# Highlights

## Dataset Cleaning Steps for Hinglish Language Corpus

Hari Thapliyal

- Dataset for Hinglish Language Sarcasm Detection.
- Cleaning Hinglish Text for NLP.

# Dataset Cleaning Steps for Hinglish Language Corpus

Hari Thapliyal (Researcher)

[a] dasarpAI, Lord Krishna Green, Doon University Road, Dehradun, 248001, Uttrakhand, India

ARTICLE INFO

*Keywords*:
Hinglish Language Text,
Text Cleaning in Hinglish Language

## Abstract

Hindi is third [1] most spoken language on our planet. Like English which is written in Roman script, Hindi also does not have its own script but almost all the Hindi speaking people write Hindi in Devanagari script. Hinglish is a mix language and it is spoken by Hindi speaking, English educated people and they can add words from other Indian languages during their conversation. Unlike Hindi Hinglish has its own script and this script is called Hinglish script. Hinglish script has characters borrowed characters from Roman and Devanagari scripts. (**?**) states that 65% of Indian population is under 35 years age. Several disruptions like low cost mobile phone, extremely cheap data, digital India initiatives by government of India has caused huge surge in Hinglish language content. Hinglish language context is available in audio, video, images, and text format. We can find Hinglish content in comment box of online product, news articles, service feedback, WhatsApp messages, social media like YouTube, Facebook, twitter etc. To engage with consumer, it is extremely important to analyse the sentiments, but to perform sentiment analysis it is not possible to read every comment or feedback using human eyes. We created corpus from the data twitter and online sarcastic articles. To create a corpus of Hinglish text we created a checklist and followed those steps. In this small article we are listing the steps followed.

## 1. Background

During our project Sarcasm Detection System in Hinglish Language (SDSHL) we are dealing with different language, different scripts. Data is collected from various blogs and twitter accounts of various native Hinglish speaker. It is challenging to get a clean text which can be used to for model building. To address that problem, we took certain steps and we are going to describe those here.

Cleaning dataset before it is used for model building is essential step in any data science project. Almost all the projects, which are scraping data from web or social media have to go through some common steps every time. Most of the time steps are common and we keep learning challenging when we are dealing with different language or data source. We wanted to maintain this list separate from our main Sarcasm detection work. The reason for that are 1- With every natural language processing project we have a new set of learning and we want to keep updating this from time to time. 2- This list should be available and handy to any researcher and community member who is working on text processing project.

Most of the text from blog was clean but twitter had uncleaned, unstructured sentences. We know that tweet text is unclean because it has text from different languages, in different scripts, extra space, emoticons, non-text sign like ” ” ”:”, ”<” etc, flag sign, line break, over used words like ”.....”, ”??????”, ”beau.....tiful”, ”!!!!!!”. Blog text may also have this kind of text but chances of that is extremely less.

## 2. Checklist of Text Cleaning

- Hyperlinks. Most of the links on twitter are tinyurl leading to other websites. These tinyurls are unique.We do not think it can help us in building any feature. So, we are removing hyperlinks.
- @name. This is used as cc to keep in loop other people. We are keeping @name, but we remove @ sign.
- #hashtag: We are keeping hashtag, but we are removing # sign.
- Space between # and hashtag. We are removing this space.
- No space between emoticon and word: We are creating space between emoticon and word.
- Frequency of Emoticon. Sometimes, people use same emoticon multiple times as continuous text to give extra emotional effect. We are preserving this as is.
- Newline: Manual line breaks are removed.
- Extra space: Extra space is removed.
- Retweet: RT is for retweet. We removed RT text.
- Punctuation: We are removing punctuation like ।, ।, । । , ।।, “.”, :, “,”, “;” But keeping punctuation like “?”, “!” .
- Replacement Rules
  - Any of these characters ,;‘’—“”:“” )(}{ is replace with space.

hari.prasad@vedavit-ps.com (H. Thapliyal)
www.dasarpai.com (H. Thapliyal)
orcid(s): 0000-0001-7907-865X (H. Thapliyal)

  - !+ with !
  - ?+ with ?
  - /, —, _ with space
  - Replace slogans like ” जय श्री राम” or any other stereo type slogan with ””

- Remove Sentence with less than 4 words
- Remove non-Hinglish sentence (i.e. any sentence which is in pure non-Devanagari script)

## 3. Data Sources

### 3.1. Existing Twitter Datasets

- https://github.com/sid573/Hindi_Sentiment_Analysis/blob/master/hi_3500.ods
- https://github.com/rkp768/hindi-pos-tagger/blob/master/News%20and%20tweets/News%20%26%20Tweets(5).xlsx

### 3.2. Twitter Handles

- #व्यंग
- #कटाक्ष

### 3.3. Twitter Account:

- Abhasin89009555
- Arnab5222
- badri_dk
- BhartiyRudr
- chitraaum
- DChaurasia2312
- DeshbhaktRosha1
- drkamleshdwivedi.wordpress
- github_rkp768
- github_sid573
- Jainritesh_rj
- kakesh.wordpress
- KapilMishra_IND
- ManojTiwariMP
- PbSwain_IND
- RajatSharmaLive
- Real_Netan
- Real_Sweta
- RealPushpendra
- Republic_Bharat
- ridhi_bose
- RubikaLiyaquat
- sandeep353055
- Shanu44339200
- ShrishtySharma_
- SushantBSinha
- SwetaSinghAT
- TeriJogan
- TheAbhishek_IND
- Twitter1
- VickyAarya007
- VinodRajotiya82
- VyangyaVahini
- whatsapp
- WiskyWala
- yatisharma111

### 3.4. Blogs

- व्यंग्य बाण —खबरों की खबर => https://www.jagranjunction.com/ajaykumarjha/----/
- पीके .......अब क्यूं बैठे मुंह ..सी के => https://www.jagranjunction.com/ajaykumarjha/----/
- पढे लिखे अशिक्षित (संदर्भ-स्मृति ईरानी विवाद) => https://www.jagranjunction.com/ajaykumarjha/---/
- पंच लाइन >वन लाइनर व्यंग्य बाण => https://www.jagranjunction.com/ajaykumarjha/----/
- पंच लाइन »> व्यंग्य बाण => https://www.jagranjunction.com/ajaykumarjha/----2
- बच्चों के प्रति क्रूर होता समाज => https://www.jagranjunction.com/ajaykumarjha/----
- मुर्गे ढोते स्कूली वैन ....... => https://www.jagranjunction.com/ajaykumarjha/---
- सीबीआई बनी “बाबा जी का ठुल्लु” => https://www.jagranjunction.com/ajaykumarjha/-----
- लोकतंत्र का पकौडा => https://www.jagranjunction.com/ajaykumarjha/--
- खडी खबड -व्यंग्य बाण ........ => https://www.jagranjunction.com/ajaykumarjha/---

- श्रीसंत को मिलेगा “दुर्योधन पुरस्कार ” ...व्यंग्य बाण => https://www.jagranjunction.com/ajaykumarjha/---
- फ़ैसले के बाद भी माकूल सज़ा के विकल्प ( संदर्भ दिल्ली बलात्कार कांड फ़ैसला) => https://www.jagranjunction.com/ajaykumarjha/-----
- बडी नहीं , खडी खबड .......... :) => https://www.jagranjunction.com/ajaykumarjha/---
- प्रिंट मीडिया जिंदाबाद :व्यंग्य बाण => https://www.jagranjunction.com/ajaykumarjha/---
- चेन्नाई एक्सप्रेस की सीक्वेल “मुगलसराय एक्सप्रेस” -पंच लाइन => https://www.jagranjunction.com/ajaykumarjha/---/
- पुत्तर प्रदेश की सैंडवादी पार्टी => https://www.jagranjunction.com/ajaykumarjha/---
- पंच लाइन -व्यंग्य बाण => https://www.jagranjunction.com/ajaykumarjha/---
- https://www.pravakta.com/category/hindi-literature/vangya/
- https://blogs.navbharattimes.indiatimes.com/satire/
- https://www.amarujala.com/humour?src=hbmenu
- https://kakesh.wordpress.com/category/-
- https://drkamleshdwivedi.wordpress.com/--/-/

## 4. Conclusion

Although this is the final checklist which we used for our SDSHL project, yet we will keep updating this based upon our learning from Hinglish or other language projects. When using this file please make note of version number mentioned on the first page of this document.

**Hari Thapliyal** is a Data Science and Project Management professional. He is a mentor, trainer, coach, consultant, meditator, philosopher and blogger. In his 28+ years professional career in software development, project management, training and consulting he has been deeply involved in all kind of roles from software design, development, quality assurance, training, mentoring, PMO head and many others. He has deep interests in diverse subjects like BFSI Sector, Artificial Intelligence, Data Mining, Data Analytics, Deep Learning, ML Modelling, NLP, Economics, Physics, Sanskrit, Vedic Chanting, Vedanta, Healing, History, Culture, Project Management, Meditation and Spirituality. This helps him to understand that how and where to discover new equilibrium among many variables like religion, culture, ethics, morality, societies, business, process automation, and new age technologies like AI, NLP, Deep Learning, GAN, Robotics, Cryptocurrency etc. He is Xpert Coach and Mentor for various AI, ML courses of upGrad and he is founder of dasarpAI an AI Training, Consulting startup.