NLP Interestship 5.0-Task 2

Logistic Regression

- It is a type of classification algorithm. It is used to predict a binary outcome unlike linear regression where the outcome predicted is contineous.

- It is used on the data which can be represented as a Linear combination of one or more independent variables. [f(x) = b0 + b1*x1 + b2*x2 + …. + bn*xn]

- A sigmoid curve is used to fit the data. The Sigmoid function is 1 / 1 + e^-f(x). It is a mathematical function that takes any real number and maps it to a probability between 1 and 0.

- The sigmoid function forms an S shaped graph, which means as x approaches infinity, the probability becomes 1, and as x approaches negative infinity, the probability becomes 0. The model sets a threshold that decides what range of probability is mapped to which binary variable.

- Suppose we have two possible outcomes, true and false, and have set the threshold as 0.5. A probability less than 0.5 would be mapped to the outcome false, and a probability greater than or equal to 0.5 would be mapped to the outcome true.

Naive Bayes

- The Naïve Bayes algorithm is comprised of two words Naïve and Bayes, Which can be described as:
    - **Naïve**: It is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features.
    - **Bayes**: It is called Bayes because it depends on the principle of Bayes Theorem.

- Bayes' theorem is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability.
    - $P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$
    - **P(A|B)** is the probability of hypothesis A on the observed event B.
    - **P(B|A)** is the probability of the evidence given that the probability of a hypothesis is true.

    - **P(A)** is the probability of hypothesis before observing the evidence.
    - **P(B)** is the probability of Evidence.
- Types of Naive Bayes models:
    - Gaussian NB
        - Each feature in the dataset will have a continuous value distributed according to the normal distribution.
    - Bernoulli NB
        - Here the features are independent boolean variables.
    - Multinomial NB
        - Here samples represent the frequencies, that is how many times the feature occures.
- Assumptions of Naive Bayes Algorithm:
    - It considers features to be Independent. Two events are called independent if the probability of occurrence of one event does not affect the probability of occurrence of the other event.
    - It assumes that each feature contributes equally to the probability.
- Applications of Naive Bayes are: Spam filtering, Sentimental analysis, Text classification, Recommendation systems.

Performance Metrics

- It is a matrix that helps us evaluate which algorithm would serve the best results i.e the results we are expecting for a problem.

- Classification Problems:
    - Confusion Matrix

	ACTUAL		
PREDICTED		POSITIVE	NEGATIVE
	POSITIVE	TRUE POSITIVE	FALSE POSITIVE

		(TP)	(FP)
	NEGATIVE	FALSE NEGATIVE (FN)	TRUE NEGATIVE (TN)

    - FP (False Positive) - Type I error
    - FN (False Negative) - Type II error
    - Trick to remember **(T/F) PREDICTED**
  - Precision
    - Tells the correctness of positive predictions.
    - How many are actually positive out of total positive prediction?
    - $Precision\ = \frac{TP}{TP + FP}$
    - If your precision is low, you might have an Imbalanced dataset or you might have not tuned your model parameters correctly.
  - Recall
    - How many actually predicted positive out of total positive in the dataset?
    - $Recall\ = \frac{TP}{TP + FN}$
    - If your recall is low, you might have an Imbalanced dataset or you might have not tuned your model parameters correctly.
  - Accuracy
    - Correct prediction divided by the total number of predictions.
    - $Accuracy\ = \frac{TN + TP}{TN + TP + FN + FP}$
  - F1 score
    - It is the harmonic mean of precision and recall. It combines both precision and recall.
    - $F1\ Score\ = \frac{2 \times (Precision \times Recall)}{(Precision + recall)}$
    - If your F1 score is high that means your models Precision and Recall are good, when it is low you cannot say whether your precision was less or your recall.