MLRD Supervision 1

Ishan Dwivedi (id360) and Neelu Saraswatibhatla (srns2)

Sentiment Lexicon

1. The articles along with our classifications of each word are in the appendix at the end of this document. The true sentiments of both articles were negative. We correctly guessed this for each other's articles based on manually classified sentiments for each word. The Simple and Improved classifiers correctly classified both articles too.
2. Words that invert the sentiment: never, barely, borderline, almost, doesn't, although, seems, isn't, questionably, hardly. Opinion adjectives can have their sentiment flipped by the presence of the word *not*.
3. Words which indicate a strong emotional involvement include words such as 'stupid', 'incredible', and swear words. Low emotional involvement may be indicated by posts mentioning just the names of places or people within a given picture.
4. accuracy = 328/412 = 0.796~~116505~~
5. If there are naturally a lot more types of a certain class then the types will have higher class probabilities so the classifier will be more likely to select this more frequently occurring class (when using the Naive Bayes Classifier). This is why we use Precision, Recall, and F1 Measures for situations with varying class frequencies.

Naive-Bayes

1.

a.

P(A | F1) = 0.5	P(B | F1) = 0.5
P(A | F2) = 0	P(B | F2) = 1
P(A | F3) = 0.1	P(B | F1) = 0.9

b. P(A) = P(A | F1) * P(F1) + P(A | F3) * P(F3) where P(F1) = frequency of F1 documents divided by the sum of frequencies of F1 and F3 (i.e. the relative probability of F1 over F1 and F3), and analogously for F3, and for class B.

$P(A|D) \propto P(A) \times P(F1|A) \times P(\text{not } F2|A) \times P(F3|A)$

In the example in the question, P(F1) = #F1/(#F1 + #F3) = 10/40 = 0.25, P(F3) = 0.75

P(A) = P(A | F1) * P(F1) + P(A | F3) * P(F3) = 0.5 * 0.25 + 0.1 * 0.75 = 0.2

c. The class probabilities for A would be far lower than for B, so during testing/querying the classifier would be far more likely to select B just because the test set happened to contain more documents of class B.

d. Feature F3 would be most useful as it has the largest disproportion of probabilities between the two classes, while not being 0 for either class. The

classifier would more reliably be able to tell whether a validation/test document is in A or B as class B documents have more of feature 3.

e. Look for the features with the greatest difference in probability for each class, while not being 0 for either class.

2. If, for example, a given text had to repeat a quote or a noun with a strong adjective in it “e.g. It’s a Wonderful Life” several times, there may be a mis-interpretation by the classifier because of all these repeats. Also, neutral words like if, and, there, etc. that don’t really contribute to the sentiment could mess with the probabilities as they appear so many times and don’t change the sentiment, so counting each word once would minimise the effect of such words.

Statistical Properties of Language

1. Most people would agree that the given words aren’t English words, and they would be correct to a certain extent. When new words appear in a language, it may be because of some large scale phenomena (like how to “Google” something is now a verb). However, it could be because of some local dialect phrase that is spreading. In the latter case, most people will not have seen these new dialect words so won’t recognise it as their own language, and they would be right to say their own lexicon doesn’t contain that word. But it doesn’t mean that word won’t be used in an English document by some other person.

   Languages tend to have somewhat consistent rules on sound combinations, especially with consonants. In English, the f sound doesn’t generally follow the p sound within the same syllable, so an English speaker could confidently claim pferd is not an English word. Similarly, x doesn’t follow k, especially since x starts with a k sound. Also, apostrophes are used either in contractions or possessives. In the latter case, they are followed by an ‘s’, while in ‘Kx’a’ the apostrophe is followed by an ‘a’. We can be confident that it isn’t a contraction either as a contraction stands in for one or more letters that do not change the word’s stress pattern significantly, and there are two consonants at the beginning so there wouldn’t be a third consonant, and there is a vowel right after it so if the uncontracted form had a vowel it would have been stressed. We can therefore be sure that ‘Kx’a’ is not an English word.

Appendix: Question 1

Article 1

Manually Classified Words (Green = Positive, Red = Negative)

["2016", "a", "about", "across", "adjustment", "against", "also", "an", "and", "anyone", "anything", "as", "before", "began", "belief", "borders", "brexit", "britain", "britain's", "bureaucracy", "business", "but", "changed", "comparison", "cycle", "depend", "diplomacy", "do", "does", "each", "economics", "else", "escalates", "european", "every", "everything", "for", "geography", "going", "grain", "gummed", "has", "in", "into", "is", "it", "its", "leads", "long", "make", "national", "neglect", "negotiation", "neighbours", "newly", "not", "nothing", "of", "old", "on", "or", "out", "painful", "partnership", "pattern", "playing", "politics", "reality", "referendum", "relationship", "required", "resented", "self-esteem", "self-sabotage", "sovereignty", "stark", "starts", "surrender", "suspicion", "test", "that", "the", "to", "trade", "trying", "union", "with", "work"]

Article

Brexit has changed everything about Britain's relationship with the European Union, and also nothing. For anyone trying to do business across borders newly gummed with bureaucracy, the comparison is stark and painful. But in politics, an old pattern is playing out – a cycle of suspicion and self-sabotage that began long before the 2016 referendum.

It starts with the belief that Britain does not depend on its neighbours for trade or anything else. That leads to neglect of the diplomacy required to make the partnership work. Going against the grain of economics and geography escalates every negotiation into a test of national self-esteem. Each adjustment for reality is resented as a surrender of sovereignty. (https://www.theguardian.com/commentisfree/2021/feb/23/brexit-machine-perpetual-grievance-britain-brussels)

Article 2

Manually Classified Words (Green = Positive, Red = Negative)

[a, accumulating, an, and, are, bar, be, braindead, but, cesspool, charles, convey, do, fault, for, genuinely, get, he, him, his, idiots, in, incoherent, internet, it, itself, james, just, like, looking, me, more, mostly, of, on, other, play, questionable, resulted, smh, soo, stans, stupid, terms, the, to, trying, tweet, twitter, under, upset, warrior, was, were, which, who, word, wrong]

Article

It's mostly James Charles' fault for trying to be an internet warrior on Twitter which just resulted in him looking stupid and accumulating a cesspool of more incoherent idiots under his tweet smh. Like don't get me wrong the bar itself was questionable in terms of the word

play he was trying to convey but the stans and other Twitter idiots who were genuinely upset are soo braindead.

(Comment from https://www.youtube.com/watch?v=tX25hdauW_o&pbjreload=101)