1)

a)

Stoplist removal- Ignoring the words which appear in a stoplist. Stoplist words are such words that don't contribute anything useful in the document.

Example-{are,in,!,some,will,to,is,does,your,many,..}

Capitalisation- Turning all words to lowercase or all words to uppercase for easier processing.

Stemming- Removal of suffix/ end part of word to reduce variations. Ex- Corrupted, Corruption to corrupt

After applying stoplist, stemming and capitalisation, the documents are-

Doc 1: dataset corrupt corrupt data not hast

Doc 2: data system transfer corrupt data file trash

Doc 3: politican corrupt develop country

b)

Inverted index-

Num	Token	Docs
1	dataset	1:1
2	corrupt	1:2, 2:1, 3:1
3	not	1:1
4	data	1:1, 2:2
5	hash	1:1
6	system	2:1
7	transfer	2:1
8	file	2:1
9	trash	2:1
10	politician	3:1
11	develop	3:1
12	country	3:1

c)

1(c)

	q_1	d_1	d_2	d_3
dataset	0	1	0	0
corrupt	1	2	1	1
not	0	1	0	0
data	1	1	2	0
hash	1	1	0	0
system	0	0	1	0
transfer	0	0	1	0
file	0	0	1	0
trash	0	0	1	0
politician	0	0	0	1
develop	0	0	0	1
country	0	0	0	1

Cosine similarity = $\frac{\sum q_i d_i}{\sqrt{\sum q_i^2}\sqrt{\sum d_i^2}}$

$\sum q_i^2$ is constant and hence not considered at ranking

$$sim(q_1, d_1) = \frac{2+1+1}{\sqrt{1+4+1+1+1}} = \frac{4}{\sqrt{8}} = \sqrt{2}$$

$$sim(q_1, d_2) = \frac{2+1}{\sqrt{1+4+1+1+1+1}} = \frac{3}{\sqrt{9}} = 1$$

$$sim(q_1, d_3) = \frac{1}{\sqrt{1+1+1+1}} = \frac{1}{\sqrt{4}} = 0.5$$

Ranking $\Rightarrow$ d_1, d_2, d_3 (decreasing similarity)

Euclidean distance =

$$dist(q_1, d_1) = \sqrt{(0-1)^2+(1-2)^2+(0-1)^2} = \sqrt{3}$$
$$dist(q_2, d_2) = \sqrt{(1-2)^2+(0-1)^2+(0-1)^2+(0-1)^2+(0-1)^2} = \sqrt{5}$$
$$dist(q_3, d_3) = \sqrt{(1-0)^2+(1-0)^2+(0-1)^2+(0-1)^2+(0-1)^2} = \sqrt{5}$$

Ranking $\Rightarrow$ d_1, d_2, d_3 (increasing distance)

d)

1. (d) TF - IDF

IDF takes occurance of terms across documents into consideration

$$IDF = \log \frac{N}{df}$$

N = number of docs in corpora
df = number of docs term is present in

We take query terms only for ranking

TF - IDF

	d_1	d_2	d_3	q_1	
corrupt	0	0	0	0	$(\log \frac{3}{3} = 0)$
data	$\log \frac{3}{2}$	$2\log \frac{3}{2}$	0	$\log \frac{3}{2}$	
hash	0	0	0	0	$(\log \frac{1}{1} = 0)$

$sim(d_1, q_1) = 0$

$sim(d_2, q_1) = 2\log \frac{3}{2} \times \log \frac{3}{2} = 2 \times 0.02 = 0.04$

$sim(d_3, q_1) = 0$

Ranking $\Rightarrow$ D_2, D_1, D_3

2)

a)

The stages of processing commonly followed within natural language generation (NLG) systems are-

- Document planning
- Microplanning:
- Realisation: grammatical details E.g. children vs. childs, an apple vs. a apple

i)

Document planning decides on the content and the structure of text in NLG. It chooses the best data to be reported out of all the available tons of data. Document planning depends on what kind of data is important to be put forward to the user. It also checks what kind of presentation makes a good narration and what can be expressed in layman terms.

Document planning also takes care of organisation of text like the order of conveying the data and the Rhetorical structure.

ii)

Microplanning decides on how to linguistically express text i.e., which words and sentences to use. It takes Lexical and syntactic choice into consideration and plans on what kind of words and linguistic structures to use. It also plays a role in distribution of information across sentences and paragraphs i.e., Aggregation.

Microplanning also does referencing, so it takes care of how the objects and entities are referred.

iii)

Realisation is all about the grammatical details and punctuation. E.g. children vs. childs, an apple vs. a apple.

Realisation plays an important role in formation of legal English sentences based on the above two stages.

Realisation should strictly follow the sub-language meaning it properly expresses the technical domain and restricted domain languages. It also makes sure that the proper genre is followed when creating the NLG like magazine writing or social media text.

Realisation also makes sure the final output format is met, like HTML, RTF formats.

b)

The study of words and how they are produced, as well as their relationship to other words in the same language, is referred to as morphology in linguistics. Prefixes and suffixes, for example, are variations of a word's base form.

There are two kinds of morphologies- inflectional morphology and derivational morphology

i) **Inflectional Morphology**: It is the variation in words which don't specifically change the meaning. Examples include plurals, superlatives and past tense.

In inflectional morphology, the parts of speech is unchanged.

Examples- run, running, ran

ii) **Derivational Morphology**: It is the variation in which the meaning changes with introduction of suffix or prefix. Examples such as prefix re- and suffix -er come under derivational morphology.

In derivational morphology, the parts of speech is changed.

Examples- draw, re-draw, slow, slow-er

c)

The three metrics to evaluate the quality of binary sentiment analysis systems are-

- Precision
- Recall
- F-measure

	Relevant	Non-relevant	Total
Retrieved	A	B	A+B
Not retrieved	C	D	C+D
Total	A+C	B+D	A+B+C+D

i) **Precision:**

The proportion of retrieved content that is relevant is known as precision. It is determined by dividing A (number of relevant documents retrieved) by the number of documents actually retrieved.

The formulae for precision is- $$\text{Precision} = \frac{A}{A+C}$$

ii) **Recall:**

The proportion of relevant material retrieved out of all relevant collection is known as recall. It's determined by dividing A by the number of papers that are relevant.

The formulae for recall is- $$\text{Recall} = \frac{A}{A+B}$$

iii) **F-measure**:

F-measure takes both precision and recall into account, cancelling out each other's negative characteristics. The (H.M) harmonic mean of recall and precision is used to calculate it. Because a high F-measure requires a high Recall and high Precision, Harmonic mean is used instead of Arithmetic mean.

The formula for F-measure is- $$\text{F-measure} = (2 \times P \times R) \Big/ (P + R)$$

d)

The intuition behind using a Naive Bayes classifier for text classification is that it is straightforward and effective if the data isn't sparse.

Naive Bayes classifier estimates the probability of each class given a text.

The equation for the classifier is-

$$\underset{c_i}{\operatorname{argmax}}\, P(c_i) \prod_{j=1}^{n} P(t_j|c_i)$$

where, P(ci)= Prior, i.e, probability of segment having class ci, and $\prod_{j=1}^{n} P(tj|ci)$ is the Likelihood i.e, product of probabilities of each feature value of segment occurring with class ci.

3)

a)

The initial step is to conduct a subjectivity analysis. This is concerned with determining whether the text contains opinions, sentiment, or simply facts(objective statement). After this, only the subjective text which contains emotions and sentiments is put forward to do the analysis.

Subjective sentences are not always used to express feelings., e.g.: I think he came yesterday. Hence it is different from sentiment alone.

b)

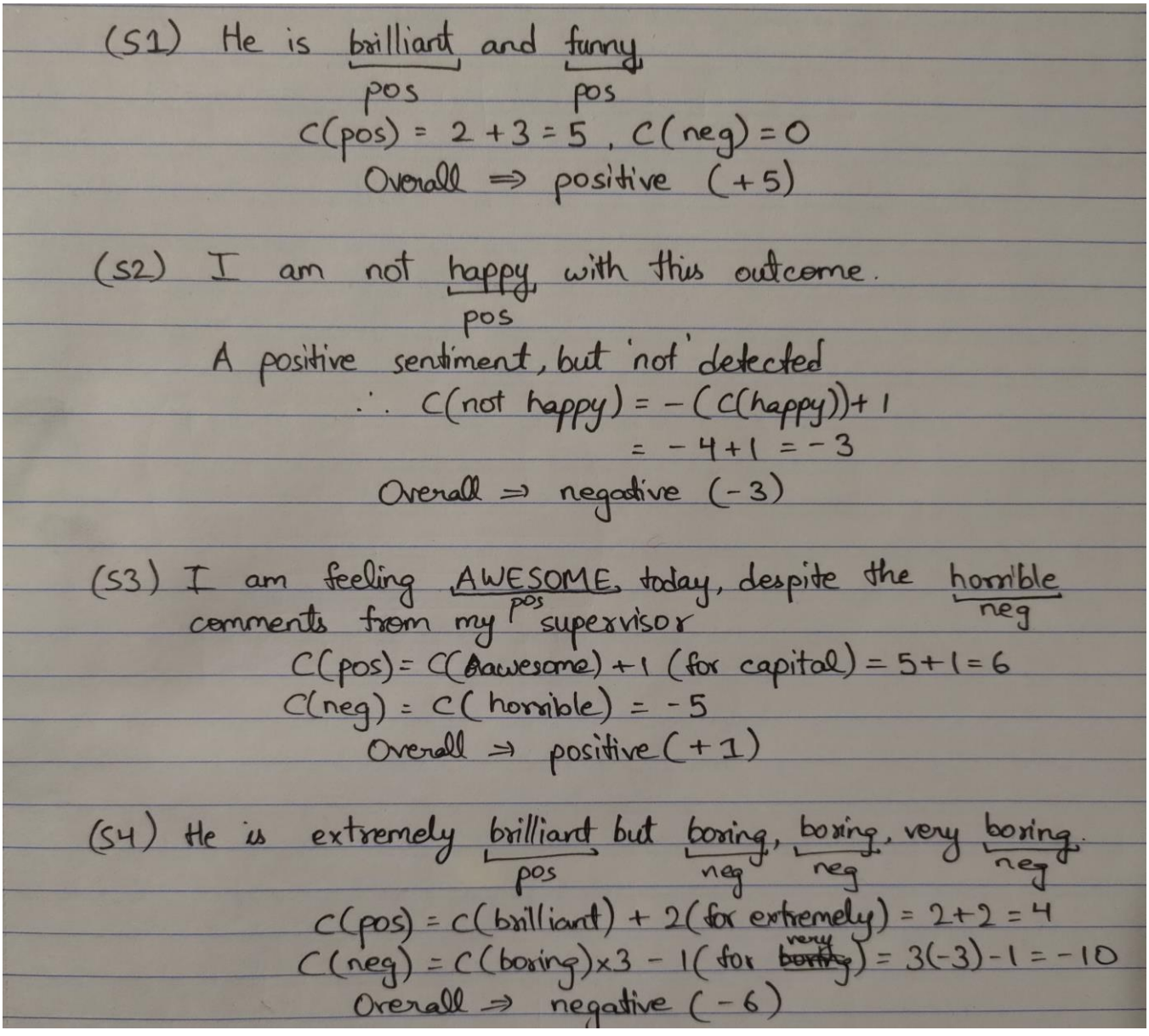
The relevance of automatic techniques for sentiment analysis for marketing purposes are-

- Users can immediately figure out what are the different positives and negatives of features in a product they are going to purchase just by looking at sentiment analysis of existing lengthy reviews of users.
- It saves a lot of human effort and is fool proof.

c)

Weighted lexical-based approach for Sentiment Analysis:

Positive weights (Cpos) and negative weights (Cneg) of words in the text are counted and weighted using the lexicon weights: If Cpos >= Cneg, the result is positive. If Cpos < Cneg, the result is negative.



d)

“I have just bought the new iPhone 13. It is a bit heavier than the iPhone 12, but it is much faster. The camera lenses are also much better, taking higher resolution pictures. The only big disadvantage is the cost: it is the most expensive phone in the market. Mark Jobs, 12/11/2021.”

An opinion is a quintuple (oj , fjk, soijkl, hi , tl), where:

• oj is a target object: in the example, iPhone 13

• fjk is a feature of the object oj : in the example, weight, speed, camera, cost

• soijkl is the sentiment value of the opinion: in the example, negative, positive, positive, negative

• hi (the author of the post): in the example, Mark Jobs

• on feature fjk of object oj at time tl : in the example, 12/11/2021.

The two language processing challenges in automating the identification of such elements are-

• To identify target objects, Named Entity Recognition(NER) is required.

• Co-reference resolution to know that ”it, that“ are needed. Ex= iPhone in above example