Name: Deep Salunkhe.
Roll.No: 21102A0014
Subjct: NLP

Assignment :- 5

Q1.] Compare different stemming techniques - Porter stemmer, Lancaster, Regex stremmer

⇒

① Porter stemmer:

→ Widely used, rule based algorith focuses on English morphology.

→ It is simple, fast, effective for English

→ It can over-stem (creating wrong roots) or under stem (missing variation). Not ideal for agglutinative. languags

→ Example: Playing → play
aggrees → agre

② Snowball stemmer

→ It is Rule based, language specific implementation more flexible than Porter

→ It can handle various languags, customizable.

→ Can be complex to implement, potentially less effictive for some languags

→ Ex Playing → play, felices → feliz.

③ Lancaster stemmer

→ It is rule based algorithm with focus on preserving morphology
→ Good for preserving meaning, handls some irregular verbs
→ less aggressive than Porter, potentially misses some variations.

④ Regexp stemmer

→ uses regular expression to remove suffixes/prefixes
→ simple to implement, fast.
→ can be inaccurate, prone to over-stemming, not ideal for complex morphology

Choosing the right stemmer:

language: Porter and Lancaster are best for English, Snowball can be adapted for various languages

Accuracy vs speed: Porter and Regexp are faster, but snowball might be more accurate for specific needs

Preserving meaning: Lancaster is better for maintaining semantic closeness

Q2] Identify the unique challenges in sentiment analysis when applied to customer reviews in Bengali, compared to other languags

⇒ Sentiment analysis in Bengali present unique challengss compared to other language

Agglutinative Language:

Bengali builds words by adding suffix, making stemming complex. stemming by general method might remove crucial information for sentiments.

Lack of Resources:

Compared to English, there are fewer sentiment lexicons and annoted Dataset for Bengali. This can affect the training and accuracy of sentiment analysis models

Sarcasm and Irony:

Bengali uses sarcasm and irony frequently which might be misinterpreted by models trained on literal language

Negation Handling:

Bengali negation is complex, requiring specific rules to identify negated sentiments.