Dataset Cleaning Steps

Prepared by: Hari Thapliyal
Reviewed by: Dr. Anil Vuppala
Date: 27-Aug-20
Version: 0.1

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 has 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- More projects of human language project processing, text processing we do more we learn 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 are tinyurl leading to other websites. Tinyurls are unique and we do not think it can help us in building any feature. So, we removed hyperlinks.
- **@name**. This is used as cc to keep in loop other people. We do not think this information will be useful in sarcasm detection. So, we removed @name.
- **Emoticons**: people use an emoticon multiple times we removed duplicate emoticons from the text. We removed duplicity information of emoticon. We feel that even after

removing the duplicate emoticons we could have preserved their frequency, but we did not do that in this experiment.

- **Newline**: Manual line breaks are removed.
- **Extra space**: Extra space is removed.
- **Punctuation**: Any overuse of punctuation is removed and single punctuation is used.
- **Retweet**: Removed if RT is there
- **Space with words and Emoticons**: Create a single space between emoticons and words.
- **Replacement Rules**
    - Replace “” with " ", “-“ with " ", “.” with “।“, “॥“ with “।“, "। " with "।", "। ।" with ""।"", "।" with "। ", ""!!"" with "!"", "??" with "?", "," with ", ", ", " with ", ", Replace "‘" with "", "’" with "", "“" with "", "”" with """, "*" with " ", """ with """, "—" with " ", "/" with " ", " " with " ", "_" with "_""
    - Replace " 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.