A LARGE-SCALE TWEET DATASET FOR URDU TEXT SENTIMENT ANALYSIS

A PREPRINT

Rakhi Batra
Department of Computer Science
Sukkur IBA University
Sukkur 65200, Pakistan
rakhi@iba-suk.edu.pk

Zenun Kastrati
Department of Informatics
Linnaeus University
351 95 Växjö, Sweden
zenun.kastrati@lnu.se

Ali Shariq Imran
Department of Computer Science
Norwegian University of Science and Technology (NTNU)
2815 Gjøvik, Norway
ali.imran@ntnu.no

Sher Muhammad Daudpota
Department of Computer Science
Sukkur IBA University
Sukkur 65200, Pakistan
sher@iba-suk.edu.pk

Abdul Ghafoor
Department of Computer Science
Sukkur IBA University
Sukkur 65200, Pakistan
aghafoor.mscsf19@iba-suk.edu.pk

March 22, 2021

ABSTRACT

This article presents a dataset of tweets in the Urdu language. There are 1,140,824 tweets in the dataset, collected from Twitter for September and October 2020. This large-scale corpus of tweets is generated by performing pre-processing which includes removing columns containing user information, retweet's count, followers information, duplicate tweets, removing unnecessary punctuation, links, symbols, and spaces, and finally extracting emojis if present in the tweet text. In the final dataset each tweet record contains columns for tweet id, text, and emoji extracted from the text with a sentiment score. Emojis are extracted to validate Machine Learning models used for the multilingual sentiment and behavior analysis. These are extracted using a Python script that searches for an emoji from the list of 751 most frequently used emojis. If an emoji is present in the text, a column with the emoji description and sentiment score is added.

Keywords Urdu Twitter Dataset · Urdu Natural language processing (NLP) · Urdu text Sentiments and Emoticons

1 Motivation

Incorporation of multilingual feature in social media platforms and other online content posting platforms has enabled people to use their own languages to converse over such platforms. Due to this, huge content is being generated and research community is focusing to make use of this data. But the challenge is unavailability of such data in an structured format that is useful for machine learning models to conduct a meaningful study. In this article, we are trying to collect a dataset of tweets in Urdu language. Because Urdu is a low resource language and is spoken by more than 600 million people from all over the world.

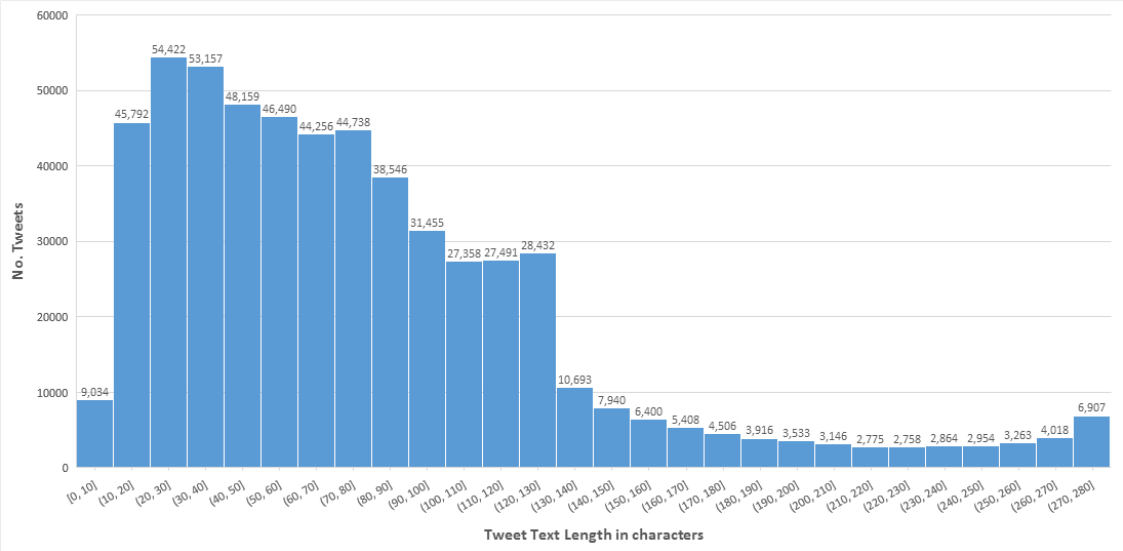

Figure 1: Distribution of tweets on the basis of tweet length

2 Data Description

The dataset [2] presented in this work has been collected by scraping twitter for tweets in Urdu language for the months of September and October 2020. There are 1,140,825 tweets in the dataset. Information about tweets includes tweet id, tweet text, emoji description with sentiment score, and emoji category. Emojis are basically extracted from tweet text by using a Python script for the purpose of sentiment analysis. Emoji description, representation and sentiment score is collected from [3]. According to twemoji[1], a project of twitter, twitter has 3245 different emojis but we have used 751 most frequently used emojis identified by [3]. Categories of selected emojis are presented in Table 1.

Table 1: Emojis' Categories

Category	No. emojis
Miscellaneous Symbols and Pictographs	414
Emoticons	76
Miscellaneous Symbols	62
Dingbats	44
Transport and Map Symbols	38
Box Drawing	30
Geometric Shapes	04
Enclosed Alphanumeric Supplement	21
Block Elements	11
Arrows	07
Miscellaneous Technical	06
Enclosed Alphanumeric	04
Miscellaneous Symbols and Arrows	04
Latin-1 Supplement	03
Specials	02
Arabic	02
Enclosed Ideographic Supplement	01
Playing Cards	01
Braille Patterns	01
Total	751

[1] https://twemoji.twitter.com/

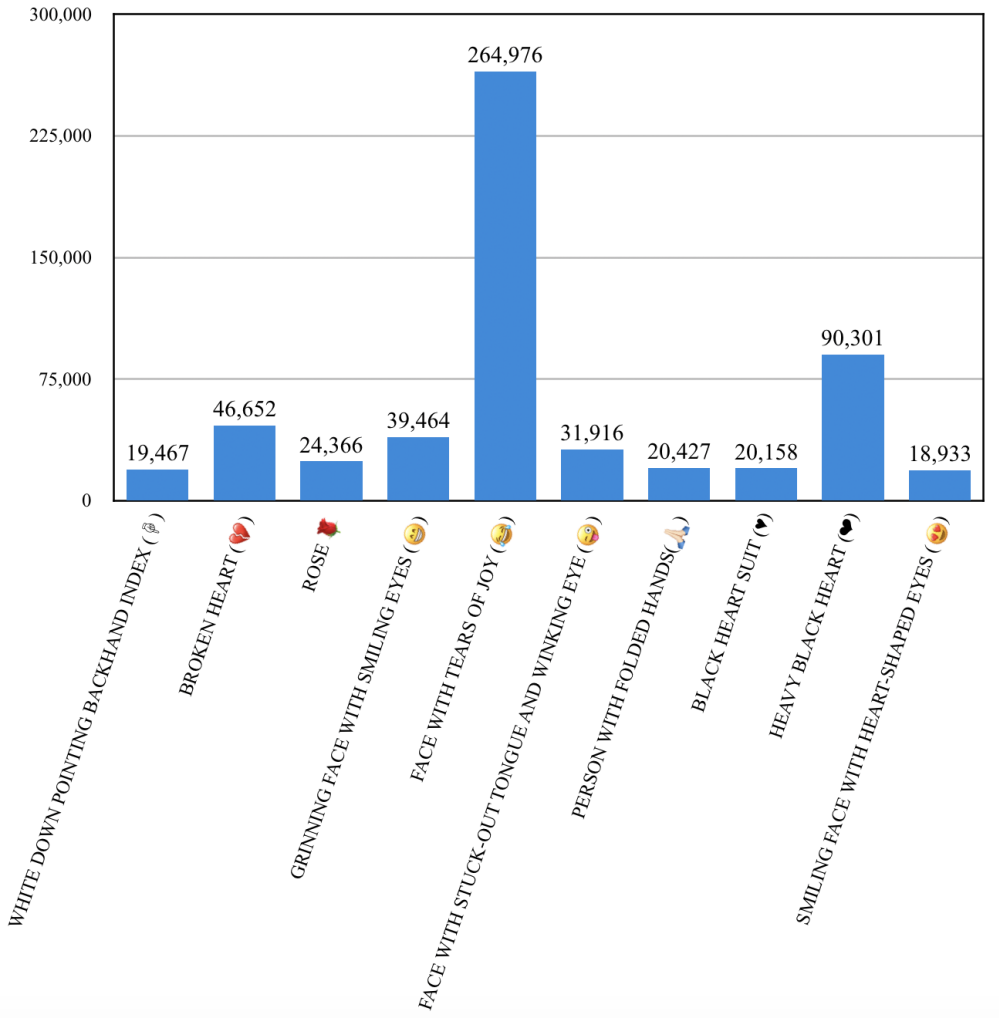

Figure 2: Top 10 frequent used emojis

The collected tweets are varying in length of text. Figure 1 depicts the distribution of tweets on the basis of length of tweet text. It shows that the most common size of tweet text ranges from 23 characters to 43 characters. The minimum length is 3 characters while maximum length is 280 characters.

As stated above, we have used a list of 751 emojis to extract from tweet text for sentiment analysis. Out of 751, the most frequently used emoji is "Face with tears of Joy" that has been used 264,976 times. Other 25 frequently used emojis are shown in Figure 2.

3 Experimental Design, Materials and Methods

Twitter is a social media platform that allows users to post small text messages called tweets. Everyday millions of tweets are being posted by millions of users. These tweets cover a wide range of subjects like politics, sports, movies, poetry, technology etc. It allows users to write a tweet in multiple languages. Along with these services, twitter provides two APIs that are used programmatically to extract data from twitter. These are Streaming API and Search API. Streaming API is used to get the live data and Search API is used to extract historical data. Different filters can be applied to get the required information such as tweets that contain a particular keyword, hashtag or from a specific user. We can also get tweets for a specific date or date range (should not be older than six days) by using Search API.

In order to use Search API, we need API key and API secret that is being provided by Twitter for authentication purposes.

The purpose of this dataset is to collect the tweets that are in Urdu language and tag each tweet with emoticons that are present in tweet text for sentiment analysis purposes. The main steps involved in collecting and creating the dataset are illustrated in Figure 3.

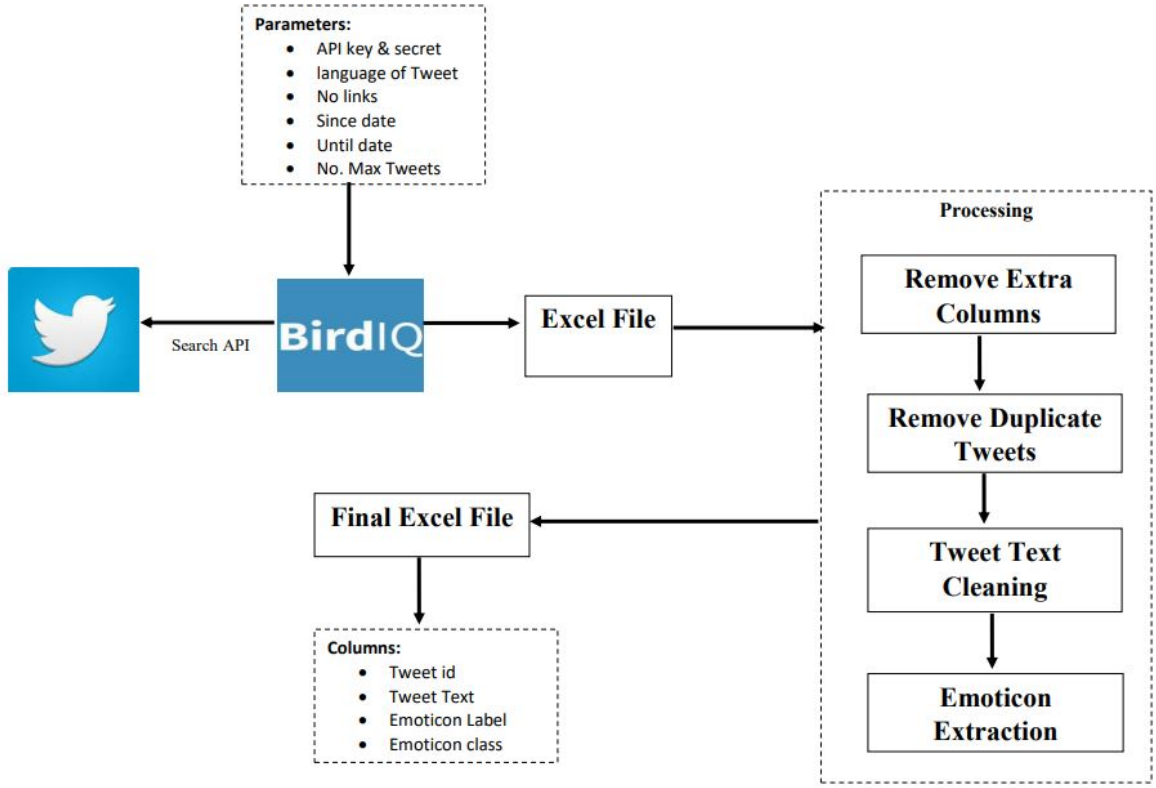

Figure 3: Dataset collection flowchart

We started by first getting an API key and secret from Twitter developer account. The API can be implemented in different programming languages, but we used a tool BirdIQ[2] . It uses the Search API of Twitter. The interface is user friendly and has features to search tweets by applying different filters. Also, it stores the extracted data in a structured format like excel.

For our purpose, we designed a query that extracts tweets for a specific date, written in Urdu language and do not contain links. Below is the query:

$$lang : ur\ until : [specify - end - date]\ since : [specify - start - date] - filter : links \tag{1}$$

We have also provided a check on the maximum number of tweets per query so that we do not lose any data for a particular date. Maximum number of tweets is set to 0.5 million. The extracted data is saved into an excel file. The file contains 72 columns, which describe tweets, users who have posted it, retweet information. We have included just three columns, suitable for the purpose of sentiment analysis on Urdu text. The included columns are tweet id. tweet text and tweet create date. After removing unnecessary columns, we checked for duplicate tweets. The tweet id column contains the unique identity information of each tweet. So, first we checked duplicates by id, if there are any, the duplicate records are deleted. Then we checked the tweet text column. Because when a user re-posts the same tweet or makes a retweet, the text remains the same but is being posted with a different id. So, this is necessary to check duplicate tweets by text.

After removing duplicates, we have cleaned the dataset. The cleaning is done by a pseudocode shown in Figure 4, it removes hashtags, mentions, RT signs, irrelevant spaces, and non- ASCII characters.

[2]https://birdiq.net/twitter-search

```
emogiToCheck = list of emojis to be searched from tweets

COLS = Column names in the output file

def clean_tweets(tweet):
    preprocessing steps for removing links, spaces, puntuations, mentions etc from the tweet text
return tweet

def write_tweets(file):
    Read excel data file
    create an emply dataframe for storing output

    new_entry = []

    for all the records in the excel data file:
        Tweet_text = read tweet text
        emoticons = The emoji/emojis found in the tweet text by checking from the list of emojis
        cleaned_tweet = processed tweet
        new_entry = Tweet related information like id, cleaned text, emojis

        put record into the dataframe
    Export dataframe to excel
    print done
```

Figure 4: Pseudocode for dataset pre-processing

We intend to make this dataset suitable for sentiment analysis, so we extracted emojis from the tweet text if available. The emojis used, their sentiment score and categories are provided in the Data description section. These emojis will help to categorize tweets in positive, negative or neutral classes depending on the sentiment score of emoji. Emoticons are extracted with a code shown in Figure 4 and a final dataset is prepared by adding emoticons description columns. A dataset snippet is shown in Figure 5

Id	Text	English Translation	Emoji	Category
1321220564384620000	اپ کیا کہنا چاہتی ہیں کہ تعلیم عورت کو بے ادب، مغرور، بدتمیز، خاوند کو غلام بناتی ہے۔ اور اس کو اپنی جیت اور کامیابی سمجھتی ہے۔ اپ کی ٹویٹ کی کوئ ٹک سمجھ نہی آئ۔ تعلیم تو شعور دیتی ہے اور معاشرے کو خوشحالی کا ذریعہ ہے۔ 😳😰😒	What do you want to say that education makes a woman uneducated, proud, abusive, her husband a slave? And she considers it her victory and success. Didn't understand any of your tweets. Education brings awareness and is a source of prosperity to the society. 😰😒 😳	['UNAMUSED FACE, -0.374', 'FEARFUL FACE, -0.14', 'FLUSHED FACE, 0.018']	Disgust, Fear
1311716677277580000	آج خوش ہوں بہت 😋	Very happy today 😋	['FACE SAVOURING DELICIOUS FOOD, 0.631']	Joy
1319297754754150000	تم بات کم کرتے ہو اور روتے زیادہ ہو، کوئی لاجک سے انسانوں کی طرح بحث کیا کرو۔ 😠	You talk less and cry more, argue with some logic like human beings 😠	['ANGRY FACE, -0.299']	Angry
1315725902647610000	میں تو کسی غریب کی زندگی بچا رہا ہوں 😢	I am saving the life of some poor person 😢	['CRYING FACE, 0.007']	Sad

Figure 5: Urdu Dataset Snippet

4 Applications

The proposed large-scale Urdu dataset will be useful in various application areas. The research community in the field of artificial intelligence such as natural language processing, machine learning, information retrieval can benefit from these data by using them in various research tasks such as: multilingual sentiment and behaviour analysis [4], cross-cultural aspects and attitude identification [5], figurative language detection (sarcasm, irony) [6], semantic and contextual text analysis [11], topic modelling [8], opinion mining [9], data balancing [10], performance analysis of deep learning models and techniques [7]. UrduNLP[3] has demonstrated that how Urdu dataset is useful for Parts of speech tagging, Named Entity Recognition, Urdu chat-bots etc. Researchers in [1] have collected and used the Urdu language text for Sentiment Analysis but they suffer the problem of unavailability of annotated dataset and were able to work on only 1000 instances. Some researchers have also used Roman Urdu Dataset[4] for sentiment analysis[5]

Another area where this dataset is useful is affective computing research community. There are two reasons: First, it is the first of its kind publicly available dataset aimed at Tweets analysis in Urdu language using NLP techniques, and second, this dataset with its large-scale corpus size could serve as a standard benchmark for these research areas as well as for testing the performance of the existing and new methods and techniques. Additionally, decision makers, (i.e. Government, state owned agencies), researchers and psychologists interested in behavioural sciences and opinion mining can make use of this data.

5 Conclusion

With the invent of new machine learning models like deep learning, low resource languages have gained attention from researchers to conduct new studies. But they face the issue of dataset unavailability. In this work, we have tried to solve this issue by collecting a large-scale dataset of tweets in Urdu language. Urdu is one of the low resource languages and research has shown a positive contribution of people in the applications of Urdu text mining. The dataset is collected through Twitter Search API for the period of two months. After applying necessary preprocessing techniques the final dataset contains 1,140,825 tweets along with emoji's and their sentiment score to make it suitable for natural language processing.

References

[1] Khan, M. Y., and Nizami, M. S. Urdu Sentiment Corpus (v1.0): Linguistic Exploration and Visualization of Labeled Dataset for Urdu Sentiment Analysis.2020 International Conference on Information Science and Communication Technology (ICISCT), Karachi, Pakistan, 2020, pp. 1-15, doi: 10.1109/ICISCT49550.2020.9080043.

[2] Batra, R., Kastrati, Z., Imran, A. S., Daudpota, S. M., Ghafoor, A. A Large Scale Tweet Dataset for Urdu Text Sentiment Analysis. Mendeley Data, V1, doi: 10.17632/rz3xg97rm5.1. 2020

[3] Kralj Novak P, Smailović J, Sluban B, Mozetič I. Sentiment of Emojis. IPLoS ONE 10(12): e0144296. https://doi.org/10.1371/journal.pone.0144296. 2015.

[4] Lo, S.L., Cambria, E., Chiong, R., and Cornforth., D. Multilingual sentiment analysis: from formal to informal and scarce resource languages. Artificial Intelligence Review, vol. 48, pp. 499–527 (2017). https://doi.org/10.1007/s10462-016-9508-4

[5] Imran, A.S., Daudpota, S.M., Kastrati, Z., and Batra, R. Cross-Cultural Polarity and Emotion Detection Using Sentiment Analysis and Deep Learning on COVID-19 Related Tweets. IEEE Access, vol. 8, pp. 181074-181090, 2020, doi: 10.1109/ACCESS.2020.3027350.

[6] Zhang, S., Zhang, X., Chan, J. and Rosso, P. Irony detection via sentiment-based transfer learning. Information processing and Management, vol. 56(5), pp. 1633-1644, 2019.

[7] Kastrati, Z., and Imran, A.S. Performance analysis of machine learning classifiers on improved concept vector space models. Future Generation Computer Systems, vol. 96, pp. 552-562, 2019.

[8] Amara, A., Hadj Taieb, M.A., and Ben Aouicha, M. Multilingual topic modeling for tracking COVID-19 trends based on Facebook data analysis. Applied Intelligence (2021). https://doi.org/10.1007/s10489-020-02033-3.

[9] Schulz, J. M., Womser-Hacker, C., and Mandl, T. Multilingual Corpus Development for Opinion Mining. International Conference on Language Resources and Evaluation, pp. 3409-3412, 2010.

[3] https://www.urdunlp.com/

[4] https://www.kaggle.com/smat26/roman-urdu-dataset

[5] https://www.kaggle.com/smat26/sentiment-analysis-on-roman-urdu/

[10] Shaikh, S., Daudpota, S.M., Imran, A.S., Kastrati, Z. Towards Improved Classification Accuracy on Highly Imbalanced Text Dataset Using Deep Neural Language Models. Applied Sciences 2021, 11, 869. doi: 10.3390/app11020869.

[11] Kastrati, Z., Imran, A.S., and Yayilgan, S.Y. The impact of deep learning on document classification using semantically rich representations. Information Processing and Management vol. 56(5), pp. 1618-1632, 2019.