Text Classification using A.I. Techniques

***Abstract* —** The aim of this paper is to apply different text mining tools on a piece of book summary and categorize it into the genre where it best fits. The book summary is preprocessed thoroughly before fitting into various models. The preprocessing task mainly involved three major steps. First, all the special characters are removed and every uppercase character is converted into lowercase counterpart. Then every word is broken down into its morphological root. At the last step, all the most commonly used words are also removed from book summaries as it will not help in the task of prediction. Feature extraction from the preprocessed book summary is accomplished using Tf-idf statistics. Finally at the last step, different statistical models, such as Logistic Regression and Support Vector Machines are used for drawing insights, in order to find the underlying relationship between the genre of the book and its summary. A prediction accuracy of 79.55% was achieved using the Support Vector Machine with radial basis function as its kernel.

***Keywords* —** Text mining, Lemmatization, Stemming, Logistic Regression, Support Vector Machines.

I. INTRODUCTION

Text Classification is the process of classifying text according to its content. It is also known as Text Categorization where tags or categories are assigned to a set of text. It is one of the basic tasks in Natural Language Processing (NLP). In this new era of computer science, NLP has become indispensable as it gives the machine the privilege to understand text which was so far an exclusive property of humans.[1] This task falls at the intersection of information retrieval and machine learning techniques and has witnessed a thriving interest from researchers and data scientists all over the world. Due to the ubiquity of textual information, text classification finds huge applications in the field of information retrieval, sentiment analysis, text summarization, question answering, and many more.

Text classification is a form of text mining. To say it broadly, text mining is a task of deriving information from text data by analyzing patterns within it. But text mining is again a subset of data mining. The concept of data mining has various utilities. One of the most important principles that employ data mining is the analysis of data. When we are given a large amount of data, for effective analysis, we need to categorize the data into sections, such that the data in each specialized section have similar features. Text is everywhere, like in every place over the web or anywhere. We can get huge insights from them if we can extract them properly.

Text classification is also a part of a relatively larger field of study, known as text analysis. Text analysis is already helping fields such as Marketing, Product Management, Academia, and Governance and they are already leveraging the process of analysing and extracting information from textual data. Classification of books in libraries and the segmentation of articles in news are essentially examples of text classification. Text classification is really useful when we have to map or tag a large set of textual data. So starting from search engines to social media, and in marketing industries, it is really helpful. As we can see the marketing industry has become very targeted, so markets are targeting people personally to grow their business, For that purpose, text analysis and getting the sentiment out of it is very important. Thus industries can get the benefit as they can look into a previous set of chat and new set and analyse everything, which helps in better engagements.

In [2], a binary text classifier had been implemented, where the newly classified instances of data and their predicted classes were added to the training data set so the enhanced data set would be able to train the classifier for future unclassified instances of data. This allows the classifier to increase the accuracy each time the classifier was used for similar problems. The primary limitation of classifier used in [2] is that it can perform binary classification only, but in the real world a classifier needs to be dynamic and a classifier should not constrain the data set to any specific number of classes. This limitation was removed in [3]. The work in [2],[3] was the direct motivation for this paper.

An addition that can be added to the classifier is to incorporate lemmatization and stemming into the training and classification processes. Lemmatization was described in [4] as "finding the Lexical Headword of a given word form" it reduces a word to its morphological root and stemming "a computational procedure which reduces all words with the same root to a common form" [5]. This would allow the classifier to give better accuracy as it would allow the classifier to treat different forms of the same word as the same concept.

This paper emphasizes on classifying a book summary into its respective genre. The objective is to build a text classifier model that first gets trained by a large set of data of book summary. It then understands the tags or labels or genres associated with the summaries. Then it is tested using a set of data. This text classification finds huge real-life applications.

The rest of the paper is structured as follows, Section 2 contains a short overview of related research work in the field of text classification and states a few examples of text classifiers implemented using different known techniques. Section 3 provides a detailed description of the theoretical principles and concepts behind the proposed text classifier. Section 4 offers a working description of the proposed classifier, stepping through the algorithms behind its implementation. Section 5 is a discussion of the results obtained using the classifier and concludes the paper with a summary of the work and its future scope.

II. BACKGROUND STUDY

Text classification or text categorization is a process of classifying a group of text into different tags which were predefined. Broadly, text classification falls under the domain of text mining which is a large domain. In paper [6], it is said that text mining, in the field of information retrieval helps to find out the pattern or the hidden search result, where it is not even explicitly known or indicated anyhow.

Text mining is intended to extract the information that was not known before by extracting it from different text-based sources [7]. Any semi-structured or unstructured data can be handled by text mining unlike data mining. Text categorization falls under the category of both information retrieval and machine learning. So tools and techniques derived from both these domains are helpful in text categorization [8].

Among different text classifiers, namely- SVM, Naive –Bayes, K-NN, etc. SVM is found to be the most optimal and useful. The SVM was first introduced by Vapnik and since then it had become an integral part in text classification, tone recognition, image classification, object classification, etc. [9]. Then depending upon the cost parameter and kernel parameter, SVM's performance is changed significantly.

Another important model which has been used is Logistic Regression. Logistic Regression (LR) is one of the most important statistical and data mining techniques used by statisticians and researchers for the analysis and classification of binary and proportional response data sets [10]. Probabilities and extend to multi-class classification problems are the advantages provided by the LR model. But it is mainly a binary classifier. LR is primarily a binary classifier which means it is a very appropriate regression analysis technique when the dependent variable is binary. One-vs-Rest (OVR) is a modification to LR where each model is trained for each of the multiple classes.

Tf-idf numeral statistic is also used on the training-testing data set. In paper [11], it is said that, tf-idf gives those specific words which help identifying a specific document. Tf or term frequency gives the frequency of a word appearing in a document and on the other hand the idf or inverse document frequency assigns larger weight to less frequent words and less weight to most frequent words.

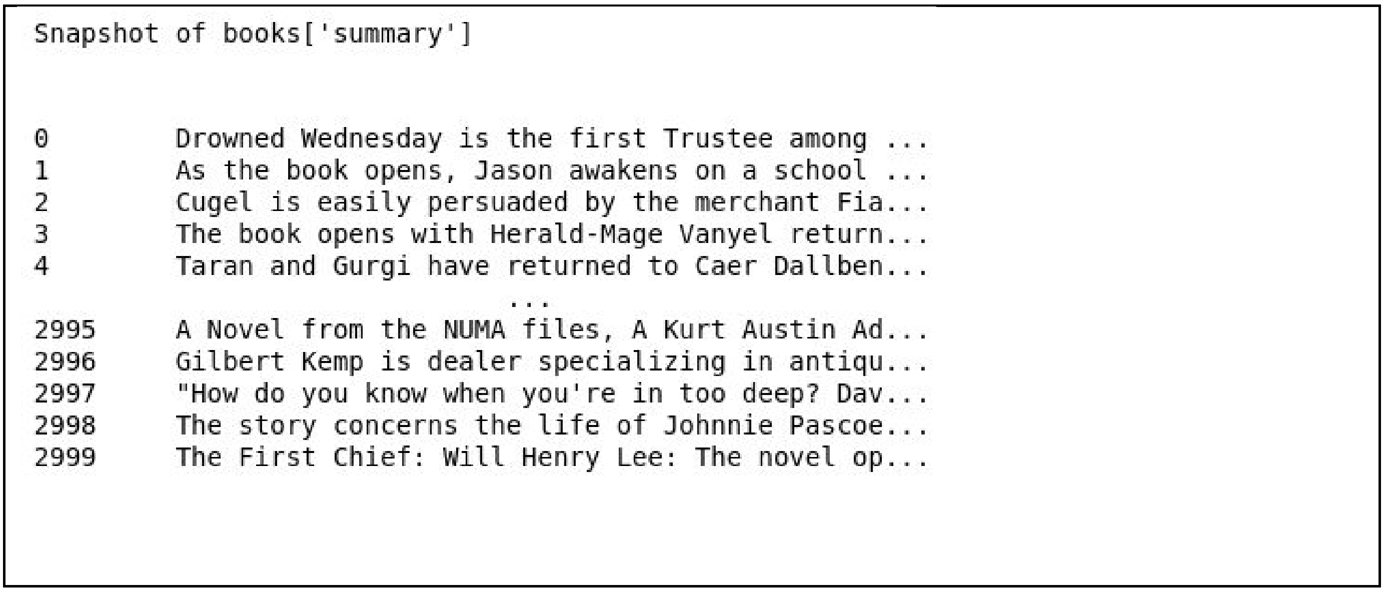
In paper [12], Joachims proposed the use of Support Vector Machines (SVMs) for text classification and explained that SVMs could be more optimal than other machine learning methods such as k-NN classifiers and Naïve-Bayes classifiers. In [13], Cristianini discussed and showed light upon the working principles and applications of SVMs in text classification.

A vital role has been played by SVM as it classifies the classes using a hyperplane and accordingly it is divided.

III. PROPOSED METHODOLOGY

Data set which has been used was originally collected by Dr. David Bamman and Dr. Noah Smith and CMU hosted it, which can be found at [17]. After performing vigorous feature engineering, the data set finally contained 500 rows dedicated to each of the six genres i.e. “Crime Fiction", "Fantasy", "Historical novel", "Horror", "Science Fiction"and "Thriller".

Figure 1 shows the snapshot of the book's summary before data set cleaning and preprocessing begins.



```
Snapshot of books['summary']


0       Drowned Wednesday is the first Trustee among ...
1       As the book opens, Jason awakens on a school ...
2       Cugel is easily persuaded by the merchant Fia...
3       The book opens with Herald-Mage Vanyel return...
4       Taran and Gurgi have returned to Caer Dallben...
                              ...
2995    A Novel from the NUMA files, A Kurt Austin Ad...
2996    Gilbert Kemp is dealer specializing in antiqu...
2997    "How do you know when you're in too deep? Dav...
2998    The story concerns the life of Johnnie Pascoe...
2999    The First Chief: Will Henry Lee: The novel op...
```

Figure 1. Snapshot of Book’s summary before data preprocessing

The working methodology starts with acquiring the data set and cleaning and preprocessing it accordingly. Everything apart from alphabets is removed from the book's summary.

In Figure 2, we see the frequency distribution of the 100 most frequently used words in the book summary across all book summaries in the data set.

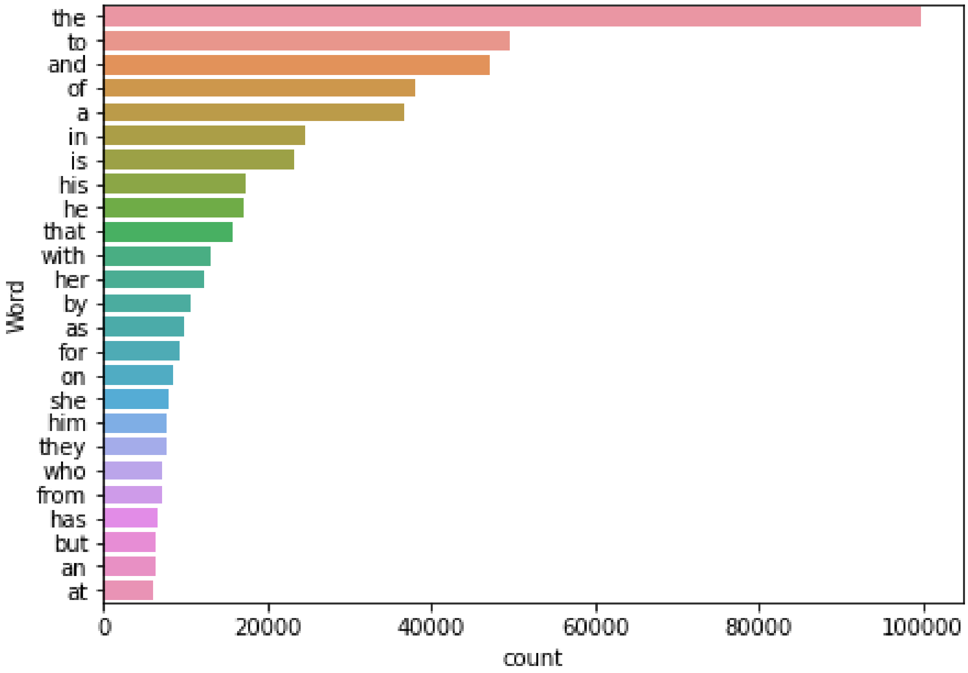


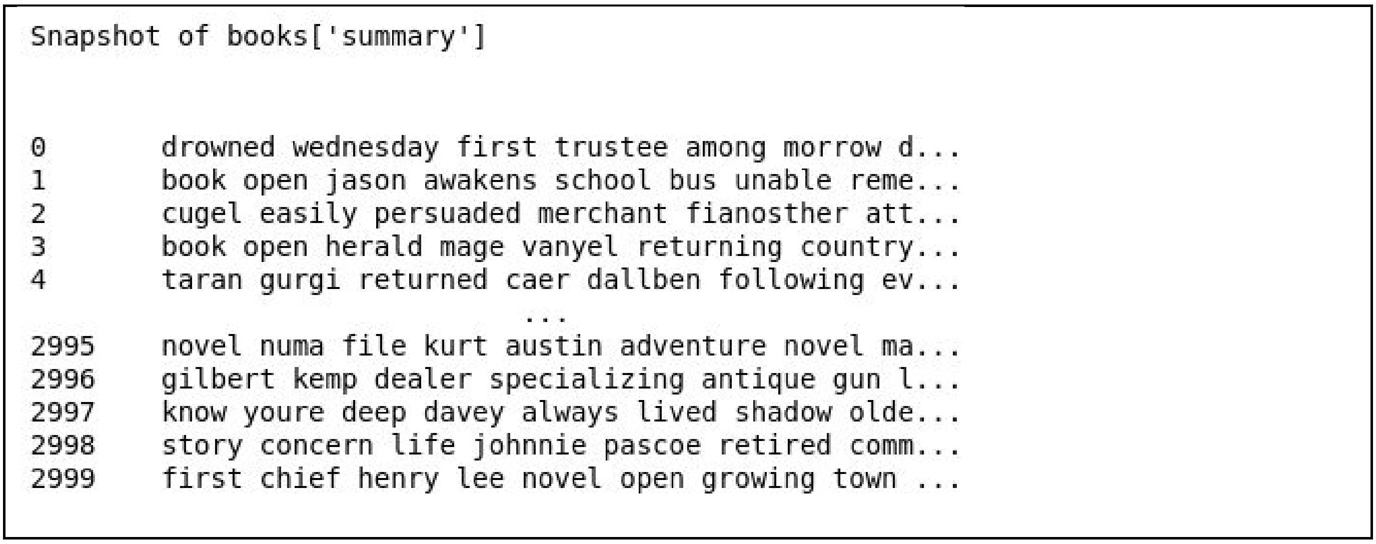
Figure 2.The frequency distribution of words before preprocessing

The most commonly used words(or stop Words) in english language that appear in almost every literary piece are also removed gradually.

Then lemmatization and stemming is performed on the data set. These two concepts have been incorporated in the data set to preprocess it.

Lemmatization falls under the domain of Natural Language Processing (NLP). In lemmatization, different versions of the same words are grouped into one. The task of the automatic lemmatization is to find a basic word form or a “lexical headword” of a given word. The use of lemmatization preprocessing has been used to be especially important for the highly inflected languages [14] in the various tasks of natural language processing, such as keyword spotting or information retrieval. The number of processed words - different forms of a particular word can be treated as one (its base form) - so it is very advantageous in highly inflected languages and lemmatization helps in achieving this.

Figure 3, shows the result of lemmatization on the book's summary.



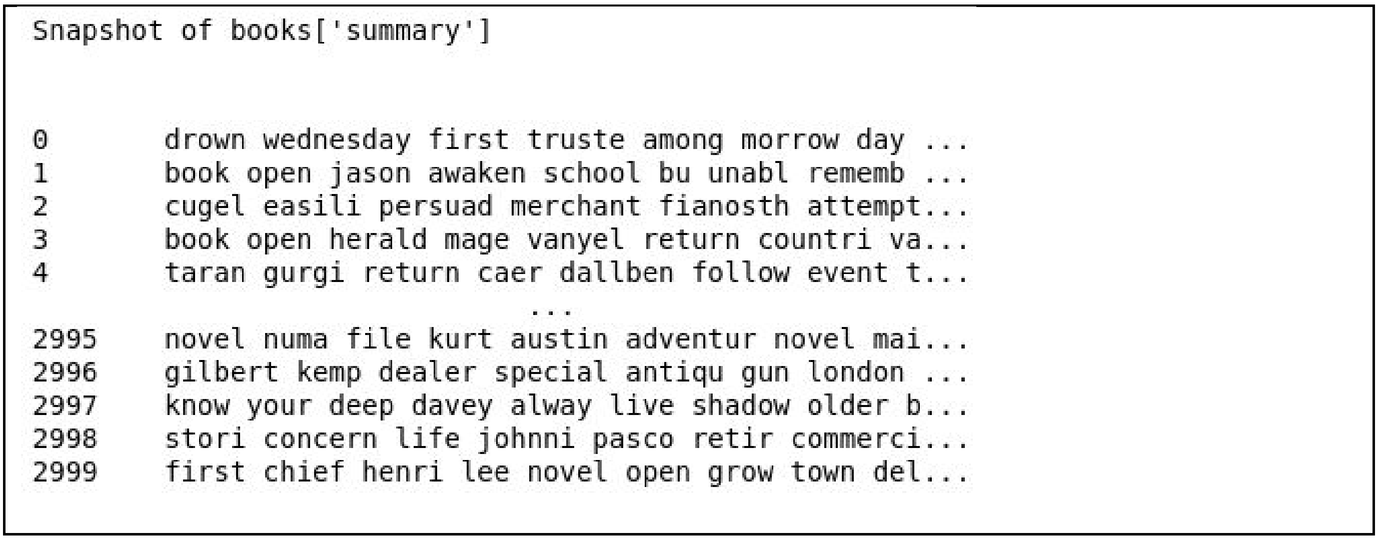
```
Snapshot of books['summary']


0       drowned wednesday first trustee among morrow d...
1       book open jason awakens school bus unable reme...
2       cugel easily persuaded merchant fianosther att...
3       book open herald mage vanyel returning country...
4       taran gurgi returned caer dallben following ev...
                              ...
2995    novel numa file kurt austin adventure novel ma...
2996    gilbert kemp dealer specializing antique gun l...
2997    know youre deep davey always lived shadow olde...
2998    story concern life johnnie pascoe retired comm...
2999    first chief henry lee novel open growing town ...
```

Figure 3. Book’s summary after performing lemmatization

Stemming is a process of producing morphological variants of the root word. It is done by mapping multiple words into one stem even if the stem doesn’t give any valid word or meaning. Stemming is an integral part in all information retrieval systems. It has been seen that most of the time the morphological variants of words have similar interpretations and can be considered as the same [15]. As the meaning is the same but the word form is different, so it is necessary to identify each word form with its root form. Porter stemmer is used for this purpose.

It is based on the idea that the suffixes in the English language are almost made up of a combination of small and simple suffixes. It has five steps, and in each step, rules are applied until one of them passes the conditions. If a rule is accepted, the suffix is removed then, and the next step is performed. The rule is something like: <Condition><suffix> □<new suffix>

Figure 4, shows the result of stemming on the book's summary.



```
Snapshot of books['summary']


0       drown wednesday first truste among morrow day ...
1       book open jason awaken school bu unabl rememb ...
2       cugel easili persuad merchant fianosth attempt...
3       book open herald mage vanyel return countri va...
4       taran gurgi return caer dallben follow event t...
                              ...
2995    novel numa file kurt austin adventur novel mai...
2996    gilbert kemp dealer special antiqu gun london ...
2997    know your deep davey alway live shadow older b...
2998    stori concern life johnni pasco retir commerci...
2999    first chief henri lee novel open grow town del...
```

Figure 4. Book’s summary after performing stemming

Figure 5 shows the most frequently used words in book summary across all book summaries in the data set, once the task of data cleaning and preprocessing has been carried out successfully.

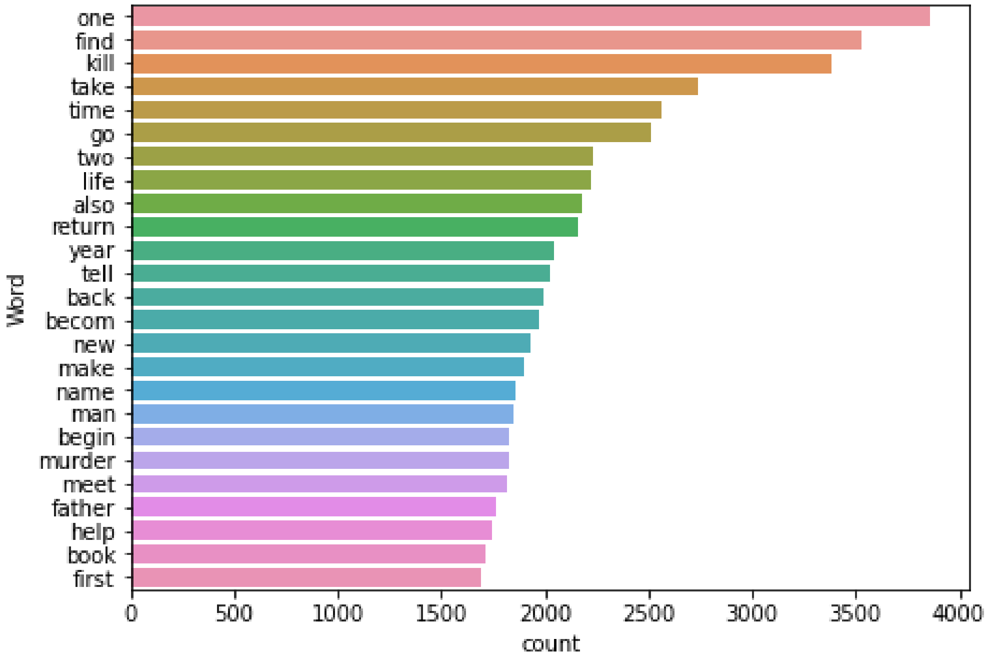


Figure 5.The frequency distribution of words after preprocessing

Tf-idf is performed on the training-testing data set then. This numerical statistic reflects the fact that how a word is important in a document.

Tf_Idf(t,d,D) = tf(t,d).idf(t,D)

Where,

Tf(t,d) = $f_{t,d}$

Idf(t,D) = log(N/{d∈D : t ∈ d})

With,

$f_{t,d}$: Frequency of the term t in document d.

N: Total number of documents.

{d∈D : t ∈ d}: Number of documents where the term t appears.

Once the data has been preprocessed thoroughly, different machine learning models were fed the preprocessed data in order to perform the task of classification.

Logistic Regression is used when the dependent variable is categorical**.** Simple Logistic regression is only a binary classifier, which means it cannot handle target vectors with more than two classes. In one-vs-rest logistic regression (OVR) a separate model is trained for each class predicting whether an observation belongs to that class or not [16].

SVMs try to find a separating line (or hyper plane) between data of two classes. SVM is an algorithm that takes the data as an input and outputs a line that separates those classes if possible. According to the SVM algorithm, points closest to the line from both the classes are found. These points are called support vectors. Then the distance between the line and the support vectors are measured. This distance is called the margin. The margin must be maximized. Thus a decision boundary is chosen in such a way that the separation between the two classes (that street) is as wide as possible.

The linear, polynomial and RBF or Gaussian kernel are simply different in case of making the hyperplane decision boundary between the classes. The kernel functions are used to map the original data set (linear/nonlinear) into a higher dimensional space with a view to making it a linear data set. Usually linear and polynomial kernels are less time consuming and provide less accuracy than the Rbf or Gaussian kernels.

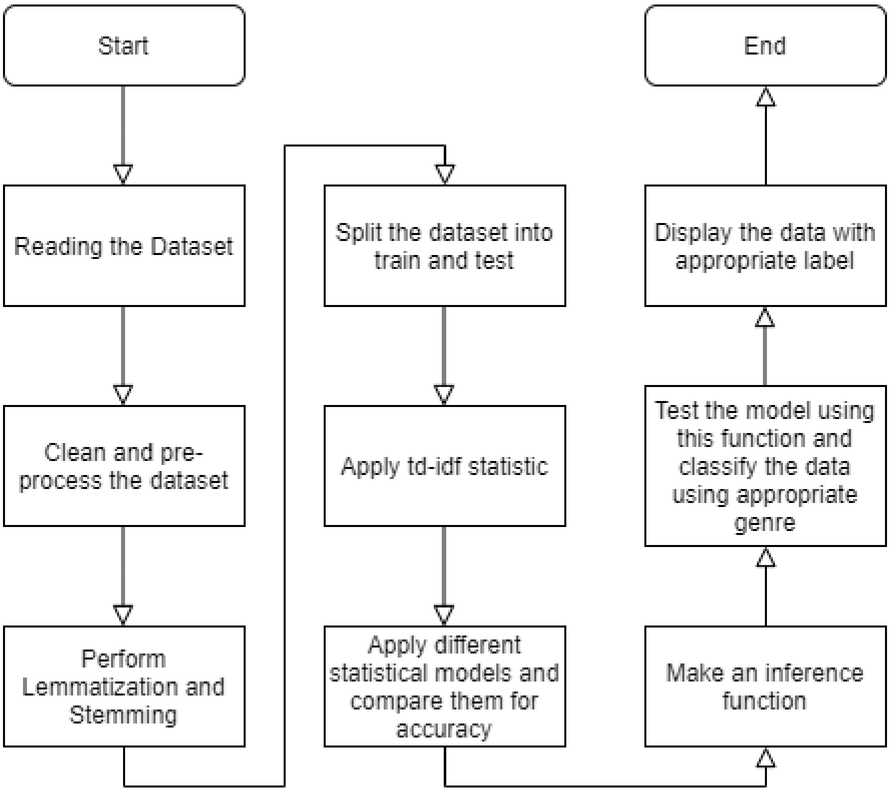

Start
Reading the Dataset
Clean and pre-process the dataset
Perform Lemmatization and Stemming
Split the dataset into train and test
Apply td-idf statistic
Apply different statistical models and compare them for accuracy
Make an inference function
Test the model using this function and classify the data using appropriate genre
Display the data with appropriate label
End

Figure 6. Flowchart representing the proposed methodology

The flowchart for the entire methodology is shown in Figure 6.

IV. RESULT AND ANALYSIS

Two different percentage ratios - an 80-20% split and an 85-15% split were opted for carrying out the training and testing phase. Once the data set was successfully split, the tf-idf statistic was used to extract features from the book summary. Since the six genres were categorical variables, they were encoded into six unique integers i.e. from 0 to 5.

A. Logistic Regression

The first classifier that has been implemented to perform multi-class classification was the Logistic Regression. In Logistic Regression for multi-class classification, a separate model is trained for each and every predicted class to determine whether an observation is in that class or not. It presumes each classification problem is independent.

Table 1 shows the classification report and accuracy score for the data set when an 80-20% split was performed for train and test phase.

Table 1. Classification Report and Accuracy score for 80-20% split

Genre	**Precision**	**Recall**	**f1-score**	**Dedicated rows**
Crime Fiction	0.79	0.82	0.80	102
Fantasy	0.69	0.78	0.73	89
Historical Novel	0.83	0.86	0.84	110
Horror	0.80	0.73	0.76	100
Science Fiction	0.82	0.80	0.81	97
Thriller	0.78	0.71	0.74	102
Accuracy	78.5%			600

Table 2 shows the classification report and accuracy score for the data set when an 85-15% split was performed for train and test phase.

Table 2. Classification Report and Accuracy score for 85-15% split

Genre	**Precision**	**Recall**	**f1-score**	**Dedicated rows**
Crime Fiction	0.73	0.78	0.75	67
Fantasy	0.80	0.83	0.81	82
Historical Novel	0.85	0.85	0.85	86
Horror	0.74	0.74	0.74	72
Science Fiction	0.82	0.86	0.84	64
Thriller	0.75	0.66	0.70	79

Accuracy	78.4%			450

B. Support Vector Machine with Linear Kernel

The second classifier that has been implemented to perform the demanding task of classifying is the Support Vector Machine with linear kernel. Support Vectors with linear kernel are used when the classes can be separated by drawing a straight line as the hyperplane.

Table 3 shows the classification report and accuracy score for the data set when an 80-20% split was performed for train and test phase.

Table 3. Classification Report and Accuracy score for 80-20% split

Genre	Precision	Recall	f1-score	Dedicated rows
Crime Fiction	0.81	0.79	0.80	102
Fantasy	0.69	0.72	0.70	89
Historical Novel	0.81	0.87	0.84	110
Horror	0.74	0.73	0.73	100
Science Fiction	0.79	0.75	0.77	97
Thriller	0.73	0.70	0.71	102
Accuracy	76.3%			600

Table 4 shows the classification report and accuracy score for the data set when an 85-15% split was performed for train and test phase.

Table 4. Classification Report and Accuracy score for 85-15% split

Genre	Precision	Recall	f1-score	Dedicated rows
Crime Fiction	0.77	0.73	0.75	67
Fantasy	0.78	0.84	0.81	82
Historical Novel	0.86	0.84	0.85	86
Horror	0.70	0.75	0.72	72
Science Fiction	0.84	0.83	0.83	64
Thriller	0.74	0.70	0.72	79
Accuracy	78.2%			450

C. *Support Vector Machine with Radial Basis Function Kernel*

The third classifier that has been implemented was the Support Vector Machine with Radial Basis Function Kernel. When the data cannot be separated with a straight line, Non-Linear Support Vector Machines are used. For this task, kernel functions are used to transform data into another dimension so that data can be classified.

Table 5 shows the classification report and accuracy score for the data set when an 80-20% split was performed for train and test phase.

Table 5. Classification Report and Accuracy score for 80-20% split

Genre	**Precision**	**Recall**	**f1-score**	**Dedicated rows**
Crime Fiction	0.84	0.78	0.81	102
Fantasy	0.67	0.76	0.71	89
Historical Novel	0.84	0.85	0.84	110
Horror	0.77	0.72	0.74	100
Science Fiction	0.83	0.77	0.80	97
Thriller	0.71	0.75	0.73	102
Accuracy	77.5%			600

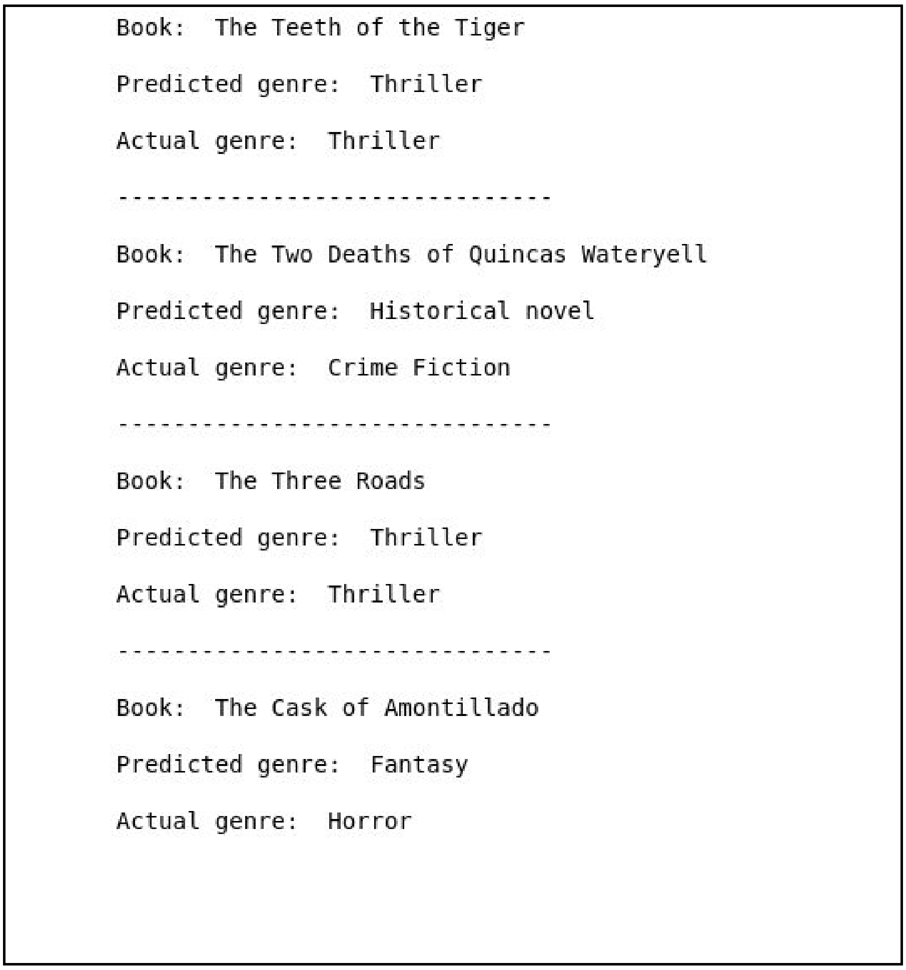
Table 6 shows the classification report and accuracy score for the data set when an 85-15% split was performed for train and test phase.

Table 6. Classification Report and Accuracy score for 85-15% split

Genre	**Precision**	**Recall**	**f1-score**	**Dedicated rows**
Crime Fiction	0.77	0.75	0.76	67
Fantasy	0.80	0.88	0.84	82
Historical Novel	0.87	0.85	0.86	86
Horror	0.73	0.75	0.74	72
Science Fiction	0.86	0.86	0.86	64
Thriller	0.74	0.68	0.71	79
Accuracy	79.55%			450

It is evident that a much higher level of accuracy is obtained in case of Support Vector Machine with radial basis function as kernel when trained on the data set with an 85-15% training and testing split.

In Figure 12, the comparison between the book's actual genre and predicted genre can be seen.



```
Book:  The Teeth of the Tiger

Predicted genre:  Thriller

Actual genre:  Thriller

------------------------------

Book:  The Two Deaths of Quincas Wateryell

Predicted genre:  Historical novel

Actual genre:  Crime Fiction

------------------------------

Book:  The Three Roads

Predicted genre:  Thriller

Actual genre:  Thriller

------------------------------

Book:  The Cask of Amontillado

Predicted genre:  Fantasy

Actual genre:  Horror
```

Figure 12. Comparison between actual genre and predicted genre

V. CONCLUSION

In this paper, accuracy was improved using Support Vector Machines with Radial Basis function as its kernel function. Linear Support Vector Machines are efficient only when classes can be separated by a straight line. In the absence of such linearity, kernel functions are used to transform the data into another dimension where a straight line can be drawn as the class boundary. The table 7 depicts a comparative study of the accuracy achieved by the algorithms that have been used - Logistic Regression and Support Vector Machines. Accuracy was also improved by incorporating advanced preprocessing techniques such as stemming, lemmatization and stop-words removal technique which was missing in [3] and we have also replaced the simple weighting factor such as simple frequency of occurrence used in [2] with advanced weighting factor such as tf-idf.

Table 7. Test results for different algorithms

Algorithm	Train-Test Split	Accuracy
Logistic Regression	80-20%	78.5%

Linear - Support Vector Machine	80-20%	76.33%
Non Linear - Support Vector Machine	80-20%	77.5%
Logistic Regression	85-15%	78.44%
Linear - Support Vector Machine	85-15%	78.22%
Non Linear - Support Vector Machine	85-15%	79.55%

The advantage of the used methodology is that it is computationally cheap and uses only the book summary to determine the genre. It uses basic natural language processing techniques for preprocessing the book summary. This paper does not limit itself to use either stemming or lemmatization for breaking every word in the book summary into its morphological root, but incorporates both. The tf-idf statistics for feature extraction emerges as the finest for this task. In the paper [18], classification of book into its genre was done using its title and cover image but using the methodology described in this paper an accuracy of 79.55% was achieved in the task of labeling the genre of the book only from its summary. However performance can still be improved if properly hyper tuned state of the art neural networks are used as the classifiers.

VI. REFERENCES

[1] Power, R., Chen, J., Karthik, T., & Subramanian, L. (2010). Document classification for focused topics. In Artificial Intelligence for Development - Papers from the AAAI Spring Symposium, Technical Report (Vol. SS-10-01, pp. 67-72)

[2] A. Sarkar, S. Chatterjee, W. Das, D. Datta, “Text Classification using Support Vector Machine”, International Journal of Engineering Science Invention, Vol. 4 Issue 11, November 2015, pp. 33 – 37.

[3] Anurag Sarkar, Debabrata Datta,"A Frequency Based Approach to Multi-Class Text Classification", International Journal of Information Technology and Computer Science(IJITCS), Vol.9, No.5, pp.15-22, 2017. DOI:10.5815/ijitcs.2017.05.03

[4] Kanis J., Skorkovská L. (2010) Comparison of Different Lemmatization Approaches through the Means of Information Retrieval Performance. In: Sojka P., Horák A., Kopeček I., Pala K. (eds) Text, Speech and Dialogue. TSD 2010. Lecture Notes in Computer Science, vol 6231. Springer, Berlin, Heidelberg

[5] LOVINS, J.B. Development of a Stemming Algorithm. Mechanical Translation and computation Linguistics. 11 (1) March 1968 pp 23-31.

[6] R. Sagayam, A survey of text mining: Retrieval, extraction and indexing techniques, International Journal of Computational Engineering Research, vol. 2, no. 5, 2012.

[7] S. M. Weiss, N. Indurkhya, T. Zhang, and F. Damerau, Text mining:predictive methods for analyzing unstructured information. Springer Science and Business Media, 2010.

[8] N. Zhong, Y. Li, and S.-T. Wu, "Effective pattern discovery for text mining," IEEE transactions on knowledge and data engineering, vol. 24, no. 1, pp. 30-44, 2012.

[9] Maalouf, Maher. (2011). Logistic regression in data analysis: An overview. International Journal of Data Analysis Techniques and Strategies. 3. 281-299. 10.1504/IJDATS.2011.041335.

[10] Indra, S. & Wikarsa, Liza & Turang, Rinaldo. (2016). Using logistic regression method to classify tweets into the selected topics. 385-390. 10.1109/ICACSIS.2016.7872727.

[11] Qaiser, Shahzad & Ali, Ramsha. (2018). Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. International Journal of Computer Applications. 181. 10.5120/ijca2018917395.

[12] T. Joachims, "Text Categorization with Support VectorMachines: Learning with Many Relevant Features",Technical Report 23, Universitat Dortmund, LS VIII,

[13] N. Cristianini, "Support Vector and Kernel Machines",Tutorial at the 18th International Conference on Machine Learning, June 28, 2001.

[14] K. Ming Leung, "Naive Bayesian Classifier", Polytechnic University Department of Computer Science/Finance and Risk Engineering, 2007.

[15] Jivani, Anjali. (2011). A Comparative Study of Stemming Algorithms. Int. J. Comp. Tech. Appl.. 2. 1930-1938.

[16] Peng, Joanne & Lee, Kuk & Ingersoll, Gary. (2002). An Introduction to Logistic Regression Analysis and Reporting. Journal of Educational Research - J EDUC RES. 96. 3-14. 10.1080/00220670209598786.

[17] http://www.cs.cmu.edu/~dbamman/booksummaries.html#:~:text=This%20dataset%20contains%20plot%20summaries,author%2C%20title%2C%20and %20genre. Access time 31st August,2020.

[18] https://pdfs.semanticscholar.org/d0d0/096d307a6da1332153b9cb8a72c29df38f87.pdf Access time 31st August,2020.