Team Members:

Lim Yi Han (A0239065B)

Low Ye Yang, Caleb (A0234305R)

Jirafe Aditya Vijay (A0239845N)

Jody Tng Jin Zi (A0238195W)

**Q2(c)**

The accuracy of the predictions is 0.659.



**Q3(a)**

To compute the right-hand side of the equation j∗ = argmax P(y = j|x = w), we need to calculate the probability P(y = j|x = w) for each class j then choose the class j with the highest probability.

Using the Naive Bayes classifier approach,

P(y = j|x = w) = P(x = w|y = j) \* P(y = j) / P(x = w),

P(x = w|y = j) is the likelihood of the input feature vector x = w given class j,

P(y = j) is the prior probability of class j,

P(x = w) is the marginal likelihood of the input feature vector.

In order to account for cases where the word is not found, add-1 smoothing was done as well. Hence, overall the function would be:

Additionally, to improve the accuracy of our model, when the word does not exist, we chose to return “@” as the tag, as the usernames are unique and are most likely to be unknown by our model.

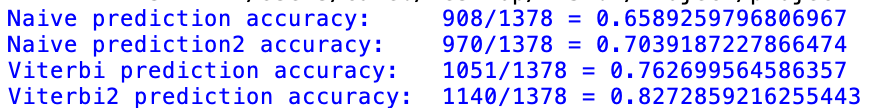
**Q3(c)**

The accuracy of the predictions is 0.704.



**Q4(c)**

The accuracy of the predictions is 0.763.



**Q5(a)**

The new POS tagger can also be improved by better handling unseen words in its output probabilities, by observing the linguistic patterns present in tweets. We classified the tweets according to the linguistic features as shown below.

@ - Username

, - Punctuation

L - Possessive Pronoun

~ - Retweet

& - Conjunction

S - Possessive Noun(‘s)

N - Noun

A - Adjective

G - Symbols

$ - Number

V - Verb

R - Adverb

X - Indefinite Pronoun

T - Phrasal Verbs

M - ["Ricochet's", '#LebronShould']

D - Determiner

O - Pronoun

Z - Name(’s)

! - Interjection

^ - Proper Noun

U - URL

P - Preposition

Y - [“there’s”]

# - Hashtag

Our new POS tagger better handles unseen words and accounts for linguistic patterns by using **Rule-Based POS Tagging**. Based on linguistic patterns that are observed commonly, we have applied these rules in *viterbi\_predict2* to specify patterns for tagging unknown words.

patterns = {

r'@\w+' : '@', # usernames

r'#\w+' : '#', # hashtags

r'RT' : '~', # retweet

r'\b((https?|ftp)://)?(www\.)?[A-Za-z0-9]+\.[A-Za-z]{2,}\b' : 'U', # url

r'^-?[0-9]+(.[0-9]+)?$' : '$', # numbers

r'.\*(?:es|ed|ing)$' : 'V', # verb

r'.\*ly$' : 'R', # adverb

r'.\*(ous|ful|able|ant|ary|ic|ive|less|like|ish|est)$' : 'A', # adjectives

r'.\*\'s$' : 'S', # possessive nouns

r'.\*s$' : 'N', # plural nouns

r'[A-Z][a-z]+' : '^', # proper nouns

}

**Rule-Based Tagging**

1. ‘@’ was tagged to usernames starting with ‘@’.
2. ‘#’ was tagged to hashtags starting with ‘#’.
3. ‘~’ was tagged to ‘RT’.
4. ‘U’ was tagged to URLs starting with ‘http’/’https’.
5. ‘$’ was tagged to numerals.
6. ‘V’ was tagged to verbs ending with ‘es’, ‘ed’, ‘ing’.
7. ‘R’ was tagged to adverbs ending with ‘ly’.
8. ‘A’ was tagged to adjectives ending with ‘ous’, ‘ful’, ‘able’, ‘ent’, ‘ant’, ‘al’, ‘ary’, ‘ic’, ‘ive’, ‘less’, ‘like’, ‘est’.
9. ‘S’ was tagged to possessive nouns ending with ‘apostrophe s’.
10. ‘N’ was tagged to plural nouns ending with ‘s’.
11. ‘^’ was tagged to proper nouns starting with capital A-Z, followed by other letters.

**Q5(c)**

The accuracy of the predictions is 0.827.

