Twitter corpus of Resource-Scarce Languages for Sentiment Analysis and Multilingual Emoji Prediction

Nurendra Choudhary, Rajat Singh, Vijjini Anvesh Rao, Manish Shrivastava
Language Technologies Research Center, Kohli Center on Intelligent Systems,
International Institute of Information Technology, Hyderabad, India
{nurendra.choudhary, rajat.singh, vijjinianvesh.rao}@research.iiit.ac.in
m.shrivastava@iiit.ac.in

Abstract

In this paper, we leverage social media platforms such as twitter for developing corpus across multiple languages. The corpus creation methodology is applicable for resource-scarce languages provided the speakers of that particular language are active users on social media platforms. We present an approach to extract social media microblogs such as tweets (Twitter). In this paper, we create corpus for multilingual sentiment analysis and emoji prediction in Hindi, Bengali and Telugu. Further, we perform and analyze multiple NLP tasks utilizing the corpus to get interesting observations.

1 Introduction

Twitter has become a valuable source of data for various NLP studies such as sentiment analysis, polarity detection and emoji prediction. Although, copious amount of research studies have been conducted on twitter data, a majority of them deal with English tweets. This is due to several factors. First, English dominates in the mix of languages on Twitter. According to (Hong et al., 2011), more than 50% of tweets are in English. Outside of the largest five Twitter languages (given in Figure 1), other languages represent just under 1% of Twitter traffic each. We primarily focus on collecting data for sentiment analysis. Additionally, taking the premise that social media devices like emojis convey sentiment of their respective tweet, we provide a methodology for collecting tweets with emojis for any language irrespective of its resource availability. Going beyond sentiment analysis, we also collect data without any preconditions on presence or absence of emojis in the tweets which could be used to draw interesting social media analytics both within and across languages or linguistic communities.

We discuss, specifically, about resource-poor languages because such discourse is available on resource-rich languages like English and Spanish, whereas resource-poor languages are largely ignored. For such resource-poor, but yet widely spoken, languages (especially in multilingual communities that have their predominant literature in a language other than their native tongue), we observe social media as a good data source. In this paper, we look at Twitter to collect data in Telugu, Hindi and Bengali, predominantly spoken in the Indian subcontinent. In these regions, English is the language of administration and is increasingly becoming the lingua franca. This, also, explains the lack of sizable corpora despite the relatively large number of speakers.

2 Related Work

Sharing of corpora or resources is important for researchers to compare results with each other, pushing the boundaries of the state-of-the-art model. Openly available corpora reduce the efforts of researchers in constructing the same corpus others developed merely for comparison. However, Twitter's terms of service, under which sharing of aggregated resources (tweets) is barred, prove an unnecessary obstacle in this regard. For example, the Edinburgh Twitter corpus (Petrovic et

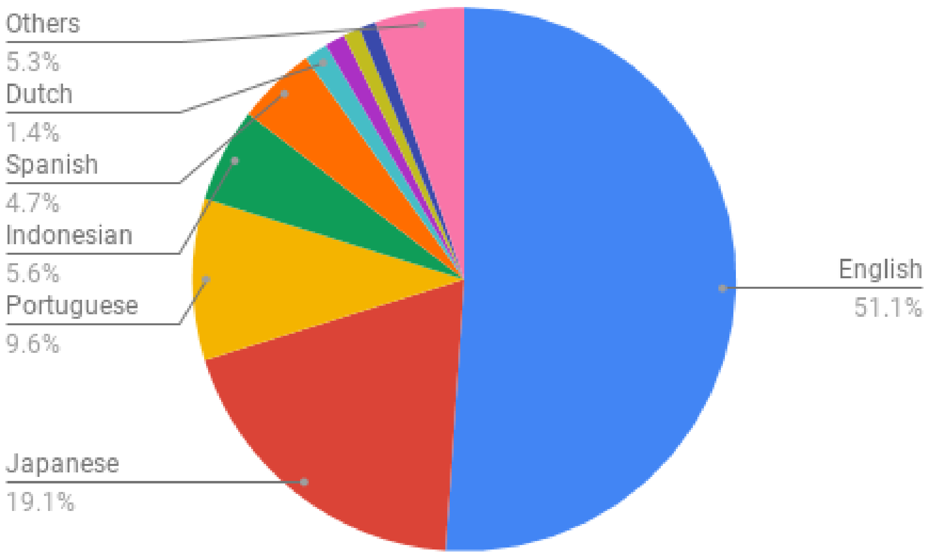

Figure 1: Top languages on Twitter. Data from (Hong et al., 2011)

al., 2010) was dissolved as a result of their policies. However, researchers found certain routes to progress even with such issues, following are some examples:

1. Distribution of only lists of tweet IDs, e.g.; in this TREC 2011 microblog task [1]

2. Distribution of the derivatives of data, e.g.; sharing n-gram counts, instead of the original tweets (Herdağdelen, 2013).

However, the second method loses important information about the data such as word order and limits the analyses of experiments. For example, in (Herdağdelen, 2013), the analysis in section 4 would require more than just n-gram counts since the model deals with tweets per day of the week. Hence, primary metadata regarding the tweets is required.

Twitter provides a streaming API under a public “gardenhouse” setting to build corpora. For example, the Edinburgh corpus (Petrovic et al., 2010), the Tweets2011 corpus from the TREC microblog shared task, and the Rovereto n-gram corpus (Herdağdelen, 2013). In this method, a considerably small fraction of tweets, over a time period are collected. However, the way in which Twitter makes these set of tweets is unclear, inducing a possible bias. Even after these restrictions, the corpora that exist are predominantly in English, hence extracting twitter for resource-scarce languages is necessary.

In their attempt to construct language specific corpora, some approaches choose certain sites to crawl based on results from using medium frequency terms of the language as search terms (Baroni and Bernardini, 2004; Schäfer and Bildhauer, 2012). Following a similar approach, we employ the top most frequent words of the required language as keywords. (Rehbein et al., 2013) proposes an efficient approach to collect German tweets using geolocation features with language filter. However, the data encounters certain biases:

1. Only a fraction of users have GPS access while tweeting. These tweets are included whereas other kinds of users’ tweets are completely ignored. Hence, the collected data is not a representative of the larger sample.

2. Tweets that do not originate from devices with geolocation features like smartphones are also completely excluded. Curated content or in-depth political discussions are not necessarily tweeted from a smartphone.

[1] http://trec.nist.gov/data/tweets

Similarly, (Scheffler, 2014) have created a German twitter corpus using Twitter APIs.

(Cui et al., 2011) analyze emotion tokens, including emotion symbols (e.g. emoticons) for sentiment analysis and emotion analysis of Twitter snapshots. (Barbieri et al., 2017) established a sentiment analysis architecture for Twitter and released a data set for the same in English. They released the data set of roughly 500K tweets for English with emojis as labels. On a similar track, our work focuses on resource-poor languages. Here, the architecture addresses sentiment analysis as a supervised multi class classification problem where each English tweet is "annotated" with its emoji. Hence their data consists of tweets with exactly a single emoji which later is treated as the sentiment label of the tweet, while the tweet itself is striped of the emoji.

We continue their task and collect data on the same lines i.e. we collect Tweets with exactly one emoji in the given aforementioned Languages and more.

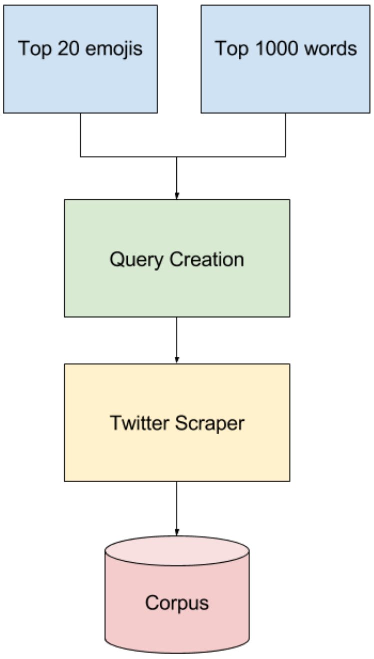

Figure 2: Corpus creation pipeline.

3 Building the corpus

While the previous work for gathering the tweets has been restricted to using twitter search APIs [2], the free usage is limited to only a history of 7 days of tweets. Therefore, we use twitter scraper for no limits with respect to temporal history, as it scraps the web search results of the given query and returns tweets from them. We attempt to scrap tweets using prominent keywords of particular language. Our methodology which is both efficient in time and memory, can easily be extended to any other language irrespective of Resource availability. Following is the description of our methodology to collect tweets with emojis for emoji prediction tasks (Barbieri et al., 2017) which can be extended to collecting just the tweets irrespective of emojis for various other NLP tasks:

[2] `https://developer.twitter.com/en/docs/tweets/search/overview`

Language	Without Emojis	With Emojis as classes	Total
Hindi	194,063	202,415	396,478
Telugu	161,851	16,990	178,841
Bengali	78,308	59,528	137,836
Total	434,222	278,933	**713,155**

Table 1: Total Number of Collected Tweets

3.1 Keywords Extraction

First, we get 1000 most frequent words of the target language for which we want to create the corpus. There are various websites which showcase most frequent used words in a particular language. We use 1000mostcommonwords.com [3] for this project. The most common words usually contain stop words and commonly occuring verbs of the language.

3.2 Query creation

To create a search query for the scraper, in preparing corpus for emoji prediction task. we use an emoji and one of the most frequent words of the language for eg.< ❤,की >. For creating a general twitter corpus irrespective of emojis, we just drop the emoji from the search query.

3.3 Scraper

We then scrap tweets using the Python package *twitterscraper* (Taspinar, 2016 2017) [4]. Our corpus collection pipeline is shown in Figure 2. The scraper returns tweets for the given query based on the condition that the tweet contains the query keyword. We also keep two hashmaps, for tweet id and tweet text's initial 30 characters. These hashmaps are used to remove repetition in incoming tweet results. Every time a tweet is processed, its availability in previously appeared tweets that are already present in corpus is checked, this ensures data is enough diverse and we avoid duplicate entries. Spam or automated tweets from bots, or tweets feed on a celebrity's birthday will be overflowing with very similar tweets like: "जन्मदिन पर बहुत बहुत शुभकामनाएँ बच्चन जी", (Huge Wishes on your Birthday Bacchan sir) and "जन्मदिन पर बहुत शुभकामनाएं अमिताभ जी (Huge Wishes on your Birthday Amitabh sir)" etc. will be avoided in this method.

Collecting such tweets as separate entries is only going to deplete the quality of our data as the later tweet doesn't give much information which the first won't. Any kind of learning on such a data, is only going to reinforce a specific data point rather than giving diversity in label or dataset.

We, also, store a preprocessed version of the collected data by removing mentions and urls because they do not carry significant information for some major tasks such as sentiment analysis and emoji prediction.

4 Corpus Analysis

In total we collected 713,155 tweets in three languages, 396K in Hindi, 178K in Telugu and 137K in Bengali (Table 1) out of which 279K tweets are with emojis as labels and remaining we collected irrespective of emojis. In both the cases we maintain two distinct versions of the data: Cleaned and Uncleaned. Where Cleaning is done with the emoji prediction task in consideration, the changes made include:

1. Trailing symbols removed, for example "????" becomes "?" or "...." becomes ".".

[3] `http://1000mostcommonwords.com/`

[4] `https://pypi.python.org/pypi/twitterscraper/`

❤️ 9.54	😊 9.51	😍 9.42	😏 8.58	😐 7.72	😠 7.50	😢 7.22	🤔 6.78	😡 6.73
😭 4.92	💕 4.32	😉 3.82	😁 2.90	💙 2.81	😘 2.56	😶 2.33	😜 1.87	💜 1.46

Table 2: Percentage of emojis in Hindi tweets

😊 15.65	❤️ 13.99	😍 10.46	😐 10.22	😠 8.66	😏 8.59	😘 7.44	😢 5.80	🤔 5.70
😶 4.64	😭 3.66	😡 1.32	😁 1.08	😜 1.04	😉 0.67	💕 0.52	💙 0.49	💜 0.07

Table 3: Percentage of emojis in Bengali tweets

😊 22.67	😠 16.55	😍 13.38	❤️ 10.77	😐 6.68	🤔 5.67	😏 5.05	😢 4.19	😭 3.91
😡 3.01	😜 2.29	😁 1.57	😉 1.36	💕 0.94	😶 0.84	😘 0.54	💜 0.31	💙 0.25

Table 4: Percentage of emojis in Telugu tweets

2. Words and symbols which are foreign to the language of the tweet are removed, for example "मुर्गी के अंडे और पा जी के फंडे रोके नही रुकते Just #ViruPanti #INDvAUS" will be cleaned to "मुर्गी के अंडे और पा जी के फंडे रोके नही रुकते"

However, As we maintain an uncleaned raw version too, depending on the application, data can be used accordingly.

The most frequent emojis in Telugu, Hindi and Bengali can be found in Table 2 , Table 3, Table 4 respectively.

4.1 Relationship between emojis and Sentiments

A random sampling of 500 tweets of each language, from the extracted tweets with emojis, are manually annotated into three sentiment classes: Positive, Negative, and Neutral to study a relationship between emojis and annotated sentiments. Results shown in Table 5.

The tabulated statistics suggest that sentiments and emojis in most of the cases share a strong correlation, which further reinforces the idea of using emoji as a sentiment label. Furthermore, we note from Tables 5,6 and 7 that even across languages the annotation of different emoji tweets into the three sentiment classes is similar, indicating that the sentiment or emoji's semantic value is preserved across languages till a large extent.

5 Applications

The entire corpus has been hosted on the link given in the footnote [5]. Potential use of the corpus can be really varied:

- `Sentiment Analysis:` As discussed in related work, we built a Sentiment Analysis tool. The emoji data be used as label for Sentiment Analysis, where we could go one step further and predict emoji for the emoji-less data we collected.

- `Enrich Resource-scarce Language:` Siamese network based approaches (Choudhary et al., 2018b; Choudhary et al., 2018c; Choudhary et al., 2018a) are capable of utilizing these

[5] `https://figshare.com/articles/Twitter_corpus_of_Resource-Scarce_Languages_for_Sentiment_Analysis_and_Multilingual_Emoji_Prediction/6477782`

Emoji	Positive	Neutral	Negative	Emoji	Positive	Neutral	Negative
😊	98	2	0	😶	10	73	7
❤️	100	0	0	😭	3	14	83
😍	99	1	0	😡	0	3	97
😐	3	84	13	😁	93	5	2
😠	0	2	98	😜	76	21	3
😏	9	79	12	😉	67	25	8
😘	96	4	0	💕	99	1	0
😢	9	3	87	💙	87	13	0
🤔	3	90	7	💜	82	8	10

Table 5: Distribution of annotated sentiment classes of Hindi tweets with emojis in percentage, symbolized by their emojis here

Emoji	Positive	Neutral	Negative	Emoji	Positive	Neutral	Negative
😊	90	9	1	😶	7	78	5
❤️	99	0	1	😭	9	80	11
😍	95	3	2	😡	0	7	93
😐	2	81	17	😁	90	7	3
😠	0	3	97	😜	71	18	11
😏	5	75	20	😉	71	27	2
😘	93	7	0	💕	100	0	0
😢	11	5	85	💙	91	9	0
🤔	5	89	6	💜	79	11	10

Table 6: Distribution of annotated sentiment classes of Bengali tweets with emojis in percentage, symbolized by their emojis here

emojis of resource-poor language concurrently with emojis of resource-rich languages to enhance the overall performance of both resource-poor and resource-rich languages.

- `Social Media Analytics:` Beyond Sentiment Analysis too, We could learn a lot, especially in Topic modeling across languages for example, what domain do most Telugu tweets belong to. Are Telugu tweets more about Movies than say Bengali tweets which might be into Politics. Social Media is always heavy with Demographic trends which can be analyzed.

6 Limitations and Future Work

Certain issues with our data which require further investigation would include

- Code Mixed Data: Because we only look for just one word to decide the language of the tweet, It's possible we may take a lot of code mixed data as compared to a corpus which is exclusively in one language, for example
  - “तू कल भी दिल में थी... और आज भी है...बस कल तक favorite list मे थी...आज block list मे है... ”, (*You were in my heart till yesterday and so today too, but yesterday you were in my favorite list and today in block list*)
  - “#Sun #Pagli.... अगर तू #doll है तो में #Dollar... अगर तू #Brand है तो में #Branded हु ”, (*Listen, O mad woman, If you are a doll, then I am Dollar, if you are a brand, then I am branded*)

Emoji	Positive	Neutral	Negative	Emoji	Positive	Neutral	Negative
😊	90	9	1	😶	5	74	11
❤️	100	0	0	😭	3	9	88
😍	100	0	0	😡	1	5	94
😐	7	83	10	😁	89	4	7
😠	0	0	100	😜	69	18	13
😏	11	82	7	😉	71	22	7
😘	91	5	4	💕	98	1	1
😢	7	3	90	💙	77	21	2
🤔	2	98	0	💜	91	1	8

Table 7: Distribution of annotated sentiment classes of Telugu tweets with emojis in percentage, symbolized by their emojis here

Such tweets may also get included in our data, even though the amount of Hindi information they have is minimal

- In any data Noise is unfavorable, Noise in such social media platforms could be
  - "सिर्फ $attitude होने से कुछ नही होता #smile ऐसी दो की हर एक #लोग बोल पड़े ", (*only having $attitude won't do anything, give a #smile which will make #people talk*)
  - "####french_fries....!!! : तले भुने आलू सदा सुखी रहो ", (*####french_fries....!!!* .. semantically non compliant)

  Such a noise will corrupt any learning to be performed on this data.

- Twitter like any other social media is a live, thriving platform, taking a snapshot at one time and presuming it to work forever is not a sound judgment. Which is why such data needs to be updated regularly.

7 Conclusion

In this paper, we present a twitter corpora for Telugu, Bengali and Hindi along with a methodology which is scalable across languages. A data which could be used for tasks like Sentiment Analysis,often done on Resource heavy languages only. Along with it we also release a regular data which can be used for tasks beyond. In addition to this, we further present that such a data provides more information than already present in the field, by providing its use for some applications. We hope this corpus can serve as a basis for more work to be done upon such Resource Scarce languages.

References

Francesco Barbieri, Miguel Ballesteros, and Horacio Saggion. 2017. Are emojis predictable? *arXiv preprint arXiv:1702.07285*.

Marco Baroni and Silvia Bernardini. 2004. Bootcat: Bootstrapping corpora and terms from the web. In *LREC*, page 1313.

Nurendra Choudhary, Rajat Singh, Ishita Bindlish, and Manish Shrivastava. 2018a. Contrastive learning of emoji-based representations for resource-poor languages. *arXiv preprint arXiv:1804.01855*.

Nurendra Choudhary, Rajat Singh, Ishita Bindlish, and Manish Shrivastava. 2018b. Emotions are universal: Learning sentiment based representations of resource-poor languages using siamese networks. *arXiv preprint arXiv:1804.00805*.

Nurendra Choudhary, Rajat Singh, Ishita Bindlish, and Manish Shrivastava. 2018c. Sentiment analysis of code-mixed languages leveraging resource rich languages. *arXiv preprint arXiv:1804.00806.*

Anqi Cui, Min Zhang, Yiqun Liu, and Shaoping Ma. 2011. Emotion tokens: Bridging the gap among multilingual twitter sentiment analysis. *Information retrieval technology*, pages 238–249.

Amaç Herdağdelen. 2013. Twitter n-gram corpus with demographic metadata. *Language resources and evaluation*, 47(4):1127–1147.

Lichan Hong, Gregorio Convertino, and Ed H Chi. 2011. Language matters in twitter: A large scale study. In *ICWSM*.

Sasa Petrovic, Miles Osborne, and Victor Lavrenko. 2010. The edinburgh twitter corpus. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, pages 25–26.

Ines Rehbein, Sören Schalowski, Nadja Reinhold, and Emiel Visser. 2013. Uhm... uh.. filled pauses in computer-mediated communication. In *Talk presented at the Workshop on" Modelling Non-Standardized Writing" at the 35th Annual Conference of the German Linguistic Society (DGfS).*

Roland Schäfer and Felix Bildhauer. 2012. Building large corpora from the web using a new efficient tool chain. In *LREC*, pages 486–493.

Tatjana Scheffler. 2014. A german twitter snapshot. In *LREC*, pages 2284–2289.

Ahmet Taspinar. 2016–2017. Python twitterscraper.