penetration, growth and profits by providing different services, and impacting reach of each of these services. Therefore, it is very important to explore the demographic composition of a sample set of social network users in order to better understand user behavior and patterns of interaction amongst a set of users.

An important characteristic of social networks is their dynamism given that there is an exponential increase in information being shared over them every minute, wherein such information is also subject to change over a period of time (for instance, age, income, relationship status) while some of it may remain consistent (for instance, ethnicity, gender). Some such demographic parameters may also coherently observe an upward trend, for instance, age, education, parental status.

The real-life, expansive social networks that form a hub of the users' vital particulars and relationships among them in a network can be used to explore the demographics of participants in such social networks characterized by voluminous information that spans a variety of factors primarily backed by interaction that is dynamically expanding by the minute while adjudging a pattern that prevails in user behavior in accordance with the fact that a high number of relations with quality interactions exist. Such social networks typically contain information about users, events, and relationships between them, which can in turn be utilized to predict user attributes given the network influence and linguistic features of the content created by them.

As far as social networks are concerned, supervised classification aims to procure a training set of annotated users to fit a model to anticipate user traits. In view of user demographics, human annotations may prove to be error-prone, while many a times, it becomes difficult to adjudge demographics of interest as there may be limitations imposed on the generalizability of classification. Moreover, network evidence must be considered as important as textual evidence to accomplish the task.

In this project, I propose an algorithm that fits a regression model to predict user demographics using both textual content and social network evidence of Twitter users across seven different parameters and applies the model to classify individual users on the basis of ethnicity, gender and political preference. My proposed algorithm aims to consider the linguistic features as well as social influence of participants in such a network.

The proposed method aims to utilize web traffic data as a form of weak supervision and use followership information as the primary source of evidence. This algorithm is able to utilize a distantly trained regression model providing classification accuracy that was competitive with a fully-supervised approach.

The main contributions of this project can be listed as follows:

- Textual content as well as social context of users is considered for representing the given datum in a better structured manner, wherein social interactions are measured by seven different demographic parameters that span across a variety of prevalent factors like gender, age, income, education, ethnicity, parental status, and political preference. These influencing factors in a social network are used to explore the demographic constitution of a sample media users. Web traffic data and followership information are utilized as a form of weak supervision and the primary source of evidence respectively. The analysis includes how accuracy varies with respect to the number of linguistic terms collected per user.
- The algorithm is applied on a distantly labeled dataset created by collecting audience measurement data for websites and the accuracy of the said algorithm is evaluated based on both linguistic features and social network features.
- The proposed method is able to utilize a distantly trained regression model providing classification accuracy that was competitive with a fully-supervised approach. This indicates that the consideration of the demographic composition of users in a network can also enhance the manner in which user behavior is comprehended.
- A significant percentage of improvement of the said algorithm, that is 0.53% and 0.79% for the value of average held-out correlation in the regression module for two of three models are obtained from a similar implementations to which comparisons are drawn, while one of the models reports 99.73% accuracy.
- In the classification module on the other hand, average F_1 scores report significant accuracies of 99.72% and 99.50% respectively in distantly supervised validation on the basis of gender and politics respectively. Such outcomes signify that the proposed distant supervision approach is comparable to a fully supervised baseline.

2. RELATED WORK

Social networks such as Facebook, Foursquare, Twitter, and Instagram have gained popularity of late and have garnered the interest of millions of users. Almost all of these social platforms have consistently exhibited dynamism and rapid growth.

An area of particular interest in the analysis of user behavior of social network participants is that of predicting user traits based on certain demographic parameters such as age, gender, race, political affiliation, occupation, or even web browsing histories.

Weakly supervised learning also known as distantly supervised learning is a variation to a standard supervised learning that relies less on hand annotated training data instead using declarative constraints to instantiate models.

Among the methods prevalent to train classifiers two namely, prior knowledge of label proportions and prior knowledge of feature-label associations dominate majority of the previous work presented in this field of research.

Distantly supervised learning provides a means to incorporate the declarative constraints available with social media, and has been leveraged in areas of document categorization, named- entity recognition, dependency parsing, language identification, and sentiment analysis.

In [2], Lars Backstrom, Eytan Bakshy, Jon Kleinberg, Thomas M. Lento, and Itamar Rosenn propose a measure for analyzing an individual's set of social contacts that address the dimension of attention across their personal network apart from its size and composition.

In [3], Ehsan Mohammady Ardehaly and Aron Culotta present a means to employ use lightly supervised learning to infer the age, ethnicity, and political orientation of Twitter users that pairs unlabeled Twitter data with constraints from country demographics, trends in first names, and exemplar Twitter accounts strongly associated with a class label.

In [4], Jacob Eisenstein, Noah A. Smith, and Eric P. Xing present a present a technique to find powerful and interpretable sociolinguistic relationship from crude geotagged content information. Utilizing total statistic insights about the creators' geographic groups, a multi-yield relapse issue amongst socioeconomics and lexical frequencies is thought about.

In [5], Prem Melville, Wojciech Gryc, and Richard D. Lawrence develop an effective framework for incorporating lexical knowledge in supervised learning for text categorization; and successfully apply the developed approach to the task of sentiment classification.

In [6], Wendy Liu and Derek Ruths perform a thorough investigation of the link between gender and first name in English tweets by characterizing how incorporating a user's name into a gender classifier improves the quality of inferred labels.

In this project, I aim to explore whether **an algorithm based on a distantly trained regression model**

can provide performance comparable to a fully supervised approach when evaluated on the basis of both linguistic features and social network features of certain users in real- life social context.

3. PROCUREMENT OF DATASET

The distantly labeled dataset as used in the implementation of [1] was procured from Dropbox by contacting the author and requesting access as it was password protected (**Link: https://www.dropbox.com/s/vtuha0pgihhxp4d/jair-2016-demographics-data.tgz?dl=0**). The way in which the dataset was created originally is detailed as follows.

A crowd of people estimation organization, Quantcast.com, tracks the demographics of guests to a huge number of sites. The utilization of treats to track the perusing movement of an extensive board of respondents helped in testing 1,532 sites from Quantcast to download measurements for seven demographic parameters:

- Gender: Female, Male
- Age: 18-24, 25-34, 35-44, 45-54, 55-64, 65+
- Income: $0-50k, $50-100k, $100-150k, $150k+
- Education: College, Grad School, No College
- Children: No Kids, Kids
- Ethnicity: Asian, Caucasian, Hispanic, African American
- Political Preference: Republican, Democrat

For every parameter, the assessed level of guests to a site with a given demographic was accounted for by Quantcast.

For each such site gathered, a script was executed to scan for its Twitter account. At that point each record got via looking was physically verified. Thus, 1,066 records from the first arrangement of 1,532 were found. A presumption for this situation was that the statistic profiles of supporters of a site on Twitter are related with the demographic profiles of guests to that site. While there are without a doubt inclinations presented here. For example, Twitter clients may skew more youthful than the web traffic board. In any case, in total these diff erences ought to have restricted effect on the final demonstrate. Each of the 1,066 Twitter accounts were given element vectors got from data about their devotees. Highlights construct both with respect to the interpersonal organization and on the semantic setting of every supporter's tweets were envisioned.

As a trademark, Twitter clients are permitted to "follow" other accounts. This thusly presents an unbalanced connection between clients. For example, if A takes after B, at that point B is a companion of A (however the switch may not be valid). For each record, the Twitter REST API to test 300 of its adherents was questioned utilizing the supporters/ids ask.

For each of these supporters, up to 5,000 of the accounts, they take after were accumulated using the friends/ids API inquire. These were implied as friends. In this way, for each of the primary locales from

Quantcast whose record is dynamic on Twitter, there are doing (300 ∗ 5K = 1.5M) additional records that are two ricochets from the main record (the partner of a fan). These discovered records were suggested as neighbors of the site's Twitter account. According to the possibility of triadic decision, countless discovered records ought to be replicated. Honestly, the middle supposition was that the amount of such duplicates addresses the nature of the closeness between the neighbors.

For each of the first records, the part of its devotees that are companions with each of its neighbors are figured and a neighbor vector was envisioned. The subsequent dataset comprised of 1.7M special neighbors of the first 1,066 records. To lessen dimensionality, neighbors with less than hundred adherents were evacuated, leaving 46,649 extraordinary neighbors with an aggregate of 178M approaching connections.

notwithstanding the neighbor vector, a practically equivalent to vector in view of the tweets of the supporters of each record was made wherein the latest 200 tweets for each of the 300 devotees of each of the 1,066 records were gathered utilizing the statuses/client course of events API ask. For each tweet, standard tokenization, expulsion of non-inside accentuation, change to bring down case, and retainment of hashtag and notices were performed. URLs and digits were each crumpled to a solitary element write and characters rehashed more than twice were considered a solitary event. Terms utilized by less than twenty diff erent clients were evacuated.

The subsequent dataset comprised of 9,427,489 tweets containing 112,642 extraordinary terms composed by 59,431 clients. For each of the first 1,066 records, a content vector like the neighbor vector was made wherein each esteem spoke to the extent of supporters of the record who utilize that term.

For endorsement purposes, sex and ethnicity data were at first assembled by the makers of as takes after.First, Twitter Streaming API was utilized to acquire an arbitrary example of clients, filtered to the United States (utilizing time zone and the place nation code from the profile). From information traversing length of six days i.e. December 6-12, 2013, 1,000 profiles were inspected indiscriminately and ordered by breaking down the profile, tweets, and profile picture for every client. 770 Twitter profiles were arranged into one of four ethnicities (Asian, African American, Hispanic, and Caucasian). Those for which ethnicity couldn't be resolved were disposed of (230/1,000; 23%). The class recurrence watched was Asian (22), African American (263), Hispanic (158), and Caucasian (327). To gauge between annotator assention, a moment annotator examined and arranged 120 clients. Among clients for which the two annotators chose one of the four Fcategories, 74/76 names concurred (97%). There was some contradiction over when the class could be resolved: for 21/120 marks (17.5%), one annotator demonstrated the classification couldn't be resolved, while the other chosen a classification.

Gender explanation was done naturally by contrasting the first name furnished in the client profile with the U.S. Statistics rundown of names by sexual orientation and questionable names were evacuated.

For every client, up to 200 of their companions were gathered utilizing the Twitter API and records that limited gets to companion data were evacuated. Asian clients were additionally evacuated because of the little example measure, leaving an aggregate of 615 clients. For classification, every client was spoken to by the character of their companions (up to 200). Just those companion accounts contained in the 46,649 records utilized for the relapse tests were held. Moreover, up to 3,200 tweets were gathered from every client and a double term vector was built, utilizing an indistinguishable tokenization from in the relapse show.

The political inclination information originates from utilization of the geo-driven part of the information, which contains Twitter clients from Maryland, Virginia, or Delaware who report their political affiliation in their Twitter profile portrayal. This contained 183 Republican clients and 230 Democratic clients. Every client has up to 5,000 companions and 200 tweets in this dataset.

4. PROBLEM STATEMENT

To determine whether an algorithm based on a distantly trained regression model can provide performance comparable to a fully supervised approach when evaluated on the basis of both linguistic features and social network features of certain users in real-life social context.

Importance associated with Problem Statement:

- To fit regression models to anticipate seven statistic parameters of Twitter clients construct both with respect to whom they take after and on the substance made by them in how a model is fit between an arrangement of clients and their normal statistic profile.

- To give extra approval to the proposed display with physically commented on Twitter accounts to investigate whether a similar model can likewise precisely anticipate the socioeconomics of individual clients separated from precisely describing socioeconomics of an arrangement of Twitter accounts.

- To expand the **held-out** connection crosswise over statistic parameters and stretch out the regression model to order singular clients.

4.1 Existing Solution

The major weakness of supervised classification approach to demographic inference is that human annotations may prove to be error-prone, while many a times, it becomes difficult to adjudge demographics of interest as there may be limitations imposed on the generalizability of classification. Moreover, network evidence must be considered as important as textual evidence to provide a solid backing to the inference being reported at any given instant.

The current methodologies for prediction of user demographics have many points of interest and admonitions. Numerous basic techniques have been appeared to perform well in anticipating age, sex and racial/ethnicity under specific suppositions. In any case, impediments exist in the adjustment of these techniques to particular areas, e.g. general wellbeing, where studies may require vast scale datasets and might be time delicate. Such techniques underlined the need to gather nitty gritty client data to enhance classifier execution without tending to issues in productivity and versatility, which limits area particular relevance.

4.2 Proposed Solution

The proposed solution takes into consideration, the textual content as well as social context of users wherein social interactions are measured by seven different demographic parameters that span across a variety of prevalent factors like gender, age, income, education, ethnicity, parental status, and political preference. These influencing factors in a social network are used to explore the demographic constitution of a sample of social media users.

Web traffic data is used as a form of weak supervision, and followership information as the primary source of evidence. The analysis includes how accuracy varies with respect to the number of linguistic terms collected per user.

The algorithm is applied on a distantly labeled dataset created by collecting audience measurement data for websites and the accuracy of the said algorithm is evaluated based on both linguistic features and social network features. The proposed method is able to utilize a distantly trained regression model providing classification accuracy that was competitive with a fully- supervised approach. This indicates that the consideration of the demographic composition of users in a network can also enhance the manner in which user behavior is comprehended.

The proposed approach works best when:

- The datum provided models a fairly connected network of users, drawing conclusions from textual content as well as network evidence and not just either of the two.
- A presumption of this work is that the statistic profiles of devotees of a site on Twitter correspond with the demographic profiles of guests to that site considering that Twitter followership is basically an unbalanced relationship.

4.3 Procedure

- The distantly labeled dataset as used in the implementation of [1] was procured from Dropbox by contacting the author and requesting access as it was password protected (**https://www.dropbox.com/s/vtuha0pgihhxp4d/jair-2016-demographics- data.tgz?dl=0**).

- Regression models were fit to predict seven demographic parameters, namely age, gender, education, income, ethnicity, parental status, and political preference of Twitter users, thus in turn fitting a model between a set of users and their expected demographic profile.

- Validation checks were performed against a set of users manually labeled by race, gender, and political preference, thus considering predicting demographics both at the aggregate level and at the individual level.

- The scripts were written in Python language, on text editor Sublime Text (Build 3.0.47) and executed on IDLE GUI (Python Shell).

4.4 F lowchart

Fig. 1: Flowchart demonstrating procedure followed by Proposed Approach

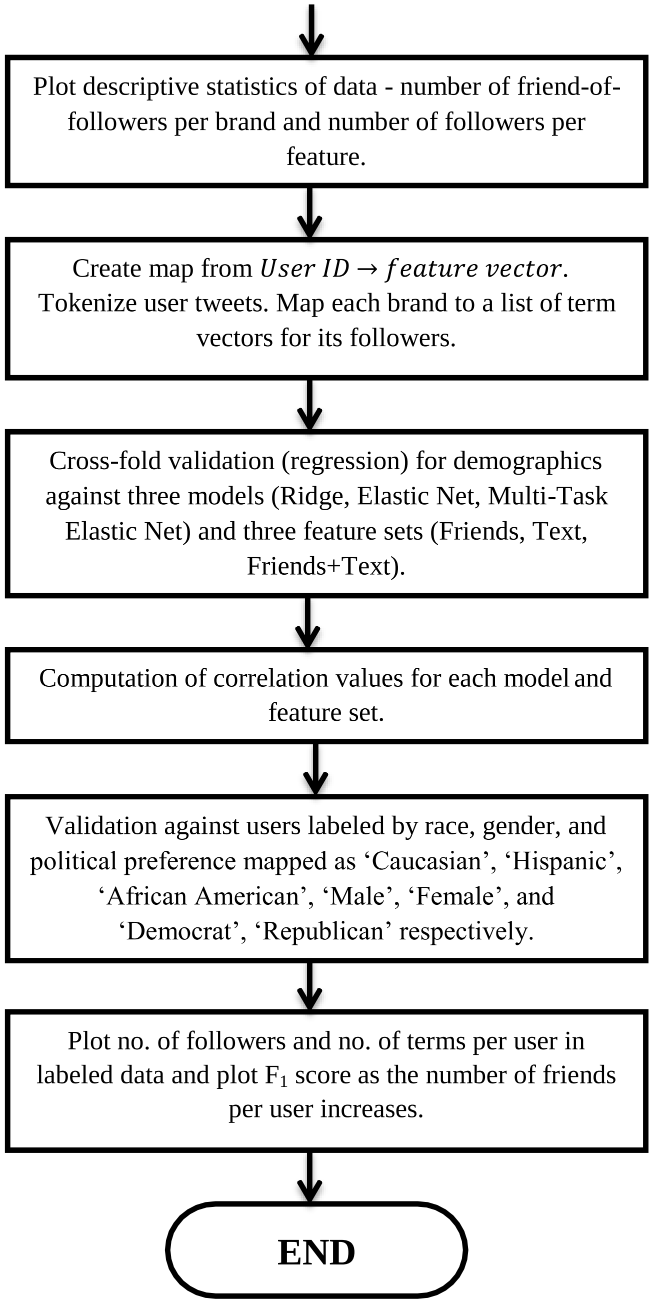

Plot descriptive statistics of data - number of friend-of-followers per brand and number of followers per feature.
Create map from $User\ ID \rightarrow feature\ vector$. Tokenize user tweets. Map each brand to a list of term vectors for its followers.
Cross-fold validation (regression) for demographics against three models (Ridge, Elastic Net, Multi-Task Elastic Net) and three feature sets (Friends, Text, Friends+Text).
Computation of correlation values for each model and feature set.
Validation against users labeled by race, gender, and political preference mapped as ‘Caucasian’, ‘Hispanic’, ‘African American’, ‘Male’, ‘Female’, and ‘Democrat’, ‘Republican’ respectively.
Plot no. of followers and no. of terms per user in labeled data and plot F_1 score as the number of friends per user increases.
END

Figure 1 depicts the procedure adopted in order to implement the proposed approach. The demographics of each brand were read. Descriptive statistics of data were plotted with number of friend of followers and number of followers per feature.

A mapping from User ID → feature vector was created and user tweets were tokenized. Each brand was mapped to a list of term vectors for its followers. Cross-validation (regression) for demographics against three models (Ridge, Elastic Net, Multi-Task Elastic Net) and three feature sets (Friends, Text, Friends+Text) was done. Correlation values for each model and feature set were computed.

Users labeled by race, gender, and political preference mapped as ‘Caucasian’, ‘Hispanic’, ‘African American’, ‘Male’, ‘Female’, and ‘Democrat’, ‘Republican’ respectively were validated. The number of followers and number of terms per user in labeled data were plotted and F_1 score was plotted as the number of friends per user increased.

4.5 Block Diagram

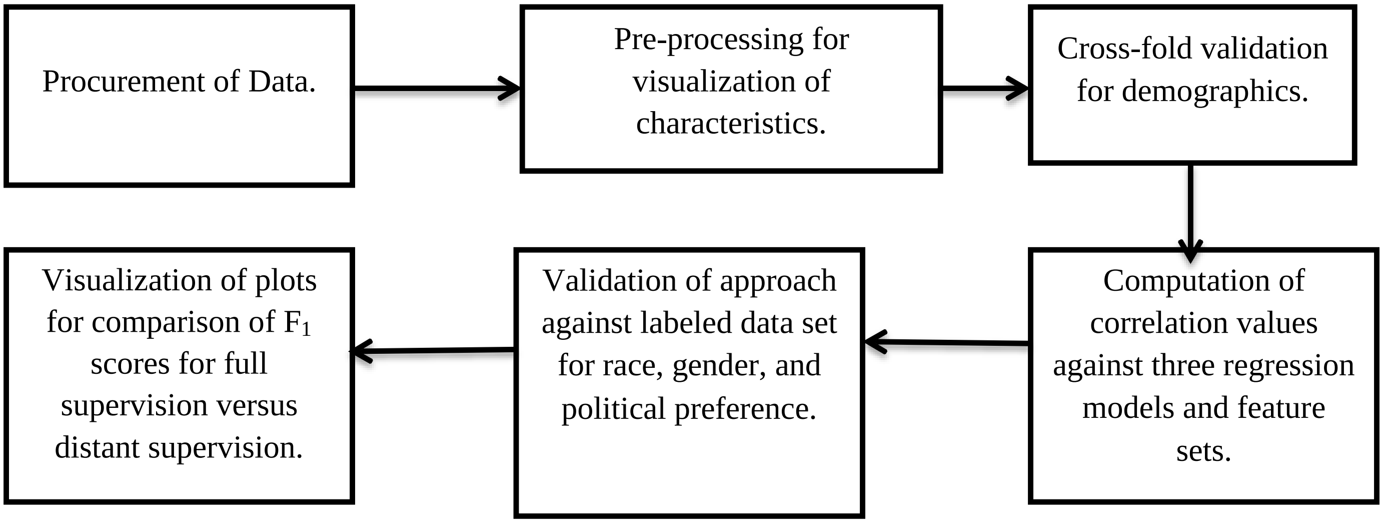

Procurement of Data.
Pre-processing for visualization of characteristics.
Cross-fold validation for demographics.
Computation of correlation values against three regression models and feature sets.
Validation of approach against labeled data set for race, gender, and political preference.
Visualization of plots for comparison of F1 scores for full supervision versus distant supervision.

Diagram depicting elements involved in implementing the Proposed Approach

Figure 2 depicts the elements involved in the proposed approach and the steps taken to achieve presentable results and comparisons.

4.6 Pseudocode

Input: Files consisting of a map from a brand's Twitter ID to its Twitter followers, Twitter ID to brand information, and Twitter username to the brand demographic information, along with a comma-separated file in the format $< brand\ twitter\ handle >< brand\ ID > [list\ of\ Twitter\ IDs\ of\ followers\ of\ this\ brand]$

Output: File containing the average correlation across all attributes, as well as the raw predictions for each demographic attribute.

Objective: Prediction of the demographics of Twitter users validated using a distant supervision approach.

```
1. Read the demographics of each brand and check whether found or not.
Initialize count to zero.
for each username $U_{i,}$ brand $B_i$
    if username $U_i$'s demographics found
        set corresponding $B_i$ values
        increment count
2. Plot descriptive statistics of data - number of friend-of-followers per brand and number of followers per feature.
3. Normalize mapping from a brand's Twitter ID to its Twitter followers and create sparse matrix.
4. Read a map from $User\ ID \rightarrow feature\ vector$, one per follower. .
for each user $U_i$
    for each tweet $T_i$
        initialize text to contents of $T_i$
            for each char $C_i$ in $T_i$
                if $C_i$ is @
                    set text to MENTION
            else if $C_i$ is #
                    set text to HASHTAG
    else if $C_i$ is one of {0,1,2,3,4,5,6,7,8,9}
    set text to DIGIT
1. Read the list of followers of each brand and map each brand to a list of term vectors for its followers.
2. Do cross-fold validation for different demographics. Display top coefficients for each demographic category.
3. Initialize output list to all demographic parameters along with respective categories.
        outputs = {'Politics': ['Democrat', 'Republican'], 'Education': ['No College', 'College', 'Grad School'],
    'Children': ['No Kids', 'Has Kids'], 'Income': ['$0-50k', '$50-100k', '$100-150k', '$150k+'], 'Gender': ['Male',
    'Female'], 'Age':
    ['18-24', '25-34', '35-44', '45-54', '55-64', '65+'], 'Ethnicity': ['Caucasian', 'Hispanic', 'African American', 'Asian']}
4. Use Multi-Task Elastic Net Model with three different feature sets: Friends, Text, Friends+Text.
5. Tune regularization parameters.
    for each User $U_i$
    l1_ratio = 0.5
    if feature_set is Friends
    $\alpha = 1e - 5$
    else if feature_set is Text or feature_set is Friends+Text
    $\alpha = 1e - 2$
    6. Obtain correlation values against each of the three regression models i.e. Ridge, Elastic Net and Multi-Task
    Elastic Net for each of the three feature sets: Friends, Text, Friends+Text.
7. Validate against a list of users manually labeled by race, gender, and political preference.
    8. Map race labels as ‘Caucasian’, ‘Hispanic’, and ‘African American’, gender labels as ‘Male’, ‘Female’, and
    political preference labels as ‘Democrat’, ‘Republican’.
9. Perform comparisons of fully supervised regression to that of distant supervision by plotting learning curves.
```

10. Plot no. of followers and no. of terms per user in labeled data and plot F_1 score as the number of friends per user increases.

11. Perform comparison of performance metrics obtained in proposed approach to existent approach.

Fig 3: Pseudocode to implement the Proposed Approach

Figure 3 elucidates the followed approach to implement first, fitting of a regression model to predict user demographics based on both linguistic features as well as social network features, and then providing additional validation to facilitate predicting the demographics of individual users. The distantly trained regression model provided performance comparable to a fully supervised approach.

In order to implement the proposed approach, demographic parameters were paired with respective friend and text feature vectors of each Quantcast website to construct a regression problem.

As a legitimization for high dimensionality (46,649 companion highlights and 112,642 content highlights) and the modest number of cases (1,066), versatile net regularization was utilized and a multi-errand variation of the flexible net was used to guarantee that similar highlights are chosen by the L1 regularizer for every needy classification in statistic parameters being considered.

Three renditions of this model were fit utilizing three diff erent include sets: Friends, Text, and Friends+Text. The regularization parameters on a held-out arrangement of 200 records for Gender forecast were tuned, setting the scikit-learn parameters l1 proportion = 0.5 for each model, alpha = 1e − 5 for the Friends model, and alpha = 1e − 2 for the Text and Friends+Text models. Five-overlap cross-approval was performed and the held-out connection coefficient (r) between the anticipated and genuine statistic parameters was watched.

For approval purposes, three parameters that can be named decently dependably for people: gender, ethnicity, and political inclination were considered.

As the proposed demonstrate was at first prepared for a relapse assignment, a couple of modifications were made to apply it to a classification undertaking. Every client in the named information was spoken to as a twofold vector of companion and content highlights, utilizing an indistinguishable tokenization from in relapse.

For instance, if a client takes after accounts An and B, at that point the component esteems were 1 for those comparing accounts; also, if the client notices terms X and Y, at that point those element esteems were 1.

To repurpose the relapse model to perform classification, the coefficients returned by relapse were changed. The z-score of each coefficient concerning alternate coefficients for that class esteem were processed.

For instance, all coefficients for the Male class were changed in accordance with have zero mean and unit difference. This made the coefficients practically identical crosswise over marks. To order every client, the spot item between the coefficients and the double component vector were processed, choosing the class with greatest esteem.

To arrange an arrangement of clients, the element coefficient dab item was registered independently for the content and companion models, at that point the z-score of the subsequent esteems was figured by class. This put the anticipated esteems for each model in a similar range. Summation of the yields of the two models was done and the class with the most extreme incentive for every client was returned.
Keeping in mind the end goal to contrast and a completely directed gauge, a strategic relapse classifier was prepared with L2 regularization, utilizing a similar element portrayal as above and three-crease cross-approval was performed to contrast precision and the remotely administered approach.

5. EXPERIMENTAL RESULTS

5.1 Dataset Characteristics

The characteristics of the data set as obtained from Dropbox and utilized for Regression and Classification modules have been tabulated as follows:

Table 1: Characteristics of Dataset

Characteristic	**Regression Phase**		**Classification Phase**		
	Friend Features	**Text Features**	**Gender**	**Ethnicity**	**Politics**
No. of websites sampled from Quantcast	1532				
No. of accounts verified from Twitter	1066				
No. of Users	1532 (1066)	59431	1000 (615)	1000 (615)	413
o. of Unique Neighbors	1.7M (46649)	**NA**	**NA**	**NA**	**NA**
No. of Links	178M	**NA**	**NA**	**NA**	**NA**
No. of Tweets	**NA**	9427489	3200	3200	200
of Unique Terms	**NA**	112642	**NA**	**NA**	**NA**

Table 1 lists the standard characteristics of the data set as provided by the author of [1].

The above characteristics are tabulated to gain a better understanding of how large the data set is and how much of it was taken into consideration (depicted in brackets) for purpose of implementation of the proposed approach.

5.2 Performance Of Algorithm

Regression Module

The pseudocode specified in Sect. 4.7 was implemented after following the procedure specified in Sect. 3.1.

Figure 4 is plotted to provide a **pictorial representation of descriptive statistics** of the data, namely number of friend-of-followers (unique neighbors) per brand and number of followers (neighbor links) per feature.

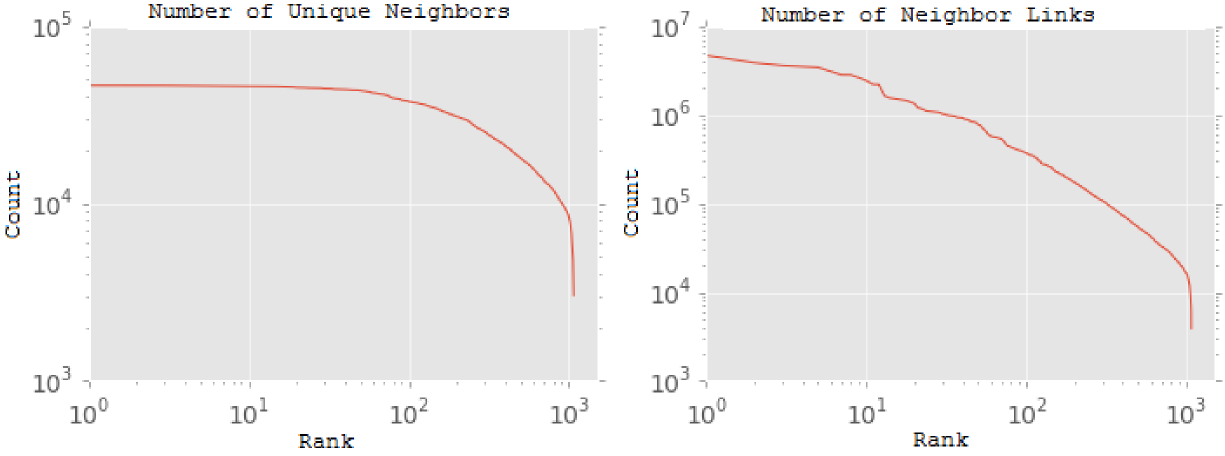


Fig. 4: Rank Order Frequency Plots of the descriptive statistics of the dataset

Figure 4 illustrates the rank order frequency plots of the number of neighbors per account and the number of links to all neighbors per account. It was observed that a plot similar to Fig. 2 of [1] was obtained.

The Twitter accounts with the highest estimated coefficients for each demographic parameter are tabulated as follows:

Accounts with the highest estimated coefficients for each demographic parameter

Demographic Parameter	**Category**	**Top Accounts**
Gender	Female	The Ellen Show, Oprah, Martha Stewart, Pinterest, Etsy
	Male	Sports Center, Adam Schefter, mortreport, WIRED, espn
Age	18-24	RockstarGames ,IGN, steam_games, PlayStation, Ubisoft
	25-34	lenadunham, azizansari, WIRED, mindykaling
	35-44	BarackObama, cnnbrk, TMZ, AP espn,
	45-54	cnnbrk, CNN, WSJ , AP FoxNews
	55-64	cnnbrk, ABC, AP, FoxNews
	65+	AP, cnnbrk , WSJ, FoxNews, DRUDGE_REPORT
Income	$0-50k	PlayStation, YouTube , RockstarGames, Drake, IGN
	$50-100k	AdamSchefter, cnnbrk, SportsCenter, espn, ErinAndrews
	$100-150k	WSJ, espn, AdamSchefter, SportsCenter, ErinAndrews

	$150k+	WSJ, TheEconomist, Forbes, nytimes, business
Education	No College College Grad School	YouTube, PlayStation, RockstarGames, Xbox, IGN StephenAtHome, WIRED, ConanOBrien, mashable nytimes, WSJ, NewYorker, TheEconomist, washingtonpost
Parental Status	No Kids Has Kids	NewYorker, StephenAtHome, nytimes, maddow, pitchfork parenting, Huff PostParents, TheEllenShow, thepioneerwoman, parentsmagazine
Political Preference	Democrat Republican	BarackObama, Oprah, NewYorker, UncleRUSH, MichelleObama FoxNews, michellemalkin, seanhannity, megynkelly, DRUDGE_REPORT
Ethnicity	Caucasian Hispanic African American Asian	jimmyfallon, FoxNews, blakeshelton, TheEllenShow, TheOnion latimes, Lakers, ABC7, Dodgers, KTLA KevinHart4real, Drake, Tip, iamdiddy, UncleRUSH TechCrunch, WIRED, BillGates, TheEconomist, SFGate

Table 2 delineates the highlights with the five biggest coefficients for every class as indicated by the relapse show fit utilizing just companion highlights. It was watched that a table like Table 1 of]was gotten, wherein a significant number of the outcomes were in concurrence with regular generalizations.

Figure 5 is observed to provide a **visualization of the scatter plots of the true demographic parameters from Quantcast versus those predicted using elastic net regression fit to friend and text features from Twitter**.

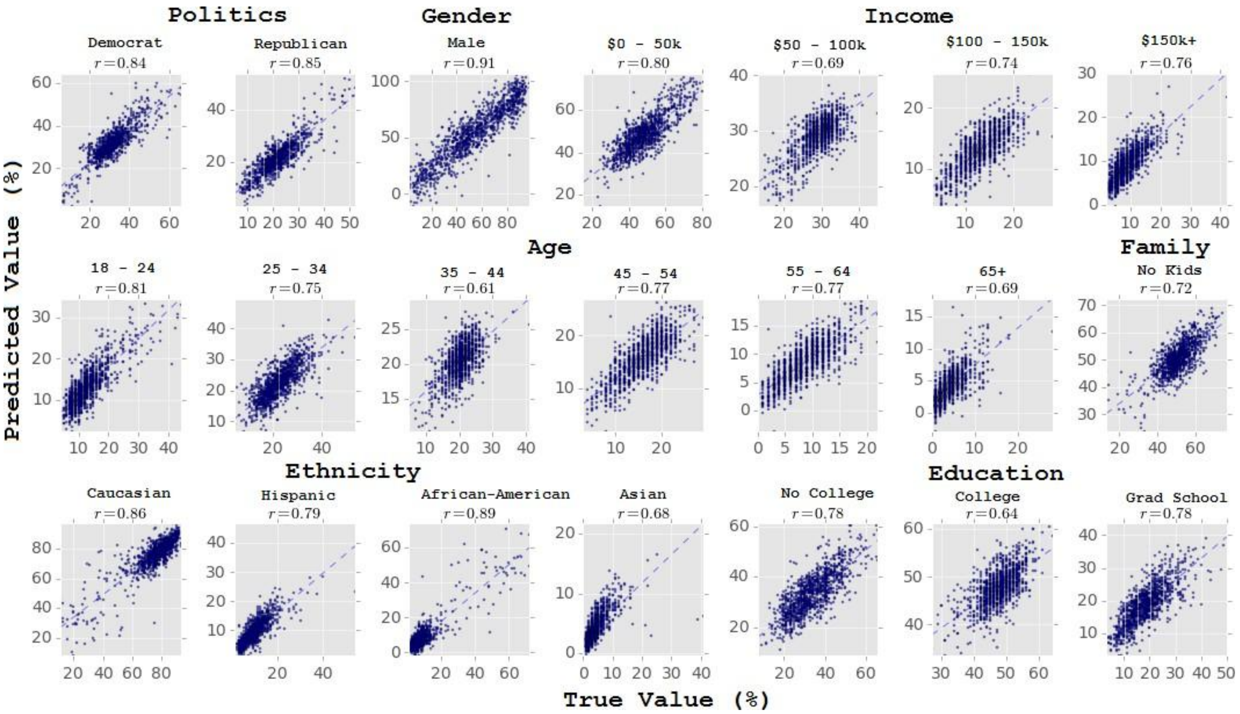


Fig. 5: Scatter Plots of True vs Predicted Demographic Parameters

Figure 5 illustrates the held-out correlation coefficient (**r**) **alongside each prediction** computed using five fold cross-validation. It was observed that a plot similar to Fig. 3 of [1] was obtained and an **average held-out correlation value of 0.768** was reported.

The terms with the highest estimated coefficients for each demographic parameter are tabulated as follows:

The highest estimated coefficients for each demographic parameter

Demographic Parameter	**Category**	**Top Terms**
Gender	Male	film, guy, gay, man, fuck, game, team, internet review, guys
	Female	hair, her, omg, family, girl, she, girls, cute, beautiful, thinking

Age	18-24	d, haha, album, x, xd, _:, actually, stream, wanna, im
	25-34	super, dc, baby, definitely, nba, pregnancy, wedding, even, entire, nyc
	35-44	star, fans, kids, tv, bike, mind, store, awesome, screen, son
	45-54	wow, vote, american, comes, ca, santa, country, boys, nice, high
	55-64	vote, golf, red, american, country, north, county, holiday, smile, 99,999
	65+	vote, golf, MENTION_foxnews, holiday, may, american, he, family, north, national
Income	$0-50k	lol, games, MENTION_youtube, damn, black, ps9, side, d, community, god
	$50-100k	great, seattle, he, performance, lose, usa, kansas, iphone, wow, cold
	$100-150k	santa, flight, nice, looks, practice, congrats, bike, dc, retweet, ride
	$150k+	dc, nyc, market, MENTION_wsj, congrats, beach, san, york, ca, looks

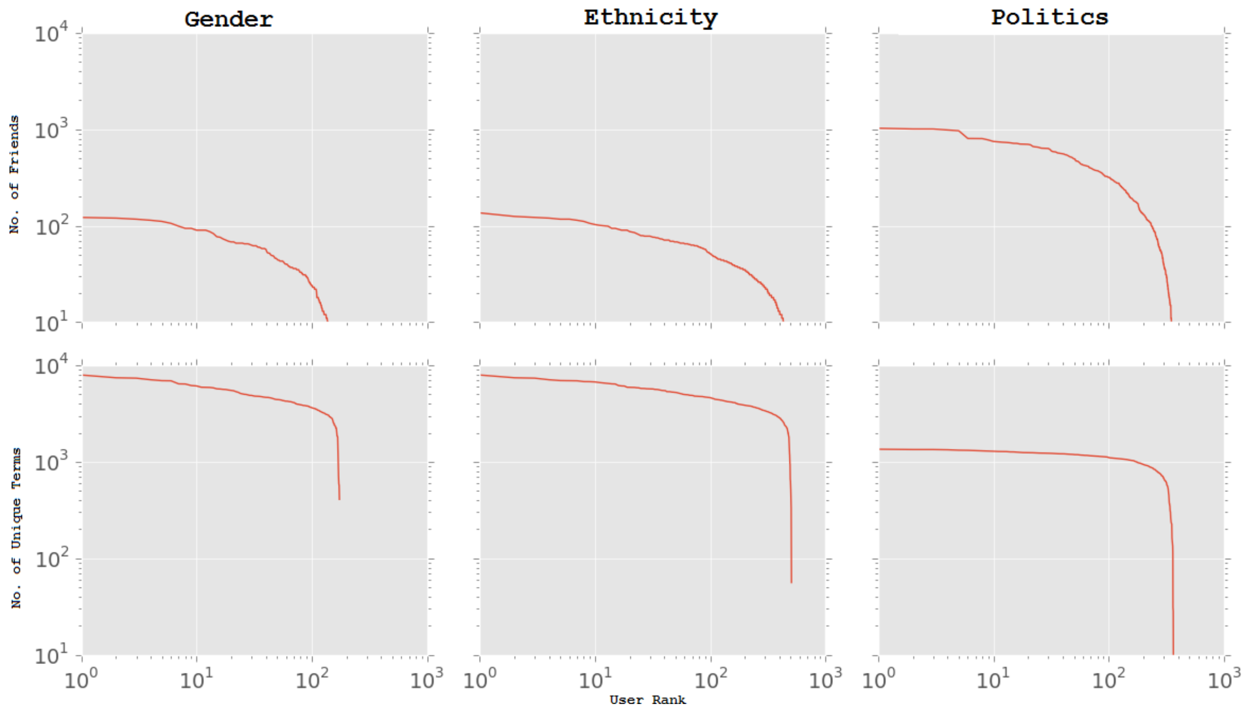
Education	No College College Grad School	lol, games, put, MENTION_youtube, county, made, ps9, xbox, videos, found our, you're, seattle, photo, MENTION_mashable, la, apple, fashion, probably, san dc, MENTION_nytimes, market, which, review, excellent, boston, also, congrats, MENTION_washingtonpost
Parental Status	No Kids Has Kids	care, street, gay, years, health, drink, dc, white, ht…, album kids, school, child, family, kid, daughter, children, utah, moms, parents
Political Preference	Democrat Republican	women, u, ain't, nyc, equality, la, voice, seattle, dc, MENTION_nytimes MENTION_foxnews, christmas, HASHTAG_tcot, football, county, morning, family, christians, country, obama's
Ethnicity	Caucasian Hispanic African American Asian	christmas, fun, dog, country, st, could, luck, guy, florida, john la, los, san, el, angeles, california, ca, lol, l.a, lakers black, lol, bout, ain't, brown, lil, african, blessed, smh, atlanta chinese, la, sf, san, china, korea, india, bay, vs,

Table 3 depicts the features with the ten largest coefficients per class according to the regression model fit using only text features. It was observed that a table similar to Table 2 of [1] was obtained, wherein many of the results were in agreement with common stereotypes as was the case with friend features but a greater level of granularity was observed.

Classification Module

The pseudocode specified in Sect. 4.7 was implemented after following the procedure specified in Sect. 3.2.

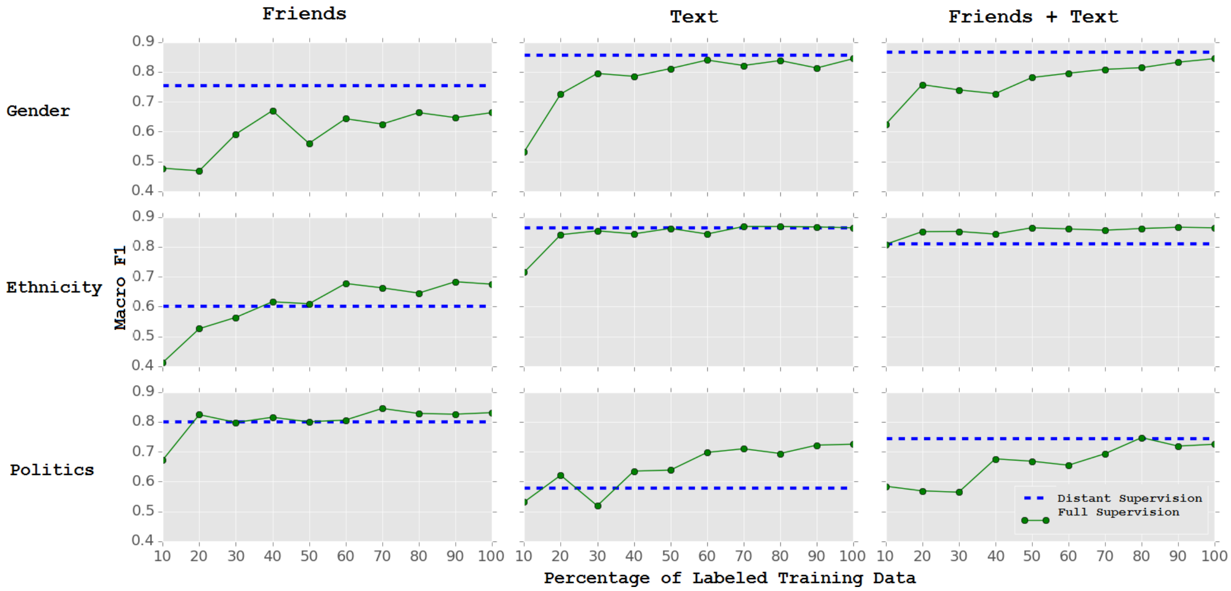
Figure 6 shows the number of friends and the number of unique terms per user for each labeled dataset according to gender, ethnicity, and politics.



Frequency Plots of the descriptive statistics of the labeled datasets

Figure 6 illustrates the rank order frequency plots of the number of friends per user and the number of unique terms per user in each of the labeled datasets. The plots were restricted to one of the 46,649 accounts and 1,12,642 terms used in the regression module. It was observed that a plot similar to Fig. 4 of [1] was obtained.

Figure 7 compares the compares the accuracy of the fully supervised approach to that of the distantly supervised one as the number of labeled data increase.



Standard logistic regression classifier to the proposed approach

Figure 7 illustrates Twitter user classification results, wherein the proposed approach is fit on statistics from Quantcast. It was observed that distant supervision is comparable to full supervision as a plot similar to Fig. 5 of [1] was obtained.

In order to ascertain how much information about a user is needed before an accurate prediction regarding demographics is made, a subset of friends and terms for each user was randomly sampled, and the F_1 score was observed as the number of selected features increased.

For friends, subsets of size {1,3,5,10,20,30,40,50} were considered while for terms, we subsets of size {10,100,1000,2000,8029} were considered.

Figure 8 is plotted to observe a **pictorial representation of the F_1 scores of the distantly supervised approach as the number of friends and number of unique terms collected per user increase**.

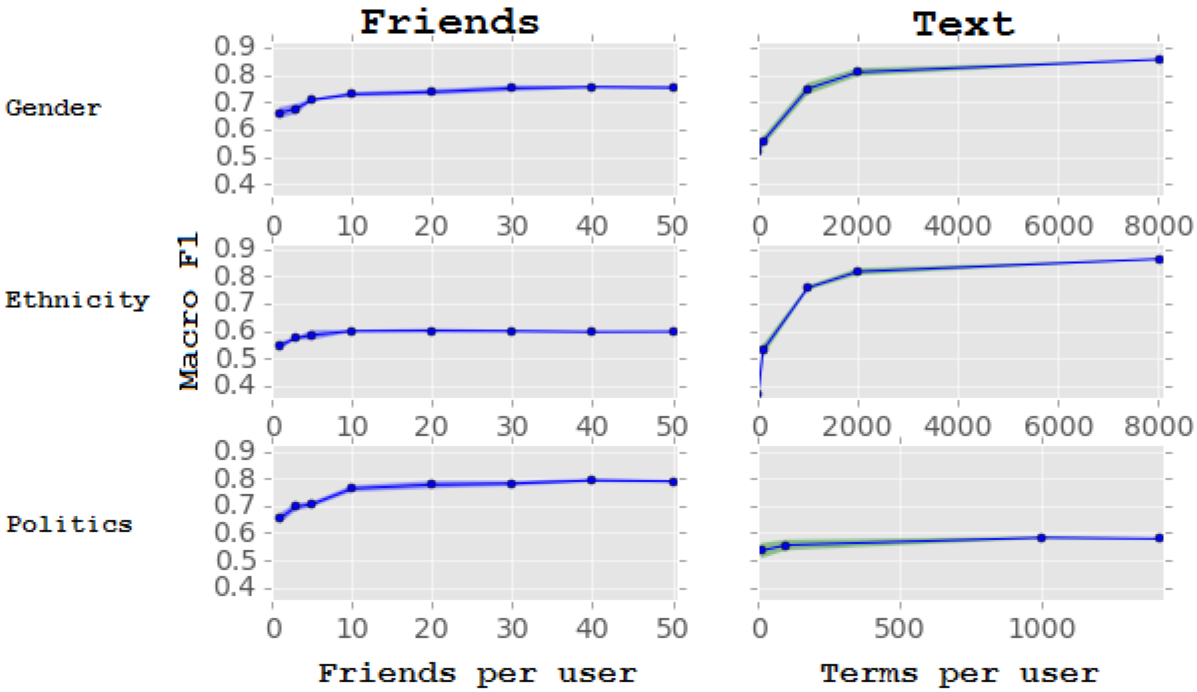


Fig. 8: Classification F_1 Scores of the proposed approach

Figure 8 illustrates how accuracy plateaus quickly using friend features while it is comparatively gradual for text features. It was observed that a plot similar to Fig. 6 of [1] was obtained.

5.2.1 Evaluation Of Performance Metrics

Table 4: Comparison of held-out correlation values as per proposed approach

Model	**Friends**	**Text**	**ends + Text**	**Average**
Multi-task Elastic Net	0.726	0.786	0.786	0.766
Elastic Net	0.724	0.776	0.762	0.754
Ridge	0.616	0.788	0.781	0.728

Table 4 depicts the computed value of performance metric Held-Out Correlation obtained as per the proposed approach for the various regression models. It was observed that a table similar to Table 3 of [1] was obtained.

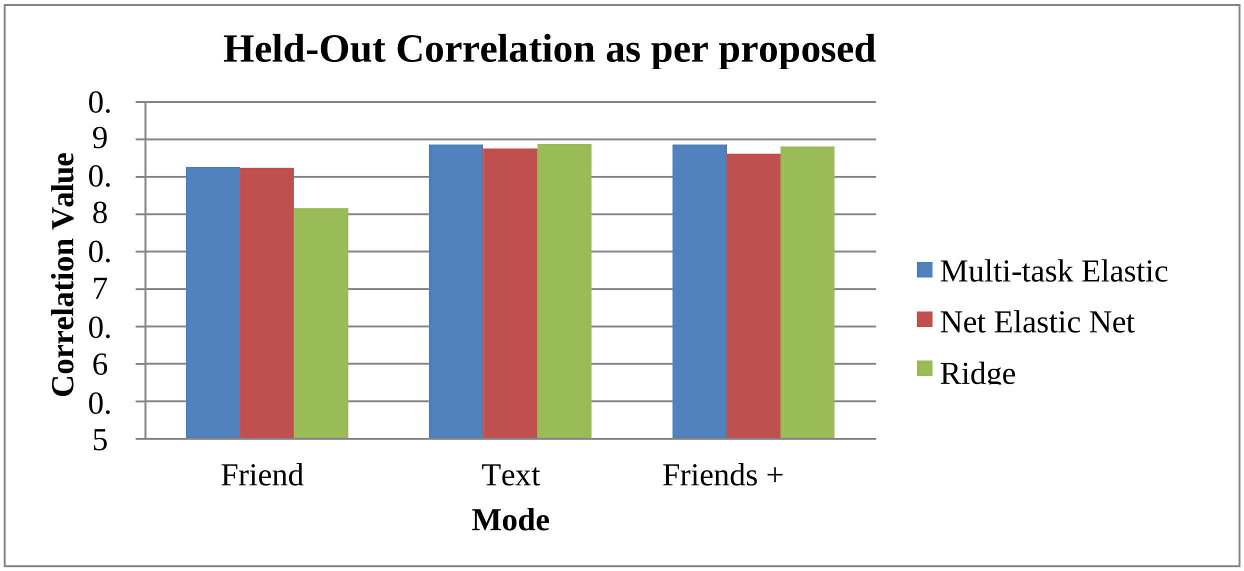


Fig. 9: Held-Out Correlation Values

Figure 9 **shows a** graphical representation **of the values of** performance metric Held-Out Correlation **obtained as per the proposed approach for the various regression models to provide a better visualization.**

Table 5: Comparison of F_1 score as per proposed approach

	Friends		**Text**		**Friends + Text**	
	Distant Supervision	Full Supervision	Distant Supervision	Full Supervision	Distant Supervision	Full Supervision
Gender	0.754	0.663	0.856	0.844	0.866	0.839
Ethnicity	0.601	0.675	0.863	0.864	0.810	0.864
Politics	0.799	0.831	0.579	0.725	0.744	0.857
Average	0.718	0.723	0.766	0.811	0.806	0.853

Table 5 depicts the computed value of performance metric F_1 Score obtained as per the proposed approach for distant supervision as compared to that obtained by full supervision. It was observed that a table similar to Table 4 of [1] was obtained.

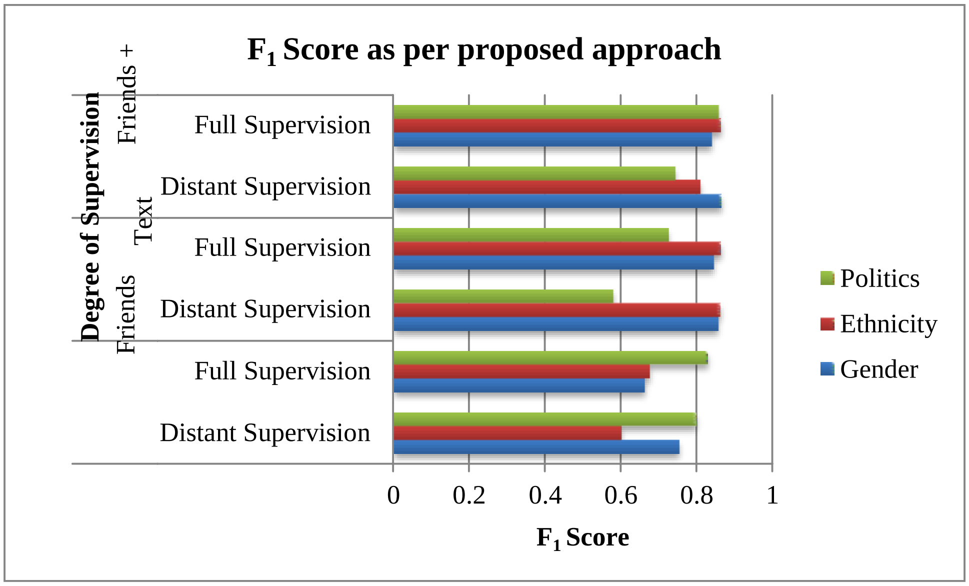


Fig. 10: F_1 Score Values

Figure 10 shows a **graphical representation** of the values of **performance metric F_1 Score** obtained as per the proposed approach for distant supervision as compared to that obtained by full supervision to provide a better visualization.

6: Improvement or Degree of Accuracy obtained over similar implementation

Model	**Proposed Approach**	**Similar Approach [1]**	**Percentage Accuracy or Improvement**
Multi-task Elastic Net	0.766	0.76	0.79 (I)
Elastic Net	0.754	0.75	0.53 (I)
Ridge	0.728	0.73	99.73 (A)

The **factor of improvement or degree of accuracy obtained over the computation of average held-out correlation in a similar approach** presented in [1] is expressed as a percentage and consolidated in Table 6. The letters I and A in brackets signify improvement and accuracy respectively.

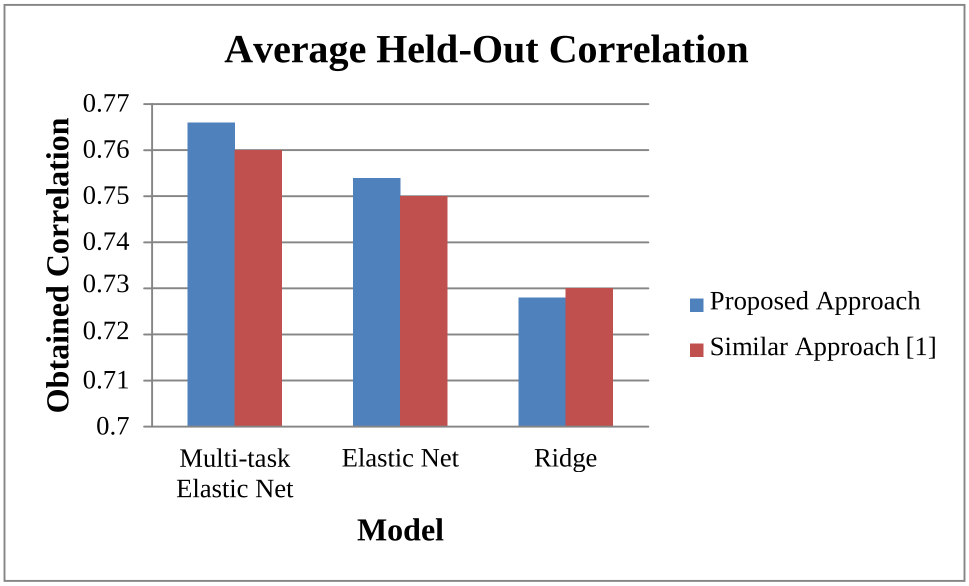


Fig. 11: Average Held-Out Correlation Values

Figure 11 plots the values of **held-out correlation averaged over the three models** obtained as per the proposed approach compared to that obtained as per the similar approach in [1] to provide a better visualization.

Table 7: Improvement or Degree of Accuracy obtained over similar implementation

	Proposed Approach		**Similar Approach [1]**		**ntage Accuracy or Improvement**	
	Distant Supervision	Full Supervision	Distant Supervision	Full Supervision	Distant Supervision	Full Supervision
Gender	0.718	0.723	0.72	0.72	99.72 (A)	0.42 (I)
Ethnicity	0.766	0.811	0.76	0.81	0.79 (I)	0.12 (I)
Politics	0.806	0.853	0.81	0.81	99.5 (A)	5.3 (I)

The **factor of improvement or degree of accuracy obtained over the computation of average F_1 score in a similar approach** presented in [1] is expressed as a percentage and consolidated in Table 7. The letters I and A in brackets signify improvement and accuracy respectively.

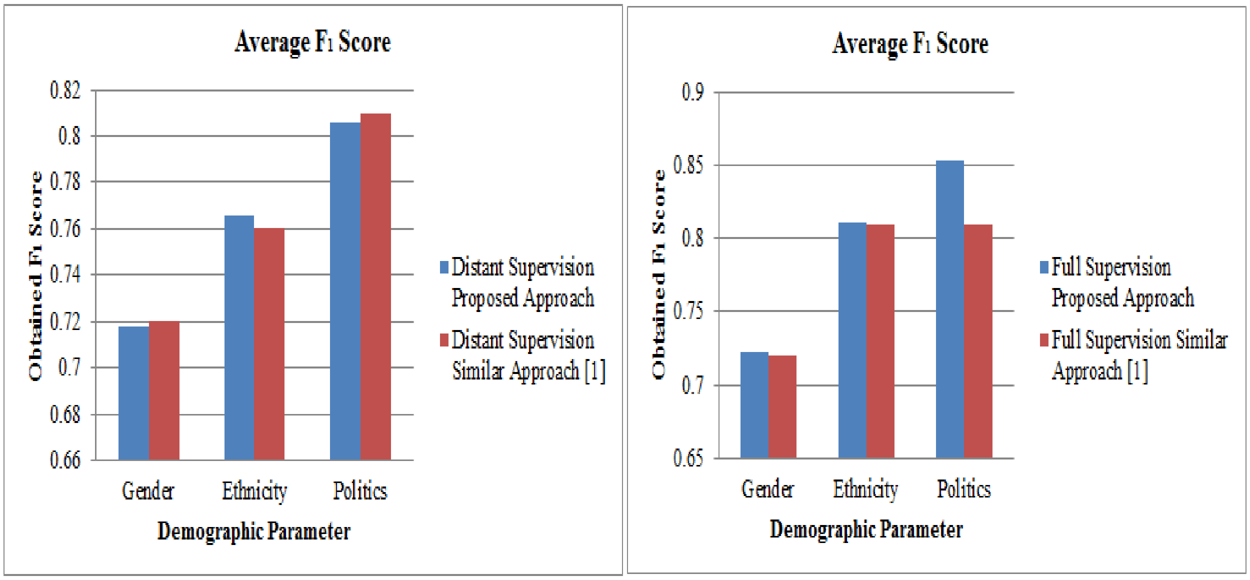


Fig. 12: Average F_1 Score Values

Figure 12 plots the values of **F_1 score averaged over the three demographic parameters** obtained as per the proposed approach compared to that obtained as per the similar approach in [1] to provide a better visualization.

6. CONCLUSION

By creating techniques for robotized location of key demographic highlights of clients, scientists have tried to develop the utilization of online networking information for look into. The suspicion that drives this examination is the clients abandon pieces of information with respect to their disconnected characters – either verifiably or expressly. And after that it takes a little measure of endeavors to create systems to recognize vital statistic hints. The key capability of this examination is to make chances to evaluate populace portrayal and furthermore to ponder incongruities in online networking information. For example, knowing key statistic data about clients could enable specialists to better comprehend sex, race or age-based disparity in data access or dispositions toward current occasion.

In this paper, I propose an algorithm that fits a regression model to predict user demographics using both textual content and social network evidence of Twitter clients crosswise over seven unique parameters and applies the model to arrange singular clients based on ethnicity, sexual orientation and political inclination. A remotely named dataset made by gathering group of onlookers estimation information for sites is used in the way that web activity demographic information from Quantcast is brought. Web activity socioeconomics of a website with the devotees of that webpage on Twitter are matched to fit a relapse show between an arrangement of Twitter clients and their normal statistic profile.

The said approach is evaluated on the basis of performance metrics for optimal inference of demographics basis textual content and network influence of users. A distantly supervised learning approach is employed for the algorithm and the results obtained are compared to a similar implementation in the recent past. A distantly labeled dataset created by collecting audience measurement data from the website of a popular audience measurement company, in the proposed approach, is utilized. To measure the effectiveness of the algorithm, accuracy evaluation on the basis of both linguistic features and social network features was employed.

The proposed method which utilized a distantly trained regression model provided classification accuracy that was competitive with a fully-supervised approach. I was able to thus establish that **an algorithm based on a distantly trained regression model can provide performance comparable to a fully supervised approach when evaluated on the basis of both linguistic features and social network features of certain users in real-life social context**.

It is proposed to consider textual content as well as social context of users for representing the given datum in a better structured manner, wherein social interactions are measured by seven different demographic parameters that span across a variety of prevalent factors like gender, age, income, education, ethnicity, parental status, and political preference. In a social network, these influencing factors are used to explore the demographic constitution of a sample of social media users.

The proposed method has an ability to utilize a distantly trained regression model providing classification accuracy that was competitive with a fully-supervised approach. This denotes that the manner in which user behavior is comprehended can get enhanced by the consideration of the demographic composition of users in a network.

The proposed approach of pairing web demographic data with user data was able to train a model for demographic inference without annotation of individual profiles and procured realistic results in terms of the textual evidence as well as network evidence of participants of the network. The results aim to be accurate and reflect true user behavior in agreement with common stereotypes.

As future work, the demographic parameters of the proposed approach can be altered in a manner so as to include new categories in existent parameters or to introduce new demographic parameters in addition to the existent ones so as to alter the inference system and analyze the modified results.