Rules for POS Tagger in Hindi:

- Adposition (ADP) as defined in universaldependencies.org, is stated as, 'A closed set of items that occur before (preposition) or after (postposition) a complement composed of a noun phrase, noun, pronoun, or clause that functions as a noun phrase, and that form a single structure with the complement to express its grammatical and semantic relation to another unit within a clause.'
  In Hindi language this is a similar notion to what is called विभक्ति and कारक, from the training data also there are certain words that specify what an ADP is, a simple utility function for checking such words was used which improved F score by a small margin from 0.38 to 0.93.
- AUX usually follows main verb. Earlier without the rule a lot of AUXs were being incorrectly tagged as VERB or PROPN. From the data it was observed that words like हैं and है were tagged as AUX. After applying the rule for these two words, since these two words are pretty abundant in any Hindi dataset (signify end of the sentence) the score improved from 0.11 to 0.58, also decreasing the number of incorrect tags in VERB.
- For subordinating conjunction (SCONJ), there are again a certain instances of words in Hindi that are most often used for linking subordinate clauses, which was also observed from the data, hence a simple function was used for checking those words. This rule improved the score from 0.21 to 0.96.
- Just like the above case similar kind of approach was used for tagging coordinating conjunctions (CCONJ). Certain fixed CCONJ words in Hindi language exists which were verified from data and a simple function detecting those words was used. This got a score of 0.38 from 0.96.
- NUM increased from 0.44 to 0.85. For this, rule created was as follows: A string containing representation of Hindi digits and English digits was created. A list containing basic numbers (0 – 10, 50, Lakh, Crore) as written in Hindi was also made. Then the words were checked against the above string and the list to detect the NUM POS tag.
- DET or Determiners are of certain type in Hindi language, Possessives, Demonstrators and Quantifiers. There are certain words which are most frequently used as Determiners in Hindi. Since such most frequent words are not too many which normally occur in Hindi texts (also observable from the dataset), a simple rule for selecting such words as Determiners was created. This improved the score from 0.11 to 0.69.
- Particles (PART) are function words that must be associated with another word or phrase to impart meaning and that do not satisfy definitions of other universal parts of speech. Again in Hindi language, being in closed class they are not often too much in number and a simple for identifying most common PART tag is created. In this case the score improved to 0.86.
- Adverbs (ADV) modify verbs and adjectives. As observed from the dataset, most of the times Adverbs were preceded by Adpositions. This rule in addition with the

checking of the word which is Adverb from a list of most common adverbs of different types (Adverbs in Hindi | Hindi Language Blog (transparent.com)) in Hindi was used. Apart from this rule another fact that was used was checking if the next word is either adjective or verb and the previous word if found in list of common adverbs, will most probably be an adverb, since this justifies the definition of Adverb itself. These mentioned points increased the score from 0.27 to 0.42.

Some important observations:

- One thing that was observed was that, in general it was much easy to improve results for close class parts of speeches, the most used being the ones that are limited in number (in Hindi language) and hence words could be checked against those particular words.
- Improvement in classification of some POS tags also resulted in improvement in correct tagging in other POS, for which rules were not implemented, since correct classification in certain tags also filtered the wrong tags.
- For open class parts of speech, there are always new words that can be added, hence, it is not easy to generalize these words across domains. Though several rules were tried like checking prefixes or suffixes in adjective creation, or adjective qualifying a noun (usually works in English language), or AUX following a VERB but these methods could not offer much improvement as compared to baseline. In Hindi there are so many different ways adjectives and verbs are used, that it becomes difficult to generalize. It is comparatively easy to identify a Proper Noun in English with an uppercase, not in Hindi though, where there is no notion of case.
- The overall score increased from 0.5125 to 0.7426 for train dataset and gave a score of 0.70 for test dataset.