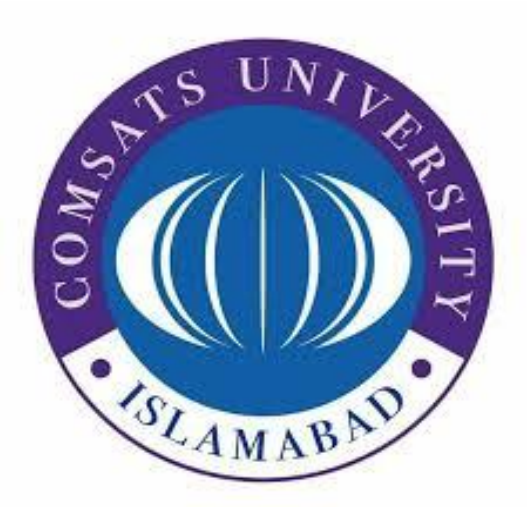
Celebrity Profiling on Short Urdu Text using Twitter Follower's Feed



By

Muhammad Hamza

CIIT/SP21-RCS-022/LHR

MS Thesis

In

"Computer Science"

COMSATS University Islamabad

Lahore Campus - Pakistan

Fall 2022

COMSATS University Islamabad

Celebrity Profiling on Short Urdu Text using Twitter Follower's Feed

A Thesis Presented to

COMSATS University Islamabad, Lahore Campus

In partial fulfillment

of the requirement of the degree

MS (CS)

By

Muhammad Hamza

CIIT/SP21-RCS-022/LHR

Fall 2022

Celebrity Profiling on Short Urdu Text using Twitter Follower's Feed

A Post Graduate Thesis submitted to the Department of Computer Science as partial fulfillment of the requirement of the award of Degree of MS (CS).

Name	Registration Number
Muhammad Hamza	CIIT/SP21-RCS-022/LHR

Supervisor

Dr. Muhammad Sharjeel
Assistant Professor,
Department of Computer Science
COMSATS University Islamabad, Lahore Campus

Submission Date: 2nd January 2023

Declaration

I Muhammad Hamza, CIIT/SP21-RCS-022/LHR hereby declare that I have produced the work presented in this thesis, during the scheduled period of study under the guidance and supervision of Dr. Muhammad Sharjeel AND Dr. Touseef Tahir. I also declare that I have not taken any material from any source except referred to wherever due that amount of plagiarism is within acceptable range. If a violation of HEC rules on research has occurred in this thesis, I shall be liable to punishable action under the plagiarism rules of HEC.

January 02, 2023

Signature of the Student

_____

Muhammad Hamza
CIIT/SP21-RCS-022/LHR

Certificate

It is certified that Muhammad Hamza, CIIT/SP21-RCS-022/LHR has carried out all the work related to this thesis under my supervision at the Department of Computer Science COMSATS Institute of Information Technology, Lahore campus and the work fulfills the requirement for the award of MS Degree.

Date: 2nd January, 2023

Supervisor:

_____

Dr. Muhammad Sharjeel

Assistant Professor

Head of Department:

_____

Dr. Muhammad Waqas Anwar

Associate Professor

Department of Computer Science,

COMSATS University Islamabad, Lahore Campus

DEDICATION

Dedicated to my loving “Parents”
and siblings.

ACKNOWLEDGEMENT

All praised are for Almighty Allah, the Most Gracious and Most Merciful, who gave me the strength to complete this research work. Nothing could have been possible without His blessings.
I would like thanks to my supervisor Mr. Muhammad Sharjeel and co-supervisor Dr. Touseef Tahir whose supervision, guidance, and mentorship help me out to complete my thesis work. Their hard work and dedication for the supervisee is incredible and led me to have success in the working of this research work. I have learned a lot from them, and I acknowledge their corporation to the fullest. I feel humbled and thankful for working under their supervision.
I am feeling grateful to admit that I owe my achievements to my loving parents. I would like to mention my truly loving father Pervaiz Ahmed, who has guided me at every dark moment in my life, who has helped me to achieve the accomplishments. I am very thankful to have a wonderful and sincere mother, who believed in me and stood firm with me at every moment of my life. I am also acknowledging the support of my siblings throughout the journey of this degree.
Furthermore, I extend my gratitude to my classmates and my friends specially Rana Rizwan, Areej Ikram, Inzmam , Ali and Kinza Tasleem who have helped me to complete this research work.

Muhammad Hamza
CIIT/SP21-RCS-022/LHR

ABSTRACT

Celebrity Profiling on Short Urdu Text using Twitter Follower's Feed

It is rightly said that current age is a digital age and social media shares a crucial chunk of it. People used to communicate, interact, and build relationships through social media. Celebrities are prolific authors and most of their personal information is public knowledge. There are some digital celebrities who exist only on social media, e.g., Twitter. Twitter is a social networking service which provides general populace as well as celebrities to interact with their fans. The demographics of celebrities could be predicted by the text of their followers as both shares same interest. However, most of the work on celebrity profiling has been performed on English and other similar languages except Urdu.

On the contrary, majority of the sub-continent celebrities and their fans tweets in Urdu. To fulfill this gap, in this research work we used Urdu tweets (short text) of 10 followers of a celebrity to build the first celebrity profiling based on followers' tweets corpus. Furthermore, the corpus was preprocessed, and Machine Learning (Logistic Regression, Support Vector Machines etc.) and Deep Learning (CNN, LSTM etc.) algorithms were used to train models for the prediction task. The trained model will be evaluated using state-of-the-art evaluation measures, i.e., precision, recall, and F1. The accuracy of the demographics of the celebrities are as follow; for the age the cumulative cRank is 0.45 , profession has the accuracy of 0.4, while the gender has cRank 0.65 and finally the cRank of fame is 0.45.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

NLP	Natural Language Processing
POS	Part of Speech
CP	Celebrity Profiling
AP	Author Profiling
CNN	Convolutional Neural Network
ML	Machine Learning
DL	Deep Learning
SLR	Systematic Literature Review
SVM	Support Vector Machines
RNN	Recurrent Neural Networks
LSA	Latent Semantic Analysis
SPM	Standard Progressive Matrices
RF	Random Forest

Chapter 1

Introduction

1.1 Introduction

Author profiling is the probing of a given text to identify the traits of an author like age, gender, and occupation based on author's writing style and features of written content [1]. Celebrity profiling is a sub-type of author profiling applied to only the celebrities to find their demographics [2]. Author profiling is analysis of an author (includes all authors) but celebrity profiling is to probe author who is only celebrity. There are two types of author profiling; first one is same-genre and second one is cross-genre. Training and testing models on the single genre (e.g., Twitter) is same-genre whereas training model on one genre (e.g., Twitter) and testing it on another (e.g., Facebook) is cross-genre. Celebrity profiling is used in forensics of linguistics, bot detection, author identification and influence tracing and marketing [3]–[5].

This behavior of celebrities and their followers is helpful in collecting written text that genuinely belong to a person. Consequently, research community have made use to of Twitter, where people use short tweets, to predict different demographics of a person like fame, age group, gender, and occupation [7]. However, majority of this work has focused on English and other languages [8]. On the other hand, most of the sub-continent celebrities and their followers use Urdu as a primarily language when they tweet. Urdu has a very different writing style and a rich morphological structure [9], [10]. Moreover, the task becomes more challenging as tweets are very short texts.

In this research work, we have used short Urdu tweets of a celebrity followers to predict certain demographics of that celebrity, e.g., age, gender, occupation, and fame. It has been observed that the followers of a celebrity have the same likes/dislikes and common interests [5]. Consequently, it is quite possible to judge the demographics of a celebrity based on its follower's feed. All the tweets of 10 followers of a celebrity would be collected, preprocessed, and Machine Learning (Logistic Regression, Support Vector Machines etc.) and Deep Learning (CNN, LSTM etc.) algorithms were used to train models for the prediction task, for this research task, celebrity profiling techniques cannot be applied to author profiling because normal author doesn't have any common interest with their followers [7], [8]. The trained model was evaluated using state-of-the-art evaluation measures, i.e., precision, recall, and F1. To the best of our knowledge, this is the first research work that was short Urdu text (tweets) from Twitter to predict age (between 20 to 60), gender (male, female), occupation (sports, politics, performer, and content creator), and fame

(high, low) of a sub-continent celebrity. The secondary objective of this research work is to foster further research in an under resourced language, i.e., Urdu.

1.2 Author profiling

Author profiling is a technique to identify the demographics of certain authors like age, profession, and gender. It is done by investigating a text written by an author. The text written by an author depicts his/her personality and it constitutes different aspects of the author's personality. The written text could be a long text like a book and a short text which has the limit on its length like tweets.

1.3 Celebrity Profiling

There exists a different type of authors, celebrities are one of them. To probe the demographics of the celebrities by investigating their written text is called the celebrity profiling. The life of a celebrity is public, most of the celebrities have their presence on the internet/ social media to interact with their followers and fans. The content and text posted by celebrities on social media and internet can be used for analysis of their demographics.

1.4 Role of Followers in celebrity profiling

The followers of a celebrity share the same interest with the author, for example a person have interest in politics always follows the politicians who are celebrities. In fact, the young followers like the young politicians because their views and energy match with each other. So, the followers of a celebrity play a key role and one can probe the demographics of a celebrity by the analysis of his/her followers, and this is the focus of this research.

1.5 Celebrity profiling on social media

Celebrities use to have their presence on every social media platform. Every social media platform has different kind of representation of one's view, for example on Facebook a celebrity can post as long as one can but on Twitter, the lengths of tweets is fixed. This research focus on tweets which has short text of celebrities' followers who use to tweets in Urdu.

1.6 Methods of Celebrity profiling

There are methods of doing celebrity profiling which are as follow:

- Same genre
- Cross genre

In same genre, the training of model is taken place on the same genre as if model is train on Twitter, it will be tested on Twitter. In cross genre, the training of the model takes place on one genre as Twitter, and it is tested on another like Facebook.

1.7 Applications of Author profiling

There are a lot of applications of celebrity profiling some of them are stated below:

Forensics

The author profiling helps to find an anonymous author and it can detect the forged text from the original content. It will help to identify the motivation and author's demographics like age, gender, and profession. It also helps in linguistic forensics to find the original author of a content in case of conflict.

Bot Detection

Author profiling is used to identify the bots on social media like on Twitter which may incur false influence on the audience by giving positive reviews on certain product, polarizing the political situation like in 2016's presidential elections of USA were influenced by the social media bots. Therefore, it is an urgency to identify the bots on social media.

Marketing

In marketing, author profiling plays an important role to identify the likes and dislikes of certain group of audience and their demographics. The companies target to the specific audience with certain demographics for better reach of their product.

1.8 Problem Statement

Celebrity profiling is a sub-branch of author profiling applied on celebrities. Applications of celebrity profiling include advertisement, security, risk assessment, and forensics. There are 88.7 million Urdu speaking community across the globe. A large chunk of sub-continent celebrities use twitter to connect with their fans. Most of them tweet in Urdu language and fans who follow them also tweet in the Urdu language. Moreover, fans share some common interests and likes/dislikes with the celebrities they follow. Presumably, when these followers tweet, they might reveal some demographic or social info about the celebrity they follow.

The aim of this thesis work is to investigate the predictive nature of socio-linguistic attributes found in short Urdu tweets on the demographics of celebrities. The main objective is to predict the demographics (age, gender, occupation, and fame) of a celebrity based on tweets (short Urdu text) of their followers using Machine Learning and Deep Learning methods.

1.9 Research Objectives

Following are the main objectives of the research work.

- Explored the celebrity profiling task for short Twitter text in the Urdu language.
- Developed a celebrity profiling corpus using tweets in the Urdu language.
- Analyzed the demographics (age, gender, fame, occupation) of the sub-continent celebrities.
- Used the information of demographics of celebrities and their followers for advertising, forensics, and to prevent hate speech in Urdu speaking community.
- Utilized the proposed technique to curb the cybercrimes and hate speech on social media.

1.10 Thesis Focus

The focus of the work is analyses of demographics of celebrities based on the tweets of their followers in Urdu language. It will help to identify the characteristics of unknown author to decrease the cybercrimes, it will also help in detecting the bots, advertisement, and negative influence tracing on social media.

1.11 Thesis Outline

The Organization of the thesis is going to state in this section. Chapter 2 will cover the Literature Review which includes the work of different authors carrying out the celebrity profiling and working of their learning techniques, corpora organization, features extraction techniques and evaluation measures. Chapter 3 is about collection of the corpus and compilation of the corpus of celebrities from Twitter. Chapter 4 will cover the training model which includes the Machine learning and deep learning models, it will also describe the evaluation measures. Chapter 5 covers the summary of the entire thesis, and the Chapter 6 is references used in the research work.

Chapter 2

Literature Review

2.1 Literature Review

The literature review from the different research community who has conducted the research on celebrity/author profiling is explained in this section.

In [6], different demographics like year of birth, gender, fame, and occupation of the celebrities are predicted based on their data available on twitter as they are most prolific personalities having a lot to share with their fans. Eight various models were evaluated and predicted the demographics of 48,335 celebrities who tweeted in English language. It worked best on predicting binary gender and fame of the celebrity along with occupation and age. The model results less accurate on predicting the non-binary gender and occupation that is not specified or single topic like science and manager. It predicts the medieval celebrities very well ranging from 1980s to 2000 (ages between 20 to 40). However, the results were less accurate on younger and older celebrities.

An information retrieval based method termed as TF-IDF based on n-grams and bigrams applied on character level is used in [7] to predict the gender, age, fame, and occupation of a celebrity. This model grouped the users in aforementioned four categories, based on given data set of tweets of a certain user. Age ranges from 1940 to 2012, fame and gender have three classes each, and occupation has eight classes.

In [8], the demographics of celebrities are identified by examining their ten followers rather than using the personal data of celebrities. Dataset of 2,380 authors were created to apply three different models including TF-IDF, LSTMs and n-grams features to solve the task in different ways. The evaluation results (F1-score) were very efficient in spite of random guessing. Follower-based model has many weaknesses and strengths like it works best on coherent classes like ‘sports’ and work less efficient on diverse like ‘creators’. Identifying age of the celebrity was found to be very difficult by this model.

Socio-linguistic approach is used to classify celebrities in [9]. There are certain challenges to finding birth year class, which produces better results in small dimensionality (ten years) but less accurate when applied to bigger cluster. There was a huge dataset of 53 million tweets used for the classification task. However, to process such huge data was a big problem. Moreover, dataset used for training had a huge imbalance in classes.

In [10], a multi-lingual analytic engine was introduced to detect the gender, age, and personality traits of an author by using Machine Learning and Natural Language Processing techniques [4]. The corpus consists of tweets in four languages Dutch, Spanish, English, and Italian. Methodology generates new features based on linguistic processing. Machine Translation is used for languages having less data. Some socio-demographic parameters faded away by Machine Translation but native language results were promising. Sentiment analysis was used to mitigate the problem of socio-demographics signal lost.

Tweets of anonymous authors were given to identify the author in [11]. Author profiling on anonymous text has a wide variety of applications like forensics, security, author identification and advertising. Companies need to know the interest and disliking of the customers on the basis of blogs and comments on socials media. Personality identification and linguistics were calculated by using n-gram and stylistic based methods. Age and combination of age and gender calculations had problems due to unbalanced dataset. Conversations were introduced to improve the results of gender identification by sexual predators.

In [12], different demographics of author were determined in cross-genre environment like age and gender in three different languages Dutch, English, and Spanish. Support Vector Machines (SVM) and Multinomial Naïve Bayes (MNB) techniques were used to detect different demographics of the author. Different languages had the different optimal configurations. Same pre-processing steps were introduced for the three languages. Average accuracy reported for English, Spanish, and Dutch is 75, 90, and 90 percent respectively.

In [13], the authors try to identify the gender of an author using a variety of languages including English, Arabic, Portuguese, and Spanish. Deep Learning algorithms such as Recurrent Neural Networks (RNN) and Convolutional Neural Network (CNN) were used for the prediction task. Gender verification had the up to 78% accuracy and language variety prediction stands at up to 97% accuracy.

In [14], anonymous data from emails is collected for the identification of different demographics of an author. The corpus consists of 9836 emails. Demographics to be predicted were age, gender, level of education, country, and native language. Machine Learning algorithms SMO, Random Forest, and LibSVM were used in the study.

In [15], the authors tried to predict the age and gender of an author based on its writings. The corpus included data from four different genres, i.e., Twitter, social media, hotel review, and blogs. Content-based and stylistic-based features were approached which includes n-gram, term vectors, frequencies, POS, and readability measures. Average accuracy was 80% on predicting the gender and age.

In [3], identifying the author demographics based on both image and text data in three different languages (Arabic, English and Spanish) were used. The textual data was collected from Twitter [16]. A total of 23 participants were evaluated using SVM and Logistic Regression algorithms. The reported accuracy was above 80% on average.

The information about socioeconomic attributes of the users of the social media like occupation and income are the main problems in computational science. In [17], probing of these demographics has many applications like personality recommending, political campaigning and targeted marketing. The demographics were evaluated by Support Vector Machines (SVM) and Gaussian Process Classifiers. The accuracy of predicted attributes was 50.54%.

Social media and user network information are useful for the geotagging on the online platform like Twitter. The tweets are the essential tools for the detection of the events and the enrichment of the events. In [18], a hybrid Gaussian mixture models are introduced to map the spatial attributes of the model with the accuracy of 85%.

In [19], political biasedness is evident in the media reporting and it is an important issue to tackle with in contemporary world. To identify the biasedness, fake news and propaganda of the media outlets, a novel approach is introduced which uses SVM model to predict the political biasedness with the accuracy up to 85.29% on different social media platforms.

In [20], the verification of the author is checked by the analyzing two texts. A subordinate task is performed that is called author obfuscation: verification prevention of an author by changing the to be obfuscated text. To test the obfuscation technique, SVM classifier is used and accuracy of 86.13% were achieved.

Gender identification on the social media is the important research area. In [21], The discrimination of the gender on the Twitter is detected by analyzing the tweets of the targeted audience. Gender discriminators is detected by the SVM model with the accuracy of 76%.

In [22], the demographics of the author and aggressiveness is detected on Mexican Spanish language tweets. The demographic of author includes his/her occupation and place of residence and aggressiveness determines between tweets weather they are aggressive or not. Classification is done using SVM and performance measures using the F1 scores.

In [23], the identification of the gender on Tweeter by using the three different languages; English, Spanish and Arabic. The Model uses the Latent Semantic Analysis (LSA) for dimensionality reduction and SVM for the classification. The accuracy is 82.21%, 82%, and 80.9% respectively in English, Spanish and Arabic.

Social media user profiling drew attention of the researchers due to its wide range applications like advertising, marketing and recruiting. In [24], a deep learning fusion technique is proposed which analysis the multiple sources of modalities like text or image user data. Accuracy was reported at 95% for prediction of the gender.

To identify the different traits of the author is termed as the author profiling and it has many applications like identification of hate speech and advertising. In [25], author profiling is being performed in Asian language Roman Urdu and English by developing a multilingual corpus. Stylistic based features were extracted using the N-gram model.

Social media automated personality prediction has gained attention of research community due to its diverse applications in Natural language processing and data science. In [26], the dataset of a diverse and huge audience of Reddit is collected and automatically extracted the features of the author. Support Vector Machine is used for the classification.

In [27], automatic identification of the identification of the gender based on the written text is carried out on 920 documents labelled for author gender. The identification of the author's gender on an unseen document is carried out by more than 80% accuracy. Simple lexical and syntactic features of text categorization were used to identify the gender of the author. K-nearest-neighbor, neural nets and SVM is used as learning methods.

In [28], Facebook likes data of 58k people were used to accurately and automatically prediction of the sensitive traits of the author like political views, gender, age, ethnicity, social connections and religious views. Attributes of the questionnaire was measured using Raven's Standard Progressive Matrices (SPM) with the accuracy of gender prediction was 62%, relationship status was 49% and religion 90%.

In [29], personality is predicted based on the language and it has many applications in data science and natural language processing. The model that proposed is hierarchal model having message and word-based attention to segregate the high- and low-level messages. It can also use for checking the relevance of two different documents. CNN and max pooling are used to predict the personality traits with up to 75% accuracy.

In recent years, social media grows rapidly and the violent and hate speech on different platforms of the social media proliferated. In [30], People who have same stereotypes and common interests tends to be abusive on social media platform. The data set is 16k tweets of the twitter users of different domains. node2vec framework[16] and LSTM are used for author profiling.

Table 1:Literature Review

Ref	Author	Year	Method	Corpus Size
[31]	Abigail Hodge, Samantha Price	2020	logistic regression, random forest and support-vector	48,335
[6]	Matti Wiegmann, Benno Stein	2019	language agnostic	49,000
[7]	Victor Radivchev, Alex Nikolov	2019	TF-IDF, n-gram	N/A
[8]	Matti Wiegmann,Benno Stein	2020	macro-averaged multi-class F1	2,380
[9]	Luis Gabriel Moreno-Sandoval , Edwin Puertas	2019	N-gram	N/A
[10]	Scott Nowson, Julien Perez, Caroline Brun	2015	language agnostic	30,000
[11]	Francisco Rangel,Paolo Rosso	2013	N-gram	236,600

[16]	Ilia Markov, Helena Gómez-Adorno,	2016	LibSVM and liblinear classifiers	N/A
[14]	Dominique Estival, Tanja Gaustad	2006	SMO, Random Forest and LibSVM	9836

Chapter 3

Proposed Urdu Corpus for Celebrity Profiling

3.1 Introduction

This chapter covers the corpus generation steps for the specified celebrity profiling task from the Tweets of their followers. It also explains the process that improve the quality of the corpus. The chapter is divided into three parts, first part explains the sources of the collection of the data and generation of the corpus. The second part explain the collection of the data from the source and the integration of the Tweets of the followers. The third part explains the pre-processing of the Tweets of the followers, it includes the extraction of the tweets which has the Urdu language and removal of the special characters and alphabets.

3.2 Corpus Generation Process

The corpus is termed as huge collection of the data in text form, specifically this term used in Natural Language Processing (NLP). The corpus is collected according to the requirements of the problem. The specified patterns in the corpus are find out according to the research focus by applying machine learning and deep learning algorithms. There are two types of data in the corpus generation, the one is annotated and other is unannotated. The output is associated with the data termed as annotated data. In this research work, the annotated data is collected to form a corpus. Supervised Learning is associated with the annotated data, the research carried out supervised learning as the data in the corpus is annotated. The Tweets of the ten followers of a specific celebrity is collected and their demographics are also associated with that data to make it a supervised corpus.

3.3 Data Collection

The main contribution of this research work is to build a high-quality data for the celebrity profiling task in Urdu language. There exist 270 million Urdu speaking community who follows their celebrities on different platforms. Celebrities interact with their followers through Twitter which is the most famous and authentic social media platform. So, there is a huge research gap, there is need to develop a state-of-the-art data set for the celebrity profiling to fill that gap. In this research, a corpus is collected which includes the data of celebrities and their followers to probe the demographics of the celebrities who tweet in Urdu and their followers who share the same linguistics and interests. For that purpose, the tweets of ten followers each celebrity is collected to probe the demographics of each celebrity.

The issue that has been faced during the research was the collection of the data. The data of the sub-continent celebrities which has their language Urdu, and they are active on twitter with substantial followers that has been used to predict their demographics. The searching for the celebrities was the hard task, most of the celebrities use to tweet in the English language. There are marginal celebrities who has tweeted in Urdu, and they have a substantial number of celebrities which also use to tweet in Urdu. Most of the followers of these celebrities use to tweets in different languages, there was hard to find those followers who use to tweet in Urdu language.

The second problem which is faced during the research was to find the suitable algorithms and feature extraction methods. For that, an extensive literature review has been done on the author and celebrity profiling and there have been chosen some algorithms to apply on the data for the celebrity profiling. Most prominent algorithms which have the highest accuracy on author profiling are decision tree and random forest. There were two features extraction techniques were applied on the data, namely TD-IDF and length of the tweet.

An example of tweets of the followers of a celebrity is shown in fig:

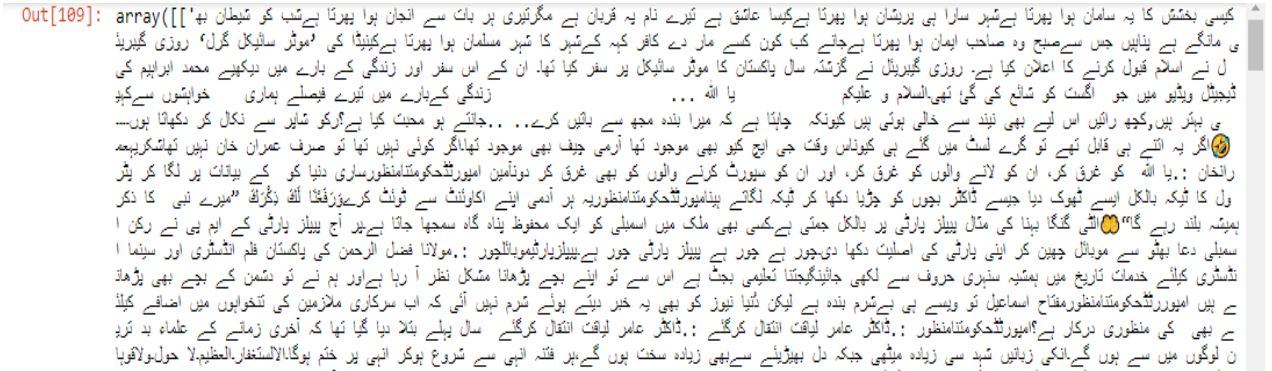


Figure 1:An example of tweets of a celebrity's' follower

There involve three main processes to download and collect the data of the celebrity and their followers which are stated as:

3.3.1 Downloading the tweets of the followers

To collect the data of followers from tweeter was the challenge, as tweeter only allow developers to download the tweets. Verification of the developer account is very difficult and involves some hidden process and policies. There is another way to download the tweets is third party websites. The tweets collected for the research is from third party website called 'VICINTAS'. The usernames of selected celebrities' followers were collected and given to website to get the data in Excel format. An interface of the website homepage is shown in fig:

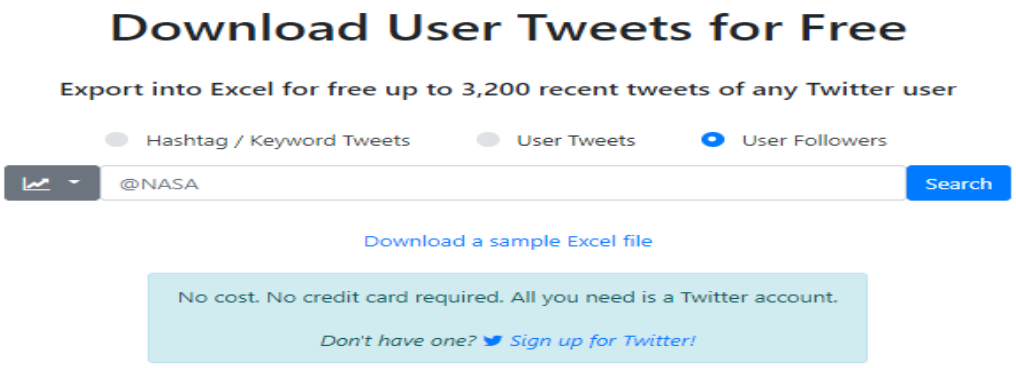


Figure 2: 'VICINITAS' website interface for downloading tweets

3.3.2 Collection of demographics of celebrities

To collect an annotated corpus besides the collection of the tweets of the followers, the collection of the demographics of the celebrities is also required. The demographics of the celebrities were collected from Tweeter and Google. Some of the demographics as age was hard to find on Tweeter and it was collected from the Google. The occupation and gender were collected from the Tweeter.

3.3.3 Integration of data

To collect the corpus, the tweets of ten followers of each celebrity is collected and integrated in one file. Each file contains the Urdu tweets of the ten followers of each celebrity and the minimum tweets of each celebrity were twenty. Table 3.1 depicts the statistics of the corpus:

Table 2: General statistics of corpus

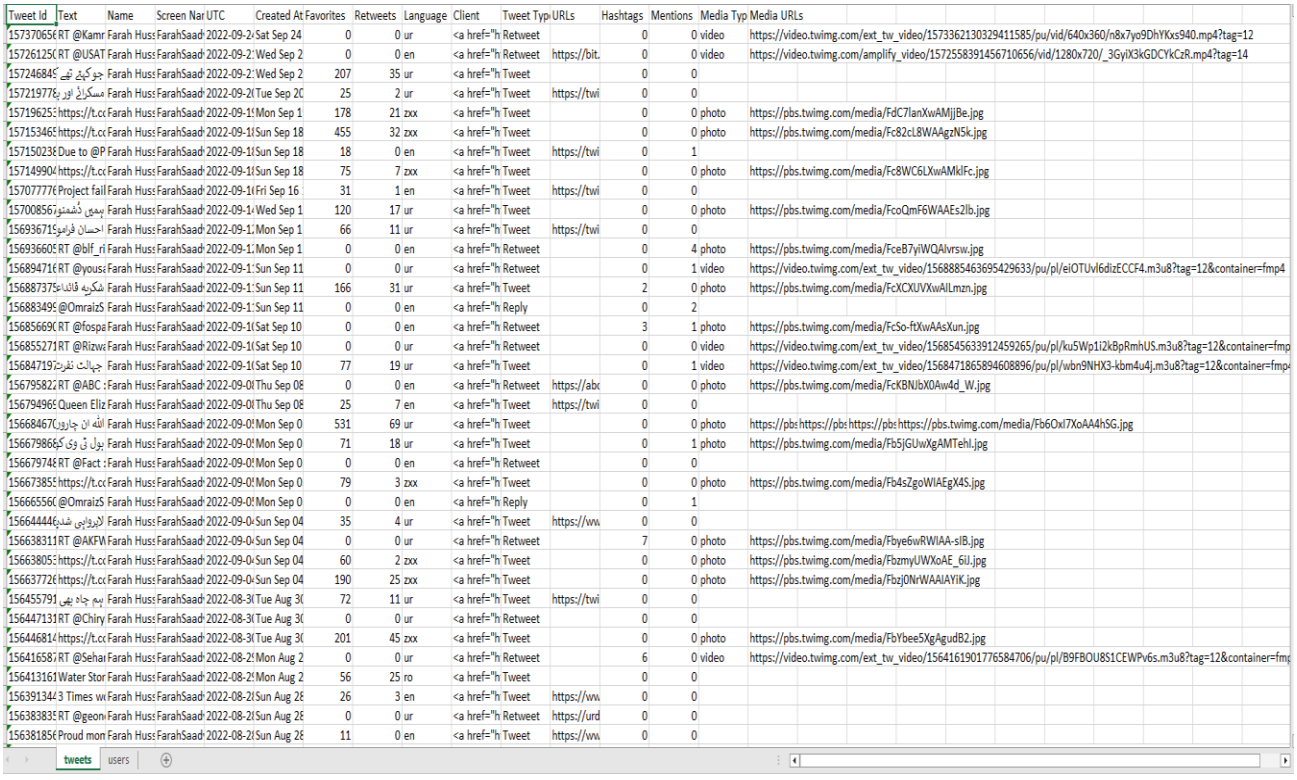
Total Number of celebrities	100
Followers of each celebrity	10
Total number of followers	1000

Minimum Number of Urdu tweets of a follower	20

3.4 Data Pre-processing and Normalization

To perform the author profiling on the celebrity tweets downloaded from the third-party website were in excel format. The data in excel format includes the tweets id, tweets, Name of the owner, date, Retweets of certain tweet, language, URL, Hashtags, Media Type (Audio, Video, or text). Data in the excel file includes all the tweets of the celebrity or its follower, it includes the tweets and retweets as well, video tweets and English tweets. To understand the data was the problem to solve, as for the require task the tweets which has only Urdu language are required. Furthermore, the tweets included URLs, emojis and special characters that was not required for the required task.

The snapshot of the excel file is shown in the Figure:



Tweet Id	Text	Name	Screen Nar	UTC	Created At	Favorites	Retweets	Language	Client	Tweet Typ	URLs	Hashtags	Mentions	Media Typ	Media URLs
15737065	RT @Kamr	Farah Huss	FarahSaad	2022-09-2	Sat Sep 24	0	0	ur	<a href="h	Retweet		0	0	video	https://video.twimg.com/ext_tw_video/1573362130329411585/pu/vid/640x360/n8x7yo9DhYKxs940.mp4?tag=12
15726125	RT @USAT	Farah Huss	FarahSaad	2022-09-2	Wed Sep 2	0	0	en	<a href="h	Retweet	https://bit.	0	0	video	https://video.twimg.com/amplify_video/1572558391456710656/vid/1280x720/_3GyiX3kGDCYkCzR.mp4?tag=14
15724684	جو کہتے تھے	Farah Huss	FarahSaad	2022-09-2	Wed Sep 2	207	35	ur	<a href="h	Tweet		0	0		
15721977	مسکراتے اور ہ	Farah Huss	FarahSaad	2022-09-2	Tue Sep 2(	25	2	ur	<a href="h	Tweet	https://twi	0	0		
15719625	https://t.co	Farah Huss	FarahSaad	2022-09-1	Mon Sep 1	178	21	zxx	<a href="h	Tweet		0	0	photo	https://pbs.twimg.com/media/FdC7lanXwAMjjBe.jpg
15715346	https://t.co	Farah Huss	FarahSaad	2022-09-1	Sun Sep 18	455	32	zxx	<a href="h	Tweet		0	0	photo	https://pbs.twimg.com/media/Fc82cL8WAAgzN5k.jpg
15715023	Due to @P	Farah Huss	FarahSaad	2022-09-1	Sun Sep 18	18	0	en	<a href="h	Tweet	https://twi	0	1		
15714990	https://t.co	Farah Huss	FarahSaad	2022-09-1	Sun Sep 18	75	7	zxx	<a href="h	Tweet		0	0	photo	https://pbs.twimg.com/media/Fc8WC6LXwAMklFc.jpg
15707777	Project fail	Farah Huss	FarahSaad	2022-09-1	Fri Sep 16	31	1	en	<a href="h	Tweet	https://twi	0	0		
15700856	ہمیں دُشمنو	Farah Huss	FarahSaad	2022-09-1	Wed Sep 1	120	17	ur	<a href="h	Tweet		0	0	photo	https://pbs.twimg.com/media/FcoQmF6WAAEs2lb.jpg
15693671	احسان فراموٹ	Farah Huss	FarahSaad	2022-09-1	Mon Sep 1	66	11	ur	<a href="h	Tweet	https://twi	0	0		
15693660	RT @blf_ri	Farah Huss	FarahSaad	2022-09-1	Mon Sep 1	0	0	en	<a href="h	Retweet		0	4	photo	https://pbs.twimg.com/media/FceB7yiWQAIvrsw.jpg
15689471	RT @yousc	Farah Huss	FarahSaad	2022-09-1	Sun Sep 11	0	0	ur	<a href="h	Retweet		0	1	video	https://video.twimg.com/ext_tw_video/1568885463695429633/pu/pl/eiOTUvl6dizECCF4.m3u8?tag=12&container=fmp4
15688737	شکریہ قائداع	Farah Huss	FarahSaad	2022-09-1	Sun Sep 11	166	31	ur	<a href="h	Tweet		2	0	photo	https://pbs.twimg.com/media/FcXCXUVXwAILmzn.jpg
15688349	@OmraizS	Farah Huss	FarahSaad	2022-09-1	Sun Sep 11	0	0	en	<a href="h	Reply		0	2		
15685669	RT @fospa	Farah Huss	FarahSaad	2022-09-1	Sat Sep 10	0	0	en	<a href="h	Retweet		3	1	photo	https://pbs.twimg.com/media/FcSo-ftXwAAsXun.jpg
15685527	RT @Rizwa	Farah Huss	FarahSaad	2022-09-1	Sat Sep 10	0	0	ur	<a href="h	Retweet		0	0	video	https://video.twimg.com/ext_tw_video/1568545633912459265/pu/pl/ku5Wp1i2kBpRmhUS.m3u8?tag=12&container=fmp
15684719	جہالت نفرت	Farah Huss	FarahSaad	2022-09-1	Sat Sep 10	77	19	ur	<a href="h	Tweet		0	1	video	https://video.twimg.com/ext_tw_video/1568471865894608896/pu/pl/wbn9NHX3-kbm4u4j.m3u8?tag=12&container=fmp4
15679582	RT @ABC :	Farah Huss	FarahSaad	2022-09-0	Thu Sep 08	0	0	en	<a href="h	Retweet	https://abc	0	0	photo	https://pbs.twimg.com/media/FcKBNJbX0Aw4d_W.jpg
15679496	Queen Eliz	Farah Huss	FarahSaad	2022-09-0	Thu Sep 08	25	7	en	<a href="h	Tweet	https://twi	0	0		
15668467	اللہ ان چاروں	Farah Huss	FarahSaad	2022-09-0	Mon Sep 0	531	69	ur	<a href="h	Tweet		0	0	photo	https://pbs https://pbs https://pbs https://pbs.twimg.com/media/Fb6Oxl7XoAA4hSG.jpg
15667986	بول ٹی وی کی	Farah Huss	FarahSaad	2022-09-0	Mon Sep 0	71	18	ur	<a href="h	Tweet		0	1	photo	https://pbs.twimg.com/media/Fb5jGUwXgAMTehI.jpg
15667974	RT @Fact :	Farah Huss	FarahSaad	2022-09-0	Mon Sep 0	0	0	en	<a href="h	Retweet		0	0		
15667385	https://t.co	Farah Huss	FarahSaad	2022-09-0	Mon Sep 0	79	3	zxx	<a href="h	Tweet		0	0	photo	https://pbs.twimg.com/media/Fb4sZgoWIAEgX4S.jpg
15666556	@OmraizS	Farah Huss	FarahSaad	2022-09-0	Mon Sep 0	0	0	en	<a href="h	Reply		0	1		
15664444	لاہور ہوائی شدی	Farah Huss	FarahSaad	2022-09-0	Sun Sep 04	35	4	ur	<a href="h	Tweet	https://ww	0	0		
15663831	RT @AKFW	Farah Huss	FarahSaad	2022-09-0	Sun Sep 04	0	0	ur	<a href="h	Retweet		7	0	photo	https://pbs.twimg.com/media/Fbye6wRWIAA-slB.jpg
15663805	https://t.co	Farah Huss	FarahSaad	2022-09-0	Sun Sep 04	60	2	zxx	<a href="h	Tweet		0	0	photo	https://pbs.twimg.com/media/FbzmyUWXoAE_6iJ.jpg
15663772	https://t.co	Farah Huss	FarahSaad	2022-09-0	Sun Sep 04	190	25	zxx	<a href="h	Tweet		0	0	photo	https://pbs.twimg.com/media/FbzJ0NrWAAlAYiK.jpg
15645579	ہم جاہ بھی	Farah Huss	FarahSaad	2022-08-3	Tue Aug 3(	72	11	ur	<a href="h	Tweet	https://twi	0	0		
15644713	RT @Chiry	Farah Huss	FarahSaad	2022-08-3	Tue Aug 3(	0	0	ur	<a href="h	Retweet		0	0		
15644681	https://t.co	Farah Huss	FarahSaad	2022-08-3	Tue Aug 3(	201	45	zxx	<a href="h	Tweet		0	0	photo	https://pbs.twimg.com/media/FbYbee5XgAgudB2.jpg
15641658	RT @Sehar	Farah Huss	FarahSaad	2022-08-2	Mon Aug 2	0	0	ur	<a href="h	Retweet		6	0	video	https://video.twimg.com/ext_tw_video/1564161901776584706/pu/pl/B9FBOU8S1CEWPv6s.m3u8?tag=12&container=fmp
15641316	Water Stor	Farah Huss	FarahSaad	2022-08-2	Mon Aug 2	56	25	ro	<a href="h	Tweet		0	0		
15639134	3 Times wi	Farah Huss	FarahSaad	2022-08-2	Sun Aug 28	26	3	en	<a href="h	Tweet	https://ww	0	0		
15638383	RT @geoni	Farah Huss	FarahSaad	2022-08-2	Sun Aug 28	0	0	ur	<a href="h	Retweet	https://urd	0	0		
15638185	Proud mon	Farah Huss	FarahSaad	2022-08-2	Sun Aug 28	11	0	en	<a href="h	Tweet	https://ww	0	0		

Figure 3: Excel file data before pre-processing

The preprocessing of the data from excel sheet required five main steps which are stated below, they include:

- Extraction of Urdu tweets
- Removal of Retweets
- Formation of text files
- Integration of tweets in pandas data frame
- Removal of emojis and special characters

3.4.1 Extraction of Urdu Tweets

Tweets in the Excel format are in many languages, it includes Urdu, English and Punjabi as well. One of the column name 'Language' specifies the type of the language of the tweets. The tweets that contain more Urdu content and its type in 'text' is termed as Urdu tweets as 'ur' in the language column. All the Urdu tweets were extracted from the file.

3.4.2 Removal of Retweets

The Retweets of the followers of a celebrity does not really depicts the thoughts and the character of the real author as it is not produced by the real author. Retweets may project the thoughts of the follower, but it does not help in the author profiling task. So, removal of the Retweets was necessary. For that purpose, the selected Urdu tweets were further filtered by the column of Tweet Type (Tweet or Retweet). The real tweets were collected by applying the filter.

Table 3: Tweets Data of a Follower

Total Tweets	3169
Number of Retweets	1194
Number of Tweets	1767
Number of Replies	208
Number of Urdu Tweets	1812
Number of English Tweets	1108

3.4.3 Formation of text Files

After the removal of the Retweets from all the tweets, the remaining Urdu tweets were required to be save in text format to get ready for the next integration of the data. Each tweet is separated from

the other line by line. The tweets of the ten followers of the celebrity are collected in the same file separated by their ids.

Figure 4: An example of text file containing preprocessed tweets

3.4.4 Integration of tweets in pandas data frame

To apply the models and to get ready the data for the modeling phase, the data must be in the format of data frame in python. All the files of Urdu preprocessed tweets were loaded in the pandas data frame and get ready for the training phase.

3.4.5 Removal of URLs, emojis and special characters

After loading the data to the data frame, the data required more preprocessing, it required to remove the stop words, URLs, emojis and special characters were need to remove for the better results, a snapshot of data before preprocessing containing the stop words, URLs and emojis is shown in Fig:

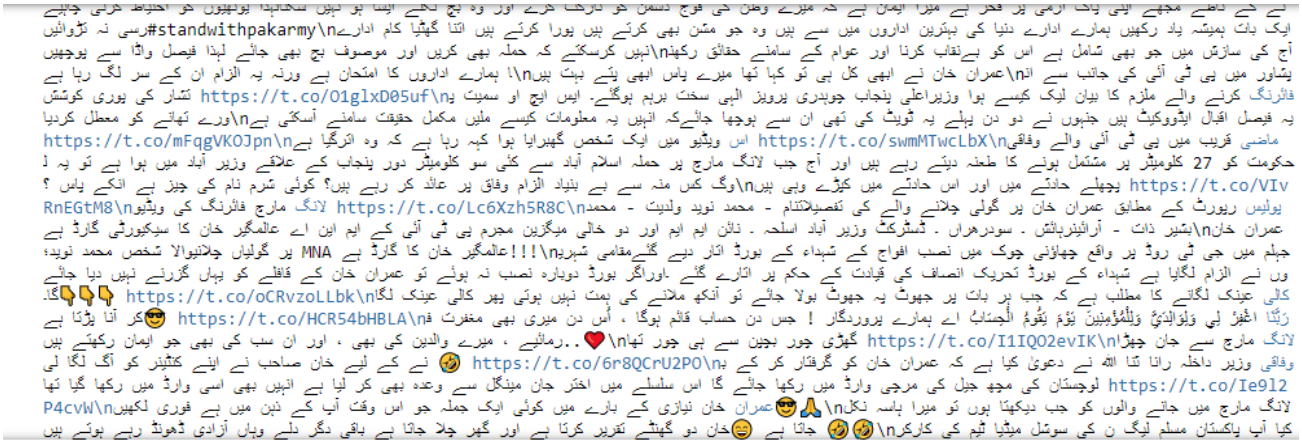


Figure 5: An example of tweets before preprocessing

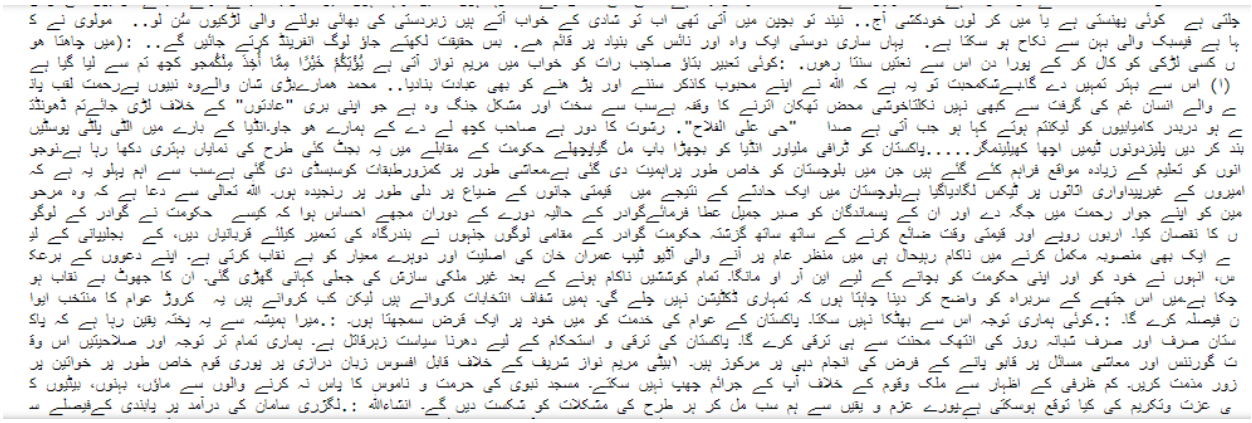


Figure 6: After Preprocessing of data example

Chapter 4

System Overview

4.1 Introduction

This chapter explains the system overview, it explains the model development, training, and evaluation techniques. The corpus built for the celebrity profiling based on the tweets of the followers will used for the development of the model. The model architecture and working are explained first, the training and development of the model is explained in the latter part of the chapter along with results and analysis.

4.2 Classification Machine Learning Techniques for Author Profiling

Author profiling is a classification problem except for identifying the age of the author and the classification techniques work excellent for the author profiling tasks. Classification techniques used to work best on the data where input and output are categorical but in case of the author profiling on Twitter, although tweets have a limited length, but the length is not fixed and this regards as the sequence [48]. The input is in the form of a sequence, but the output is in categorical format. Most of the classification machine learning algorithms perform better on the author profiling problems like SVC and Random Forest.

4.3 Text Classification Models for Celebrity Profiling

The models which are proposed for the classification of different traits of the celebrities have the same common architecture. The classical classification machine learning algorithms share the similar architecture. The input for the classifier is in the form of the tweets which have a short length text. The input is in the form of a sequence of text includes all the tweets of ten followers of a celebrity. The input, which is given to the algorithm convert into the collection of the feature vectors.

All the feature vectors filtered through the feature extraction techniques. Multiple feature extraction techniques were used to extract the features from the tweets. The feature vector techniques include length of tweets and count vectorizer. The change in the feature extraction method also effect the results of the algorithms.

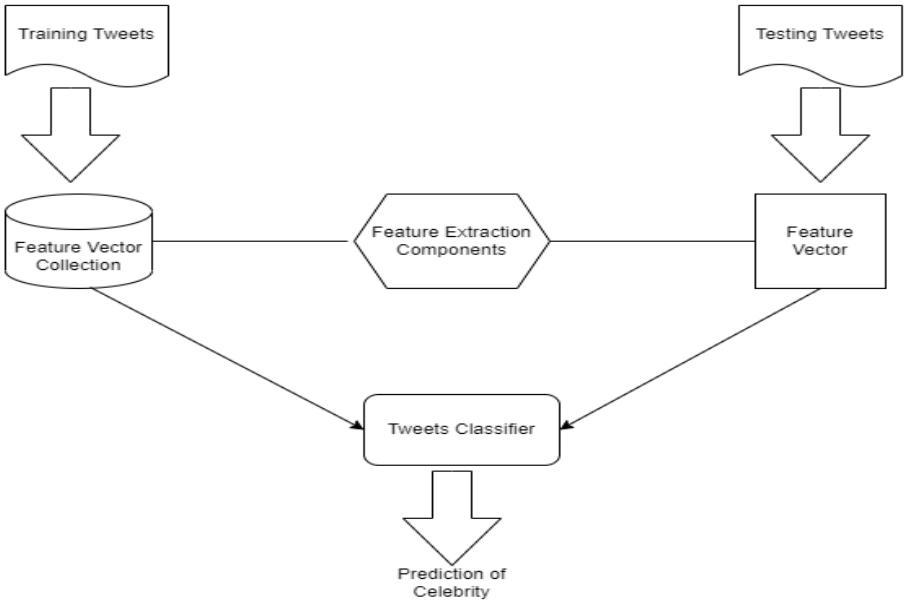


Figure 7: Celebrity Profiling text classification model architecture

Multiple classical machine learning algorithms were applied on the data. The algorithms include Logistic Classifier, Decision Tree, Random Forest, and Support Vector Classifier (SVC) [47]. For the prediction of the age, Logistic Regression and Support Vector Machine (SVM) were used to get the better results. The classification methodology for this research is depicted in the figure:

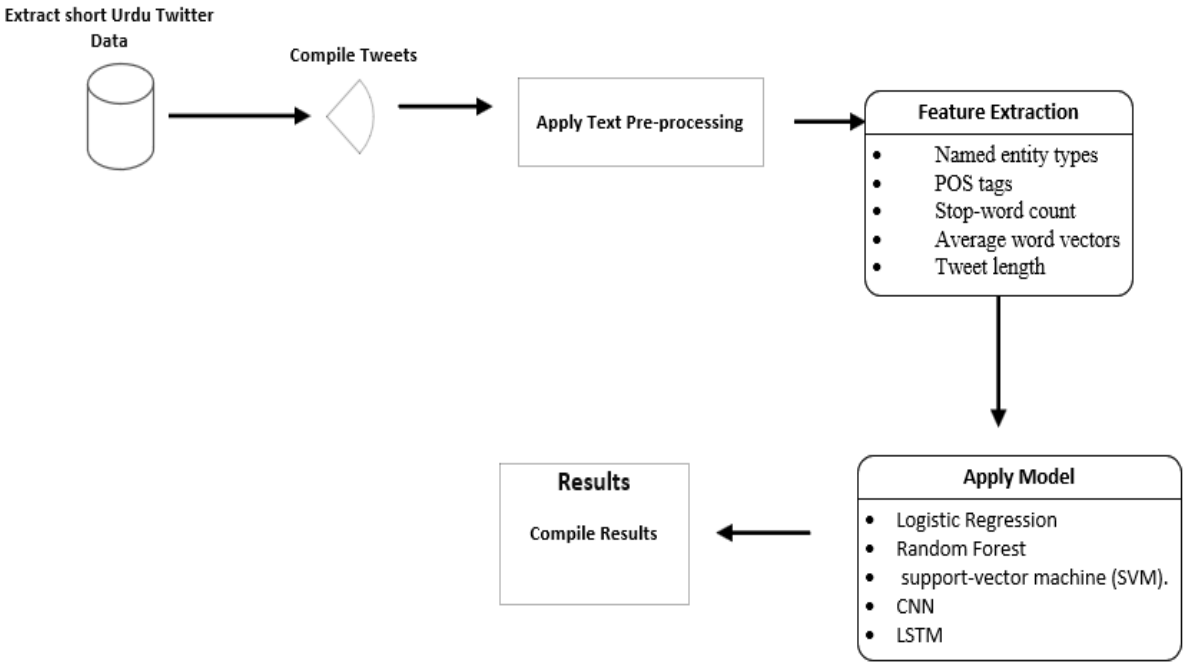


Figure 8: System Architecture Diagram

4.4 Demographics to be predicted of Celebrities

There is total four demographics are selected to be predicted by the tweets of the followers of the celebrities. The demographics includes age of celebrity, occupation, gender, and fame of the celebrity.

Age

In author profiling, age prediction is an important demographic, and it helps in forensics to find the age of an unknown author and in advertisement to target the audience with specific age. For the prediction of age, the range is from 20 to 80 years. The study shows that [30], prediction on authors below 20 years showed the results which are not promising. So, the age groups are divided into 3. First group is 20-40, second is 40-60 years and the last group is from 60-80 years.

Occupation

Occupation of the author is relevant for the purpose of advertisement, to target occupational group and for the optimization of social media for the same occupational groups. The occupation is categories into four categories named as politics, entertainment, journalism, and sports.

Gender

Gender identification in author profiling helps in finding the real author of an artifact. Gender identification helps to categories the gender and ultimately reduce the search space [46]. The gender is categorized into male and female.

Fame

The celebrities are categorized into three categories for the prediction of fame. Celebrities who have followers equal to or lower than 1 million are categorized into 'Rising' category. Celebrities who have followers between 1 and 2.5 million are categorized into 'Star' category and 'Super Star' category is for the celebrities having more than 2.5 million followers.

4.4 Feature Extraction Methods

For the classification in classical machine learning algorithms for author profiling, the feature extraction techniques have a significant important for the performance of the model. There are

several feature extraction techniques were used for the extraction of the features from the tweets data of the followers [32]**.** Term frequency–inverse document frequency (TD-IDF) is used to give weightage to the words, to check how important a word is in the document. TD-IDF is a vectorizer zo, it makes the vectors from the words in the document and then the algorithms learn from the vectors. The formula fo the TD-IDF is:

$$TF{-}IDF\ (t,d)=TF\ (t,d) * IDF(t)$$

$$IDF(t)=\log\ [\ n/DF(t)]+1$$

TF−IDF (t,d) is the frequency of term **t** within the document **d**.

Tweets Length

The length of tweet is an important feature extraction method in author profiling. Although the length of a tweet is fixed but the style of the author and expression of writing and length of tweet can help in classification of important traits of an author [45]. The words of the tweets of a follower are given converted into the vectors and given to the algorithm to learn from. The algorithm learns from the vectorized tweet length and classify the traits of author.

4.5 Classification Machine Learning Models

After applying the feature extraction algorithms, the vectorized data has been used for the development of the models using machine learning which includes KNN, Decision Tree, Random Forest, and Logistic Regression. Deep learning algorithms includes LSTM and CNN.

K-Nearest Neighbor

The working of KNN is simple, it computes the of all the data instances and a query, it computes all the distances of the query and select the closet to the query. The most important variable in KNN is k, it determines the number of closest neighbors to be selected for the data instances [42]. It also determines the performance of the algorithm.

Decision Tree

Decision tree model is a supervised learning model, and it constructs the rules in tree like structure. It has different nodes which represents the attributes, and the leaf node holds the class label or the outcome of the rule. To predict the value of the target variable, tree is constructed based on the information gain on each node and the decision tree is constructed through these rules.

Random Forest

The random forest is also a supervised learning algorithm and it used for the classification and regression as well, in our research, prediction of age of the celebrity is a regression problem and random forest is applied to predict the age of a celebrity [43]. It constructs several trees to build a construct decision rules for prediction.

Logistic Regression

Logistic regression is a supervised learning algorithm, use for the regression problem. It used to predict the dependent variable from a data of the collection of dependent variables. It used to predict the output of dependent categorical variable from an independent variable.

4.6 Deep Learning Classification Models

Deep learning algorithms are use for the classification and regression problems. Deep learning algorithms do not need the feature extraction methods, it extracts the feature from the data by itself and them build the model based on the features vector. For the deep learning models, it requires to have more data to build an efficient model.

Convolutional Neural Network

Convolutional neural network is a supervised deep learning model, it uses for both the classification and regression problems [40]. It has three layers, the input layer, hidden layer, and the output layer. The hidden layer performs the convolution to the input data and predict the output on the output layer.

Long Short-Term Memory

LSTM is an Artificial and deep learning algorithm use for the classification and regression problems. It is a supervised learning model; it uses the labeled data for the classification [35-37]. It is feedback connected model. It doesn't work on single data point but on the entire data sequence. It provides the short-term memory to the recurrent neural networks.

4.7 Evaluation Measures

After applying the different machine learning and deep learning models, the evaluation measures are use to evaluate the performance of the models [38], there are different evaluation measures used for the evaluation of the models. It includes:

- Confusion Matrix
- Accuracy
- Precision
- Recall
- F1-Score

4.7.1 Confusion Matrix

Confusion matrix is use in the classification problems as an evaluation measure. It represents the accuracy of the objects in the validation set [42]. It also provides the information about the performance of the classifier algorithm.

Table 4: Confusion Matrix

	Predict values	
Actual Values	TP	FP
	FN	TN

The True Positive (TP) is the value that is true in actual, and it is predicted as true. False Positive (FP) is the value that is false, and it is predicted wrongly projected as the true. False Negative (FN) is the value that is false, and it is categorized wrongly. True Negative (TN) has been labeled negatively.

4.7.2 Accuracy

Accuracy is the main evaluation measure for the classification model, it depicts the overall accuracy of the classification model.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

4.7.3 Precision

The precision is calculated for the sum of the values that are positively classify by the classifier.

$$\text{Precision} = \frac{TP}{TP+FP}$$

4.7.4 Recall

Recall is the calculation of the sum of the negatively predicted values by the classifier.

$$\text{Recall} = \frac{TP}{TP+FN}$$

4.7.5 F1 Score

The F1 score is calculated by both precision and the recall. It depends on both, if the values of the precision and the recall are high then the value of the F1score would be high else it would be low.

$$\text{F1}-\text{Score} = \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

4.8 Results and Analysis

In this section, the results of the proposed machine learning and deep learning models are discussed. The performance of these models is evaluated using the F1 score, precision, recall and accuracy score. The results of different models are discussed, which are applied on followers' data to predict the demographics of celebrities.

There are four demographics that are probed for the classification of celebrity profiling on the twitter data for this research. The demographics are age, occupation, gender, and fame. For profession there are five models applied KNN have accuracy of 0.55, Logistic Regression has an accuracy of 0.60, Decision tree has accuracy of 0.40. Two deep learning modules were applied on the data, from which CNN clock the accuracy of 0.45 and LSTM had 0.50. Average of all the algorithm that is mean of accuracy is 0.525 for the profession prediction.

Table 5: F1-Scores/Accuracy of machine learning models on profession

Evaluation Measures/Algorithms	**KNN**	**Logistic Regression**	**Decision Tree**	**Random Forest**	**SVC**	**NB**	**cRank/ Mean**
TD-IDF	0.61	0.66	0.44	0.78	0.57	0.57	0.588
Count Vectorizer	0.48	0.60	0.52	0.56	0.57	0.67	0.548
Hash Vectorizer	0.50	0.66	0.63	0.77	0.52	N/A	0.601
Length of Tweet	0.44	0.57	0.36	0.36	0.47	N/A	0.428
Accuracy	0.55	0.60	0.40	0.80	0.40	0.40	**0.525**

Table 6: F1-Scores/Accuracy of deep learning models on profession

Models	**F1-Score**	**Accuracy**
LSTM	0.62	0.45
CNN	0.44	0.50
cRank/Mean	0.515	0.475

There was total six machine learning models applied on the data, Table 5 depicts the F1-scores and accuracy of these models along with cRank. The cRank for the profession demographic for

the KNN was 0.588 for TD-IDF, 0.548 for the count vectorizer, 0.601 for hash vectorizer and 0.428 for the length of tweet. Two deep learning algorithms were applied, cRank of them is 0.515 and mean accuracy is 0.475. The lowest cRank for the profession demographic was 0.428 for length of tweets because the only length of tweets was passed to algorithms as vectors. The length of tweets are just number that doesn't contain too much information to train a model more efficiently.

The second demographic of the celebrity is age prediction, which is a multiclass problem and there was total six algorithms were applied on the data for the prediction of the age. KNN has the accuracy of 0.50 for the age prediction. Logistic regression has the accuracy of 0.45 for the age prediction, Decision tree has 0.40, Random Forest 0.40 and the deep learning models CNN has 0.40 and LSTM had 0.30. The cumulative average of the age prediction accuracy is 0.475.

Table 7: F1-Scores/Accuracy of machine learning models on Age

Evaluation Measures/Algorithms	**KNN**	**Logistic Regression**	**Decision Tree**	**Random Forest**	**SVC**	**NB**	**cRank/ Mean**
TD-IDF	0.67	0.62	0.51	0.43	0.57	0.67	0.564
Count Vectorizer	0.67	0.62	0.45	0.39	0.57	0.67	0.539
Hash Vectorizer	0.67	0.62	0.53	0.39	0.62	0.67	0.563
Length of Tweet	0.49	0.67	0.52	0.52	0.57	N/A	0.547
Accuracy	0.50	0.45	0.40	0.40	0.60	0.50	0.475

Table 8: F1-Scores/Accuracy of deep learning models on Age

Models	**F1-Score**	**Accuracy**
LSTM	0.46	0.30
CNN	0.45	0.40
cRank/Mean	0.455	0.350

There were six machine learning models applied on the data for age prediction, Table 7 depicts the F1-scores and accuracy of these models along with cRank. The cRank for the age demographic

for the KNN was 0.564 for TD-IDF, 0.539 for the count vectorizer, 0.563 for hash vectorizer and 0.547 for the length of tweet. Two deep learning algorithms were applied, cRank of them is 0.455 and mean accuracy is 0.350. The lowest cRank for the age demographic was 0.539 for count vectorizer. The cRanks for the age prediction were more promising on machine learning algorithms because the classes for the prediction of age were balanced. The accuracy mean was 0.475 for machine learning algorithms.

The third demographic which is probed is gender identification. It is a binary class problem. There is total 6 models applied for the prediction of the gender on the twitter data of the celebrities' followers. The accuracy of the KNN is 0.55, Logistic regression has 0.60, Decision tree has 0.30, Random Forest 0.70 and the deep learning algorithms, CNN clocked 0.65 and LSTM has 0.65. The average accuracy score of the algorithms is 0.516.

Table 9: F1-Scores/Accuracy of machine learning models on Gender

Evaluation Measures/Algorithms	**KNN**	**Logistic Regression**	**Decision Tree**	**Random Forest**	**SVC**	**NB**	**cRank/ Mean**
TD-IDF	0.71	0.71	0.41	0.43	0.71	0.71	0.578
Count Vectorizer	0.71	0.56	0.56	0.71	0.62	0.67	0.634
Hash Vectorizer	0.71	0.71	0.48	0.60	0.33	N/A	0.522
Length of Tweet	0.52	0.71	0.58	0.58	0.67	N/A	0.606
Accuracy	0.55	0.60	0.30	0.70	0.40	0.55	0.516

Table 10: F1-Scores/Accuracy of deep learning models on Gender

Models	**F1-Score**	**Accuracy**
LSTM	0.79	0.65
CNN	0.79	0.65
cRank/Mean	0.79	0.65

For machine learning, there were six machine learning models applied on the data for age prediction, Table 10 depicts the F1-scores and accuracy of these models along with cRank. The cRank for the age demographic for the KNN was 0.578 for TD-IDF, 0.634 for the count vectorizer, 0.522 for hash vectorizer and 0.606 for the length of tweet. Two deep learning algorithms were applied, cRank of them is 0.790 and mean accuracy is 0.65. The lowest cRank for the age demographic was 0.522 for hash vectorizer. The highest cRank is 0.634 and for deep learning is 0.79 among all demographics is for the gender prediction because gender prediction is a binary classification problem.

The last demographic of the celebrity profiling is fame prediction, which is a multiclass problem. There are six models applied on the data. The classical machine learning models includes KNN has 0.30, Logistic Regression has 0.30, Decision tree has 0.25, Random forest 0.40, SVM 0.55 and deep learning algorithms CNN has 0.35, LSTM 0.45 and the average accuracy of the six algorithms is 0.358.

Table 11: F1-Scores/Accuracy of deep learning models on Fame

Models	**F1-Score**	**Accuracy**
LSTM	0.62	0.45
CNN	0.34	0.35
cRank/Mean	0.439	0.40

Table 12: F1-Scores/Accuracy of machine learning models on Fame

Evaluation Measures/Algorithms	**KNN**	**Logistic Regression**	**Decision Tree**	**Random Forest**	**SVC**	**NB**	**cRank/ Mean**
TD-IDF	0.33	0.30	0.35	0.41	0.53	0.52	0.388
Count Vectorizer	0.33	0.30	0.28	0.50	0.62	0.52	0.388
Hash Vectorizer	0.33	0.30	0.35	0.47	0.33	0.52	0.368
Length of Tweet	0.30	0.52	0.46	0.46	0.67	N/A	0.451
Accuracy	0.30	0.30	0.25	0.40	0.55	0.35	0.358

For the last demographics, the cRank for the six machine learning algorithms is shown in the Table 12. The cRank for the age demographic for the KNN was 0.388 for TD-IDF, 0.388 for the count vectorizer, 0.368 for hash vectorizer and 0.451 for the length of tweet. Two deep learning algorithms were applied, cRank of them is 0.439 and mean accuracy is 0.40. The lowest cRank for the age demographic was 0.388 for count vectorizer. The lowest cRank among all the demographics is 0.388 for the fame prediction because the classes were unbalanced.

Figure 9 shows the confusion matrix for the gender prediction on the LSTM. The confusion matrix represents the crux of the results of the classification problems. The figure shows the confusion matrix in which class 0 is termed as female and class 1 is termed as male. Each row of the matrix have the values of the prediction. The values of 2 and 3 are from the class female and it provides 2+3=5 multiclassification. The second row shows the class 1 which is the male class, it has the values of 12 and 3.

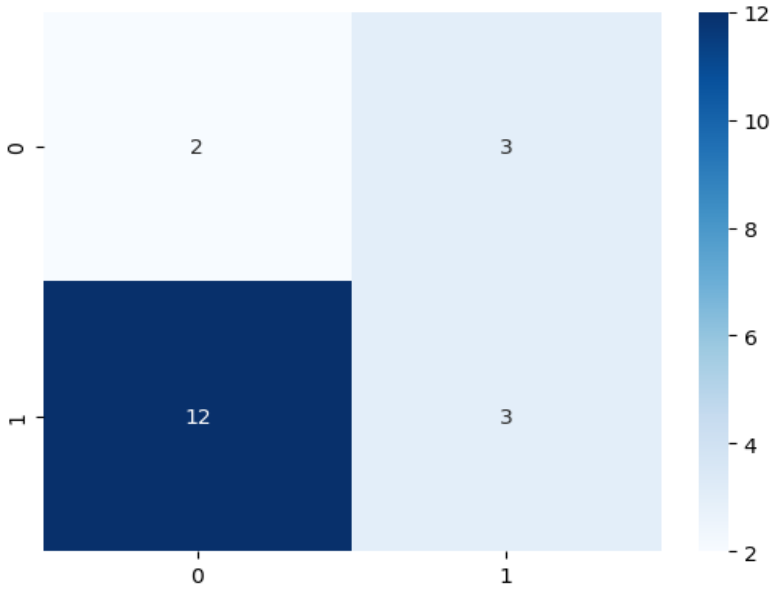


Figure 9:Confusion matrix for the gender prediction on LSTM

Figure 10 shows the training vs validation accuracy and training vs validation error for the LSTM model on the gender identification. The training accuracy is the accuracy when the model is used to train on the training data which is shown in the figure in blue color and the validation accuracy is when model is validated in the validation data. In deep learning models, the feature extraction

is done automatically, and the validation data is selected randomly to avoid the over or under fitting of the model.

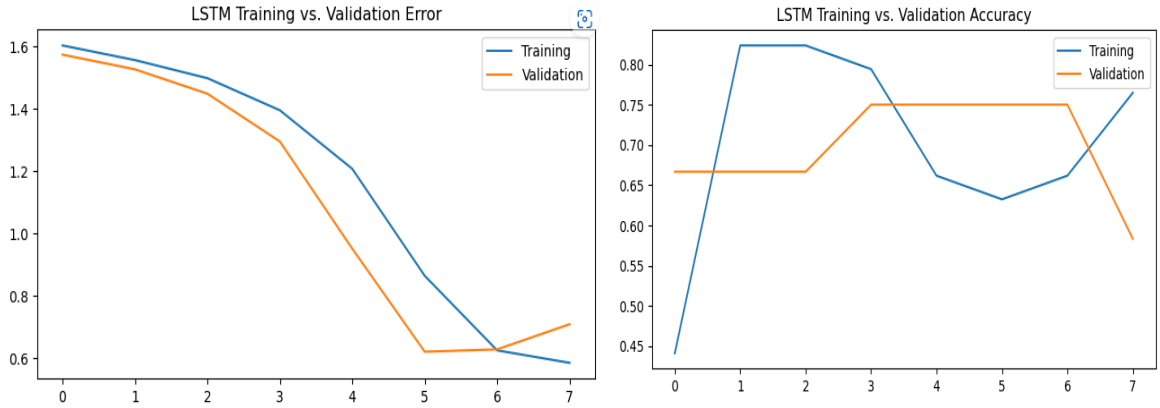


Figure 10: Accuracy-Error Curve for CNN Model on gender identification

Summary

There are four demographics that were probed for the classification of celebrity profiling on the twitter data for this research. There were only 100 celebrities, due to small amount of data, the classes are unbalanced. The demographics are age, occupation, gender, and fame. The highest cRank among the demographics were clocked for the gender prediction which is 0.634 for the machine learning algorithms and 0.79 for the deep learning algorithms. The cRank for the gender prediction was highest because it was a binary classification problem. The lowest cRank was observed for the fame demographic which is 0.388. The data is divided into 20/80 Ratio. 20% was used for the testing and the results discussed are from testing data. The fame class was highly unbalanced that's why it didn't show the promising results.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

Currently people used to communicate, interact, and build relationships through social media. Celebrities are prolific authors and most of their personal information is public knowledge. There are some digital celebrities who exist only on social media, e.g., Twitter. Twitter is a social networking service which provides general populace as well as celebrities to interact with their fans.

The demographics of celebrities could be predicted by the text of their followers as both shares same interest. However, most of the work on celebrity profiling has been performed on English and other similar languages except Urdu. On the contrary, majority of the sub-continent celebrities and their fans tweets in Urdu. To fulfill this gap, in this research work we were used Urdu tweets (short text) of 10 followers of a celebrity to build the first celebrity profiling based on followers' tweets corpus.

There has been some issue which is faced during the research work, firstly the collection of data was a hard task for the celebrities who only tweets in Urdu and to find their followers who do the same. Secondly, for the selection of the algorithms. It was a hectic task to find out the best algorithms and the best features extraction techniques for the celebrity profiling. The feature extraction techniques include the TD-IDF and the length of the tweets,

The corpus is be preprocessed, and Machine Learning (Logistic Regression, Support Vector Machines etc.) and Deep Learning (CNN, LSTM etc.) algorithms were used to train models for the prediction task. The trained model were evaluated using state-of-the-art evaluation measures, i.e., precision, recall, and F1. The highest accuracy which is clocked is for the gender identification with 0.75 on CNN and the second best was for the KNN and it was 0.60.

5.2 Future Work

There are a lot of applications of the author and celebrity profiling. Author profiling playing an important role in forensics, advertisement, and the reduction of cybercrimes. This research is limited to the twitter and short text. The social media is emerging as the best tool for the assistant of the human being and in the future this research might expand to the other social media platforms include:

- Facebook
- E-commerce Reviews profiling
- LinkedIn

The research for the Urdu speaking community in these platforms will help in different fields, like to find out the fake accounts on Facebook, to probe about the fake reviews on E-commerce sites.

Chapter 6

References

[1] E. R. D. Weren and A. U. Kauer, et al., “Examining Multiple Features for Author Profiling,” J. Inf. Data Manag., vol. 5, no. 03, pp. 266–279, Oct. 2014.

[2] F. Chiu Hsieh and R. F. Sandroni Dias, et al., “Author Profiling from Facebook Corpora,” presented at the LREC 2018, Miyazaki, May 2018.

[3] F. Rangel and P. Rosso, et al., “Multimodal Gender Identification in Twitter,” presented at the CLEF 2018 Evaluation Labs and Workshop, France, Sep. 2018.

[4] H. Van Halteren, et al., “Author Verification by Linguistic Profiling: An Exploration of the Parameter Space,” ACM Trans. Speech Lang. Process., vol. 4, no. 1, pp. 1–17, Jan. 2007.

[5] S. Argamon and M. Koppel, et al., “Automatically Profiling the Author of an Anonymous Text,” Commun. ACM, vol. 52, no. 2, pp. 119–123, Feb. 2009.

[6] A. Hodge and S. Price, et al., “Celebrity Profiling using Twitter Follower Feeds,” presented at the CLEF 2020 Labs and Workshops, Sep. 2020.

[7] M. Wiegmann and S. Benno, et al., “Overview of the Celebrity Profiling Task at PAN 2019,” presented at the CLEF 2019 Labs and Workshops, Sep. 2019.

[8] M. Wiegmann and B. Stein, et al., “Overview of the Celebrity Profiling Task at PAN 2020,” presented at the CLEF 2020 Labs and Workshops, Sep. 2020.

[9] S. Hussain, et al., “Resources for Urdu Language Processing,” presented at the The 6th Workshop on Asian Languae Resources, 2008.

[10] A. Daud and W. Khan, et al., “Urdu language processing: a survey,” Artif. Intell. Rev., vol. 47, no. 3, pp. 279–311, Mar. 2017.

[11] V. Radivchev and A. Nikolov, et al., “Celebrity Profiling using TF-IDF, Logistic Regression, and SVM—Notebook for PAN at CLEF 2019,” presented at the CLEF 2019 Labs and Workshops, Sep. 2019.

[12] M.-S. uis Gabriel and E. Puertas, et al., “Celebrity Profiling on Twitter using Sociolinguistic Features—Notebook for PAN at CLEF 2019,” presented at the CLEF 2019 Labs and Workshops, Sep. 2019.

[13] S. Nowson and J. Perez, et al., “XRCE Personal Language Analytics Engine for Multilingual Author Profiling—Notebook for PAN at CLEF 2015,” presented at the CLEF 2015 Evaluation Labs and Workshop, France, Sep. 2015.

[14] F. Rangel, and P. Rosso, et al., “Overview of the Author Profiling Task at PAN 2013,” presented at the CLEF 2013 Evaluation Labs and Workshop, Spain, Sep. 2013.

[15] I. Markov and H. Gómez-Adorno, et al., “Adapting Cross-Genre Author Profiling to Language and Corpus—Notebook for PAN at CLEF 2016,” presented at the CLEF 2016 Evaluation Labs and Workshop, Portugal, Sep. 2016.

[16] M. Martinc and I. Škrjanec, et al., “Author Profiling - Gender and Language Variety Prediction,” presented at the CLEF 2017 Evaluation Labs and Workshop, Ireland, Sep. 2017.

[17] D. Estival and T. Gaustad, et al., “Author Profiling for English Emails,” presented at the Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics, Sep. 2007.

[18] F. Rangel and P. Rosso, et al., “Overview of the 2nd Author Profiling Task at PAN 2014,” presented at the Working Notes Papers of the CLEF 2014 Evaluation Labs, Sep. 2014.

[19] H. G.-A. Ilia Markov, et al., “Author Profiling with doc2vec Neural Network-Based Document Embeddings,” presented at the Mexican International Conference on Artificial Intelligence, Oct. 2016.

[20] N. Aletras and B. Paul, et al., “Predicting twitter user socioeconomic attributes with network and language information,” Jul. 2018, pp. 20–24.

[21] J. Bakerman and K. Pazdernik, et al., “Twitter Geolocation: A Hybrid Approach,” ACM Trans. Knowl. Discov. Data, vol. 12, no. 3, pp. 1–17, Jun. 2018.

[22] R. Baly and G. Karadzhov, et al., “What Was Written vs. Who Read It: News Media Profiling Using Text Analysis and Social Media Context.” arXiv preprint arXiv:2005.04518, May 2020.

[23] J. Bevendorff and M. Potthast, et al., “Heuristic Authorship Obfuscation,” presented at the Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Italy, Jul. 2019.

[24] B. Josh and J. D, et al., “Discriminating gender on Twitter,” presented at the Conference on Empirical Methods in Natural Language Processing. -Association for Computational Linguistics, Aug. 2011.

[25] M. A. Alvarez-Carmona and E. Guzm´an-Falc´on, et al., “Overview of MEX-A3T at IberEval 2018: Authorship and aggressiveness analysis in Mexican Spanish tweets,” presented at the Notebook papers of 3rd sepln workshop on evaluation of human language technologies for iberian languages (ibereval), Spain, Sep. 2018.

[26] S. Daneshvar and D. Inkpen, et al., “Gender Identification in Twitter using N-grams and LSA Notebook for PAN at CLEF 2018,” presented at the Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018), Sep. 2018.

[27] G. Farnadi and J. Tang, et al., “User Profiling through Deep Multimodal Fusion,” presented at the Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, Feb. 2018.

[28] M. Fatima and K. Hasan, et al., “Multilingual author profiling on Facebook,” presented at the Information Processing & Management 53.4, Jul. 2017.

[29] M. Gjurkovic and J. Snajdar, et al., “Reddit: A Gold Mine for Personality Prediction,” presented at the PEOPLES@ NAACL-HTL, Jun. 2018.

[30] M. Koppel and S. Argamon, et al., “Automatically Categorizing Written Texts by Author Gender,” presented at the Literary and linguistic computing 17.4, Nov. 2002.

[31] K. Michal and D. Stillwell, et al., “Private traits and attributes are predictable from digital records of human behavior,” presented at the Proceedings of the national academy of sciences, Apr. 2013.

[32] V. E. Lynn and N. Balasubramanian, et al., “Hierarchical Modeling for User Personality Prediction: The Role of Message-Level Attention,” presented at the Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Jul. 2020.

[33] P. Mishra and M. Del Tredici, et al., "Author Profiling for Abuse Detection," presented at the Proceedings of the 27th international conference on computational linguistics, Aug. 2018.

[34] B. Koloski, S. Pollak, and B. Škrlj, "Know your Neighbors: Efficient Author Profiling via Follower Tweets Notebook for PAN at CLEF 2020," *CEUR Workshop Proc.*, vol. 2696, no. Section 4, 2020.

[35] R. Alroobaea, A. H. Almulihi, F. S. Alharithi, S. Mechti, M. Krichen, and L. H. Belguith, "A Deep learning Model to predict gender, age and occupation of the celebrities based on tweets followers Notebook for PAN at CLEF 2020," *CEUR Workshop Proc.*, vol. 2696, no. September, pp. 22–25, 2020.

[36] S. Maharjan, P. Shrestha, and T. Solorio, "A simple approach to author profiling in MapReduce: Notebook for PAN at CLEF 2014," *CEUR Workshop Proc.*, vol. 1180, pp. 1121–1128, 2014.

[37] J. Schler, M. Koppel, S. Argamon, and J. Pennebaker, "Effects of age and gender on blogging," *AAAI Spring Symp. - Tech. Rep.*, vol. SS-06-03, pp. 191–197, 2006.

[38] F. Rangel and P. Rosso, "Overview of the 7th author profiling task at Pan 2019: Bots and gender profiling in twitter," *CEUR Workshop Proc.*, vol. 2380, 2019.

[39] M. Busger op Vollenbroek *et al.*, "GronUP: Groningen User Profiling---Notebook for PAN at CLEF 2016," *CLEF 2016 Eval. Labs Work. -- Work. Notes Pap. 5-8 Sept. Évora, Port.*, 2016.

[40] R. R. R. Merugu and S. R. Chinnam, "Automated cloud service based quality requirement classification for software requirement specification*," Evol. Intell.*, vol. 14, pp. 389–394, 2021

[41] V. Lynn, N. Balasubramanian, and H. A. Schwartz, "Hierarchical Modeling for User Personality Prediction: The Role of Message-Level Attention," pp. 5306–5316, 2020.

[42] S. Argamon, M. Koppel, J. W. Pennebaker, and J. Schler, "Automatically profiling the author of an anonymous text," *Communication of the ACM*, vol. 52, pp. 119–123, 2009.

[43] M. E. Aragón *et al.*, "Overview of mex-a3t at iberlef 2020: Fake news and aggressiveness analysis in mexican Spanish," *CEUR Workshop Proceeding.*, vol. 2664, pp. 222–235, 2020.

[44] J. D. Burger, J. Henderson, G. Kim, and G. Zarrella, "Discriminating gender on Twitter," *Proceeding of the 2011 Conference on Empirical Methods Natural Language Processing*, pp. 1301–1309, 2011.

[45] J. Bevendorff, M. Potthast, M. Hagen, and B. Stein, "Heuristic authorship obfuscation," Proceeding of the *57th Annual Meeting of the Association for Computational Linguistics..*, pp. 1098–1108, 2020.

[46] R. Baly *et al.*, "What Was Written vs. Who Read It: News Media Profiling Using Text Analysis and Social Media Context," pp. 3364–3374, 2020.

[47] N. Aletras and B. P. Chamberlain, "Predicting twitter user socioeconomic attributes with network

and language information," *Proceeding of the 29th ACM Conference on Hypertext and Social Media*, pp. 20–24, 2018.

[48] M. Martinc, I. Skrjanec, K. Zupan, and S. Pollak, "PAN 2017: Author Profiling-Gender and Language Variety Prediction.," *Working Notes on* Conference and Labs of the Evaluation Forum *2017-Conference and Labs of the Evaluation Forum*, 2017.