Dataset Cleaning Steps

Prepared by: Hari Thapliyal

Reviewed by: Dr. Anil Vuppala

Date: 18-Oct-20

Version: 0.2

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.

2. Introduction

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.

3. 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.
  - !+ 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)

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.