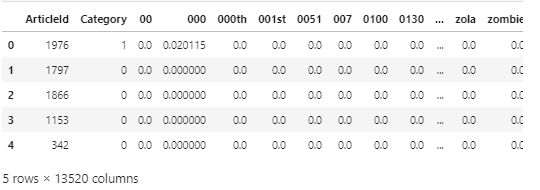
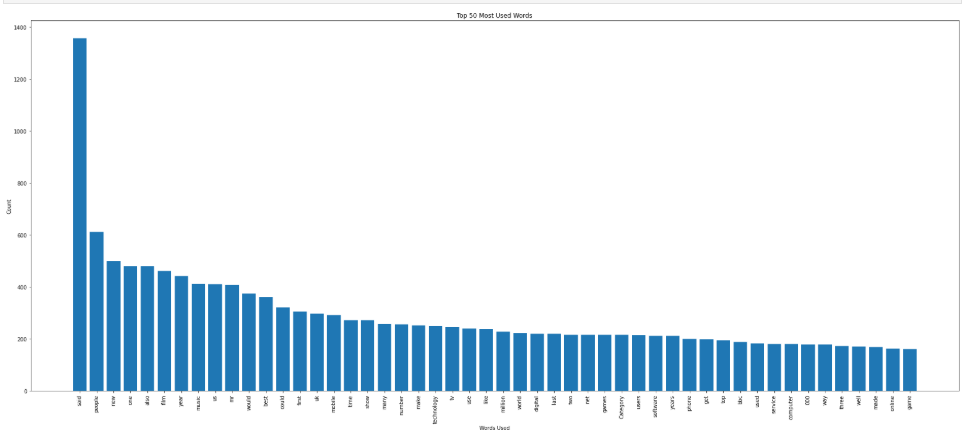
**Task One: Exploratory Data Analytics**

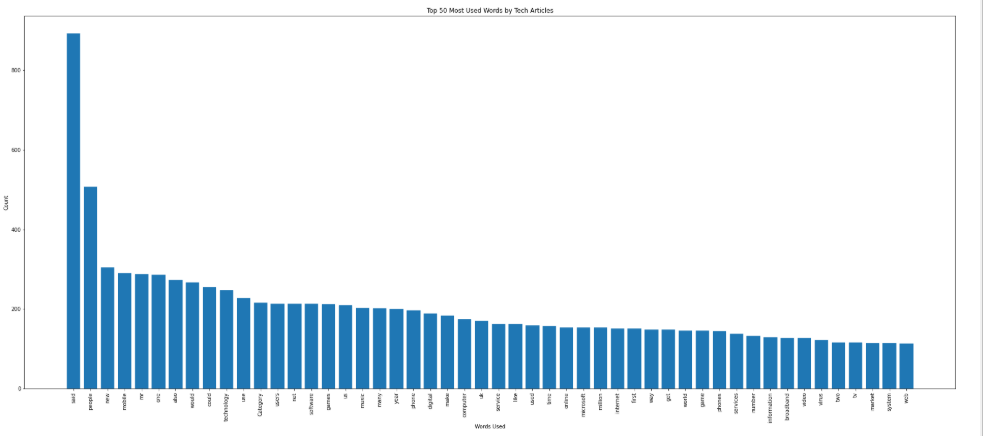
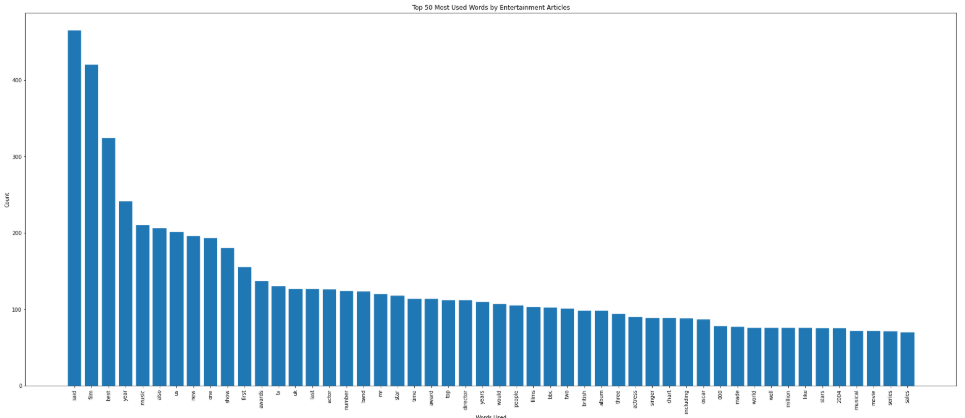
1. This dataset consists of 428 articles and there are 13520 features extracted from it. 5 example articles with their extracted features are displayed below.



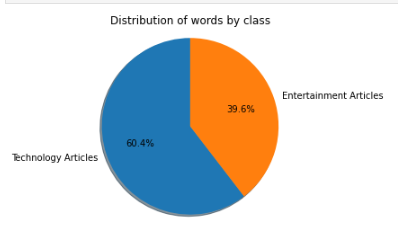
1. i)The figure below shows the top-50 term frequency distribution across the entire dataset



ii)Term frequency distribution for respective class of articles



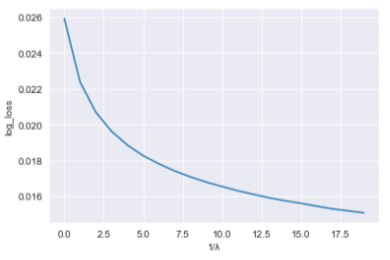
iii) class distribution



**Task Two: Classification Models Learning**

1. **LR**

* The plot below shows how the training loss varies with λ increasing.



* The impact of λ on bias and variance: As we increase λ from 0, there is an inverse/negative relationship between bias and variance because as the variance decreases, the bias increases.

1. **NB**
2. Top 20 words that are most likely to occur in entertainment news articles: ['film' 'best' 'said' 'show' 'music' 'band' 'year' 'awards' 'us' 'award' 'actor' 'album' 'star' 'tv' 'chart' 'number' 'also' 'new' 'oscar' 'top']

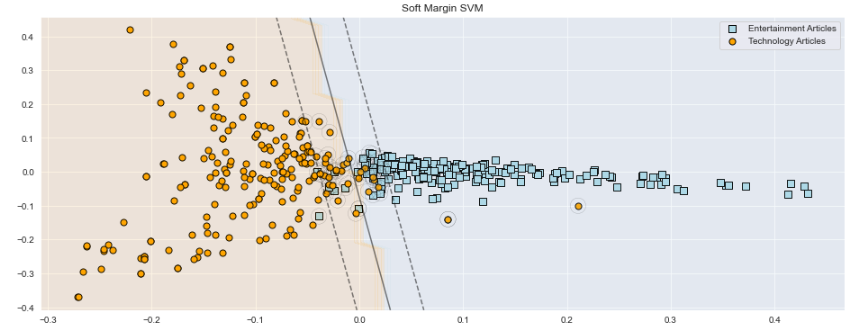
Top 20 words that are most likely to occur in tech news articles: ['said' 'people' 'mobile' 'games' 'software' 'phone' 'net' 'users' 'technology' 'mr' 'microsoft' 'computer' 'broadband' 'virus' 'use' 'new' 'game' 'could' 'digital' 'service']

1. Top 20 words that are most likely to occur in entertainment news articles while maximising constraint: ['film' 'band' 'best' 'actor' 'album' 'chart' 'oscar' 'singer' 'award' 'actress' 'musical' 'star' 'stars' 'comedy' 'awards' 'festival' 'aviator' 'theatre' 'nominated' 'rock']

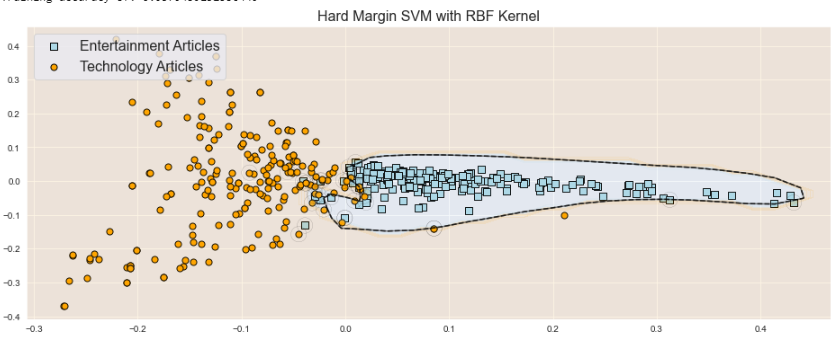
Top 20 words that are most likely to occur in tech news articles while maximising constraint: ['mobile' 'software' 'users' 'games' 'microsoft' 'net' 'technology' 'broadband' 'virus' 'phone' 'computer' 'phones' 'spam' 'mail' 'firms' ‘spyware' 'use' 'online' 'services' 'internet']

By maximising the constraint, we can see that each top 20 list contain words that are more specific/ related to their corresponding news article category. This makes sense as the constraint gives us the probability of seeing a word in one type of news article compared to another type, which in turn generated top 20 words that have more chance of appearing in one news article type over the other. This can be compared to the top 20 words lists generated initally that contain words that are more general and have broader usage in English sentences (Not category specific). Hence, we can say that the top 20 words list generated with the constraint in place describes the two classes better.

1. **SVM**
2. Soft-Margin Linear SVM

****

1. Hard-Margin RBF Kernel

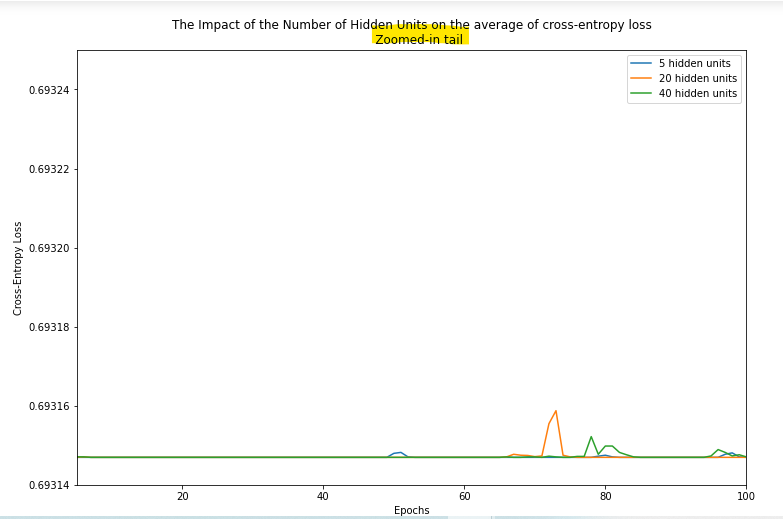
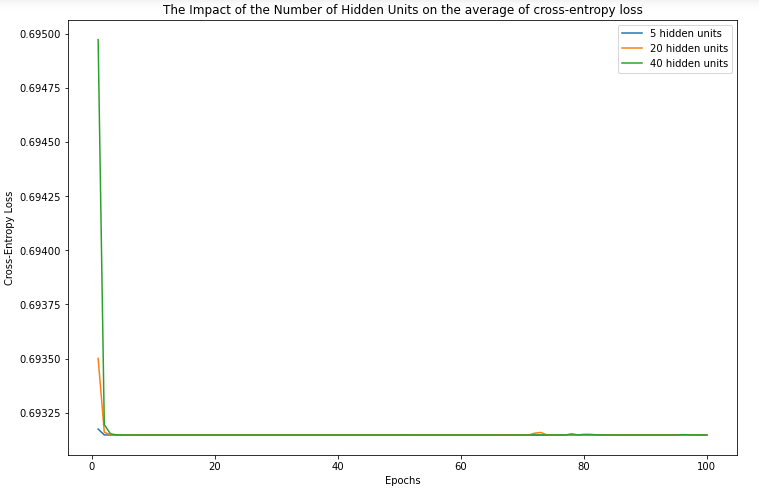
****

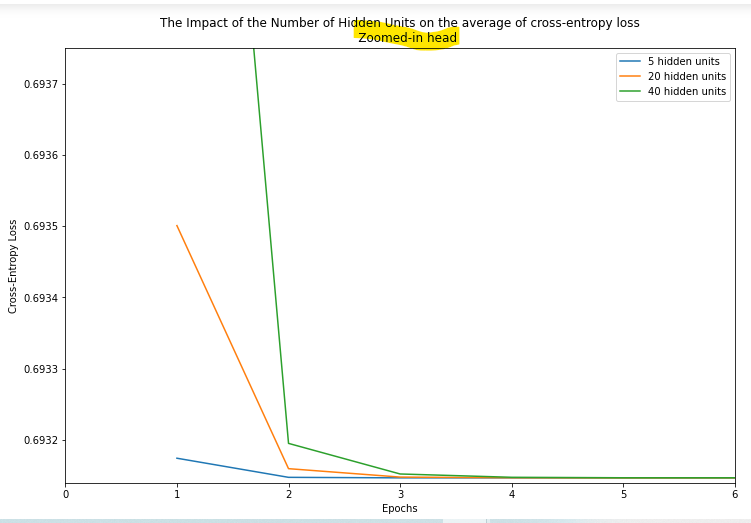
1. Explanation of Results of the Differing SVM Methods

The Soft Margin SVM approach is heavily affected by the parameter C. The misclassification penalty. If C is low then the penalty for a misclassification is light and that they are acceptable for the model. This will result in a larger margin. This can be seen in the above graphs. When we increase the penalty, the margins become narrower. The Hard Margin SVM approach is effected by the kernel that it uses. It is clear that with a very complicated dataset such as the one we have now, that using a linear kernel to define a decision boundary would not achieve great accuracy as there is no straight line we can draw to accurately separate the classes. Especially as the dataset we are in has high dimensionality. RBF uses normal curves around the data points, and sums these so that the decision boundary can be defined by a type of topology condition. Whereas the hard margin SVM the sigmoid kernel, it sections the decision areas in a much more different way.

1. **NN**

* Please refer to the source code for the construction for the required neural network.
* The following plots show the average training cross-entropy loss across different numbers of hidden units.





**The Effect of the Number of Hidden Units**

A higher number of hidden units in the hidden layer seems to have a greater impact on reducing the average entropy loss after the first iteration. After that, the loss incurred by different numbers of hidden units converge at the third iteration. Till around the 46th iteration, the number of hidden units have little influence on the amount of loss. However, there is a surge of cross-entropy loss between the 46th and 61st iterations when 40 hidden units are deployed. Subsequently, there is a minor surge of loss between the 75th and the 18th itertions when 20 hidden units are used, followed by a tiny surge when 5 hidden units are used. Therefore, it appears that neural network with less hidden units tend to be more stable with respect to average cross-entropy loss.

Task Three: Classification Quality Evaluation

1. (i) Training accuracy

(ii)Testing accuracy