Genre Identification on 19th century English Fiction Books, a subset of Gutenberg Corpus

1st Aishwarya Suresh, 2nd Deeksha Ramakrishna, 3rdPriyamvada Bhardwaj, 4thSathya Sudha Murugan, and 5th Snigdha Mohapatra

***Abstract*—This report is based on the task of identification the genre of the given fiction books as part of semester team project.The given data-set is 19th century English fiction books which are subset of the Gutenberg's corpus. In order to build a classification model which is able to predict the genre of any given English fiction book , this project builds Bag of words as a base line model and another model with the combination of hand crafted features as meta data features, sentiment analysis and emotion analysis. This also handled the issue of huge class imbalance problem in the given data-set.At the end ,for the handcrafted feature based model , model selection and evaluation using cross validation is performed and is compared with the bag of words model.**

***Index Terms*—Text Classification, Supervised Machine Learning, Genre Identification, NLP**

I. Introduction

Since books are lengthy in nature labelling the genre of the book after reading the complete book without the presence of preface/abstract etc is time consuming and we often relies on human experts for these task.As part of this project our team builds a predictive model which is able to auto label the genre of any given English fiction book without the help of human expert thus saving the mundane human efforts.There exists such models in real life such as "Audible" from Amazon where each book is automatically classified in the correct genre or digital libraries of books.The given data set is given with the labels with the books making it suitable for supervised learning, in addition it has Author Name and Book names as well.There are total 9 Genres of the fictional books are provided which makes its a multi-class problem, no book belongs to more than one genre as per the given labels hence ruling out multi label classification problem.

In order to teach the classification model the features of the fiction books we decided to take two different approaches, 1. We decided to use the bag of words approach as assuming any book which has key words like "horse", "man", "rider" shall most probably a Western story than a romantic book while a romantic book shall have words like "love", "romance", "forever" more than in any genre of the book. 2.In the other method we decided to use the features such as meta data features like average length of a book , average number of common words, average number of sentence /words etc in a book to use as a distinguishable features of a fictional book. The assumption was that romantic books tends to be shorter than the Literary books. In addition to these we tried to run the emotion and sentiment analysis on the books and combined all these as hand crafted features to feed to models to learn. In these our team tried to extract those meaningful features which were fitting with the elements of the fiction [9].

In addition to it additional problem of class imbalance has been addressed.

We would like to see if second method produces reasonable accuracy and classification performance and does it perform better than bag of words method.

II. Data Set

The data-sets are the books which are in the form of HTML files which mostly consists of the pages of the books rather than the complete books. There are total of 996 books for which the labels are provided along with the Author Names and Book Names in a separate file.

For bag of words methods we prepossessed data, since books are lengthy and there were lot of common words [1]like "is", "your", "be", "is" and common English names like "Merry", "John" etc which were crowding the key words, we removed it along with the HTML tags and kept only the nouns and verbs.WE removed all the punctuation and numbers or any extract spaces etc to extract the nouns and verbs only which was done using the nltk's "pos tag" and "lemmatization" technique since we wanted the word as "sail" wherever "sailed", "sailing" was mentioned.

To analyse and extract the meta data features we wanted to preserve the original text as it is, hence we only removed the HTML tags to process the data further.

For text based Features extraction like sentiment analysis or emotion analysis etc also we needed pre processed data hence we followed the same steps as we did for bag of words approach.

Another challenge here was the Length of the book pages for which we used pandas to process the cleaning . Common exploratory analysis to check for data types, null values etc was performed.

While trying to get insights of the data, we first get to see the target data imbalance. There is a huge data imbalance in between majority and minority class. The majority class has around 700 books of Literary genre whereas the minority class Allegories has only 2 and the rest of the classes were falling in between.

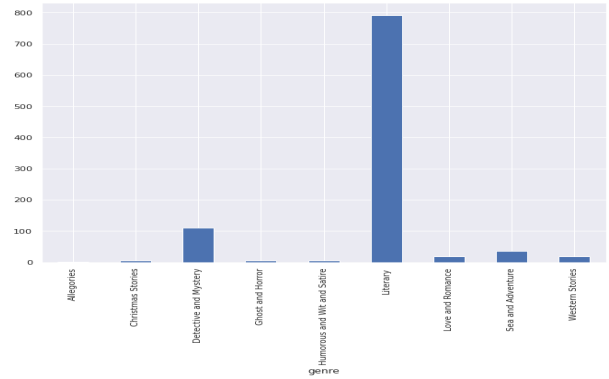

Fig: Total number of books in each genre

Another insight was found after cleaning the data by using word clouds based on only extracting verbs, nouns, etc, which we wanted to use potentially for emotion analysis and bag of words.By using verbs the horror genre books only has ”suspect”, ”fear” , ”hallucination”, ”priest”, ”question” like words that are more common and can predict the better sentiments of the book. Deductive and mystery genre has ”mystery” as a keyword.We generated the word clouds for verbs and nouns separately to get more useful keywords.

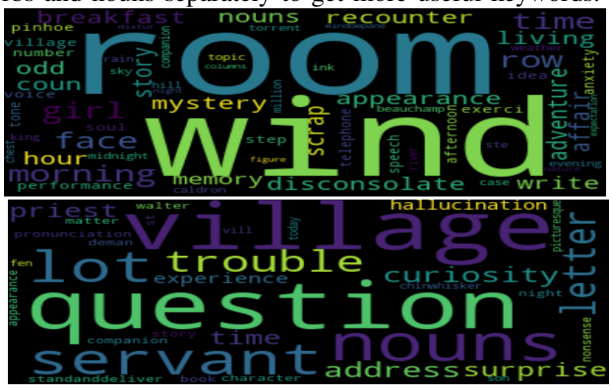

Fig: Word cloud of detective books and horror books

One more insight about the data is that, all the genres except 2 contains only a single book by a author however in literary genre there are authors who has written more than 15 books as well. However in most of the genre, the number of books present are mostly 1 and each written by individual authors.

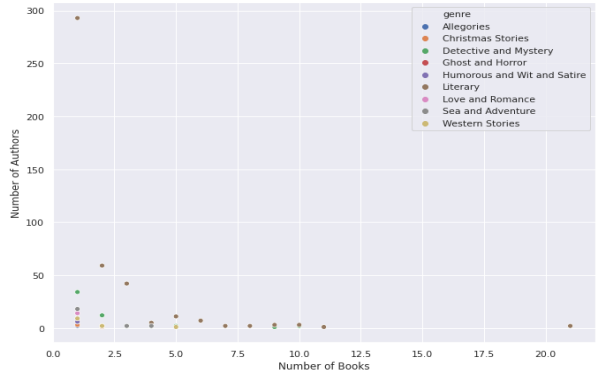

Fig: Number of authors vs number of books written in each genre

In the below graph it is shown that for Literary genre , number of books written by a single author are far greater than 1

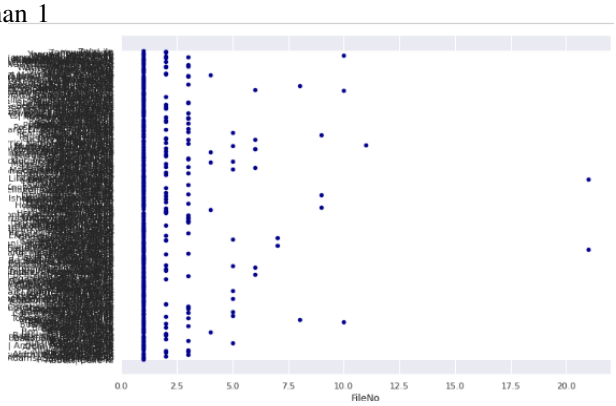

Fig: Number of authors vs number of books written in Literary

Also, there are 45 authors who had written books in more than one genres.

III. Concept

Initially, we decided to extract features from the books which makes it a fictional book and in turn captures the characteristics of the genre of the fiction.

We considered the given 5 elements of fiction, Plot, Setting, Character, Point of View,Style and Theme.

Considering all these we decided to try 2 kind of methods- 1. Bag of words only, where we try to find the keywords in each genre to distinguish.Since this gives the keywords specific to the books this is a useful method to test the data upon, as we have already seen in the word clouds. 2. We considered extracting meta data features and combining those in qualitative form. This also can give some useful features as we have seen in EDA and also it can be used as a characteristics as element of fiction. [10] [11] 3. We tried to capture the emotion/sentiments of the books and capture this also in quantity.If fits into one of the elements of fiction as can give plot information or overall sentiment of the book, such as horror genre shall have more negative emotions than positive and romantic books shall have more joy or positive sentiments. [11]

For each element we considered extracting features for every methods,

Character - there are flat and round characters type in fiction. we decided to extract the proper nouns which shall include the names of the characters and places and run the sentiment /emotion analysis on it to capture how major character is feeling to approximate the character analysis. Assumption here is that horror fiction genre shall have fewer characters and most of the emotions shall be fear and sadness, on the other hand literary genre shall have more number of characters and there would be range of emotions felt by the dominant characters which shall be a distinct point to learn for the model [3] [4].

Plot- We tried to capture the plot setting in terms of context. However we had to use the bag of words for it. The assumption was that the adventure books shall have key words like ’sailing’ and western stories shall have words like horse, rider , man etc, which combined with the sentiment analysis

on overall book/page which shall give us the feeling /plot of the books.

Point of View - To capture this the assumption was to get the number of occurrence of pronouns, nouns and proper nouns, and that in first person directive pronouns like "I" count shall be higher than the number of nouns or proper nouns.

Setting- For general setting the assumption was to get the proper nouns and digits which shall include the places(names and time periods) .

Style- For author writings style, we tried to capture the number of stop words used, number of punctuation used, number of dialogue breaks, etc

Themes- We tried to run the bag of words approach on the last 500, 100 words of the book to get the theme, the assumption was to get the emotions and key words from the conclusion of the book to get the Theme since in the end the dominate idea is concluded.

Hence we decided to build a model using only Bag of words and another one with meta data features and sentiment/emotion analysis.

We decided to represent the features first in bag of words(BOW) approach to establish a base line model for the comparison. Since there was a huge class imbalance problem to deal with, we decided to run it with cost sensitive classification models.The meta data feature were represented by each feature such as number of "!" present per sentence in a book and sentiment analysis and emotion analysis had a score for each emotion and sentiment.

We then combined all these features(of method 2 and 3) and put as a union feature set for classification models to run .

The advantage of the meta-data features and sentiment/emotion analysis is that at the end we shall have quantitative features to build a classification model upon.On the other hand, this can be a very large feature set in order to process if all the elements are considered for meta data features and capturing all emotions and its score.

However BOW shall have advantage of having key words for each genre which shall be more efficient as a classifier to identify the genre but we are loosing the order of the words.

IV. Implementation

For Meta data features extraction we primarily used the NLTK library.

For the emotion analysis on the book page we used the NRC emotion lexicon and for the sentiment analysis we used Vader sentiment analysis provided by the NLTK.

For bag of words we used TF-IDF from sklearn.[2]

For meta data feature extractions and sentiment analysis in nltk libraries we used lemmatization process after tokenizing and used pos tags.[3][4]

A. Class Imbalance Problem [6]

To address the class imbalance problem we chose to use imblearn to combine under sampling and oversampling methods.[5]

We started with the bag of words approach to establish a base line, hence with the given data first we tried models like Support Vector Machine(SVM), Logistic regression and Naive Bayes with the available data and captured the metric with accuracy, balanced accuracy, precision and recall.

Since the models were biased towards the majority class, we tried with the cost sensitive versions of the logistic regression and SVM where class weight parameter gives the balanced weights to all the classes with number of weights to proportion of instances per class.Unfortunately, there was not much improvement in the balanced accuracy/classification report.

As a result we decided to balance the data and then apply further feature extractions. We also tried to do the balancing after the feature extraction but due to highly imbalanced data we were struggling with the RAM issues and were not able to process all the data.

Hence decided to do the class imbalance before feature extraction.

Since the majority class had around 700 books we decided to do the under sampling on the majority class . WE first tried the bag of words methods and tried to measure the effectiveness of under sample methods such as RandomUnderSampling, Condensed Neasert Neighbour, Near Miss Undersampling,Tomek Links Undersampling, Edited Nearest Neighbors Rule (ENN), One-Sided Selection (OSS) and Neighborhood Cleaning Rule (NCR) from imblearn [5]library.

However when random under sampler was crashing because of memory error the other methods worked but did not reduce the number of books by much margin and hence the data was still imbalanced. Then we decided to randomly sample the books in majority class "Literary" for which we have considered the case where 1 author has written only 1 book, in order to keep it consistent with the other genre. Since we had seen it in EDA and it reduced the number of books in both majority classes to 40.

After under sampling there were classes where genre like allegories only had 2 books and few had only 5. Even with using cost sensitive classification algorithms like SVM and Logistic regression it was not yielding any good results hence we had a choice of over sampling. Even if we don't over sample the minority classes it was such a small data set that it was going to be over-fit. In case we over-sample those classes in which it just randomly copied in proportion to the other classes it again was leading to over fit, but we still chose to over sample those minority classes since it was making the model selection rather easier.

We were aware of the draw backs of under sampling that it was removing some of the useful information and oversampling was leading to over fitting.

For Oversampling we tried, random oversampling, SMOTE-NC and Adaptive Synthetic Sampling (ADASYN) , however it all produced the same results and copied the texts as it is.

B. Assumptions

While extracting meta data features, we considered the number of sentence and length of the book as the assumption was

the romantic books, allegories books are tends to be shorted in size rather than Literary books. Hence we counted the number of average sentences, words and average book length.

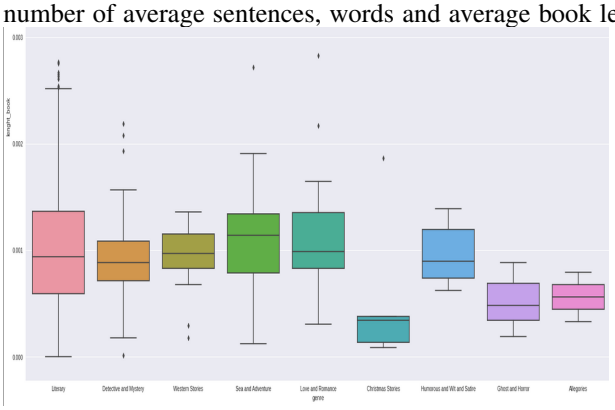

Fig: Average book length of all genres

To get the writing style we assumed that the number of common words such as ”is”, ”are”, ”am”,”I”, ”you”, ”spoke”, ”told”, ”yours” etc shall be more in romantic books rather than in allegories , since allegories has more poetic kind of style and indirect references hence we counted the number of stop words provided in nltk.

We also tried to get the number of characters both for plot settings and also other features, for this we assumed more number for proper nouns corresponds to more number of characters and hence the plot is complex.Literary books shall have more number of characters and romantic , horror books shall have few characters and allegories shall have even fewer. Similarly we considered number of digits, nouns, verbs and adjectives also to be distinguishable features.This could lead to the author’s writing style as well. Like if Charles Dikson likes to use more adjectives and some other author in another genre do not.

To dig deeper on author’s writing style we assumed that usage of foreign words lie ”gr8” , ” ersatz” etc would be peculiar to a book. Hence we counted number of foreign words as well.

Similarly we counted the number of period, punctuation, exclamation marks, colons etc. in order to identify the genre. As more period means more number of sentences and more exclamation means much more surprise or humor.

We also checked for dialogue breaks also assuming allegories with poems shall have fewer dialogue breaks and literary genre shall have more longer dialogues.

Similarly we assumed that number of masculine words such as ”he his man mr himself boy men gentleman gentlemen king prince son sir husband” etc shall be more in western stories and feminine word like ”she her woman herself girl women lady queen princess daughter madam madame wife” etc shall be more in romantic novels.

We also tried to capture the average number of unique words in a book for which we did not have a valid assumption.

Emotion Analysis [8]

We initially tried to run the emotion analysis on the dominated characters in the book however we were unable to implement it hence we tried to run it on the whole book.

We used NRC word lexicon to capture 8 available emotions- ’anger’, ’anticipation’, ’disgust’, ’fear’, ’joy’, ’negative’,’positive’, ’sadness’, ’surprise’, ’trust’. As in the list it also captures the 2 sentiments positive and negative.

It basically has list of words associated with each emotions and on the provided data each word in input data is compared and calculated against and thus for each emotion a score is calculated based on its presence.

Since books also can have range of emotions we thought to capture those to get better plot/context-setting , it also could be useful to capture in combination of proper nouns to see how each of the character is feeling.

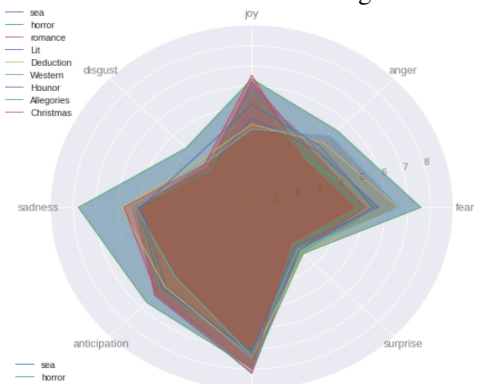

Fig: Radar Chart of 8 Emotions for all genres

Sentiment Analysis [9] [12]

We used Vader sentiment analysis to capture the overall sentiment of the book. We initially tried passing the first 100, last 100 and randomly 100 in middle words but that did not yield any good results when plotted on Radar charts. Hence we decided to use the completely book to capture the sentiments in hope to pass more information to it.

VADER (Valence Aware Dictionary and sentiment Reasoned) is a lexicon and rule based sentiment analyzer. It captures the presence of punctuation etc. and also capture how positive or negative is the score in compound score. Hence it seems like a good candidate to capture the sentiments.

However while analyzing for all the classes we assumed that for genre like Christmas stories/ romance positive sentiment shall be higher and for genre like horror negative emotions shall be higher. However while plotting there wasn’t much difference but the difference between positive sentiment in horror and Christmas was definitely higher. Also, it was all pointing towards the neutral sentiment more. Again the difference in neutral score among all genre was significant to distinguish.

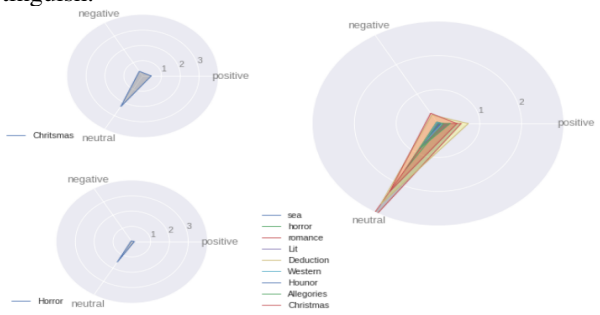

Fig: Radar Chart of 2 sentiments

If assumption made while implementing are wrong then we might not be able to learn the true distinction between all the classes and it may mis-lead the models to learn while trying to discriminate and leading to higher classification.

For bag of words we used TfidfVectorizer with unigrams only.

V. Evaluation

After extracting the features for the quantitative features we scaled the features using max-min scalar to standardise it and then ran the PCA on it to reduce the dimensional. On running 5-fold cross validation to get the number of component of PCA we found 15 as the best .

A. Model Selection

In order to select one final model we were primarily looking for the model which was giving better balanced accuracy(since after sampling we had balanced data), precision and recall. Since it's a multi class classification problem we were going through available predictive models.

We tried range of models such as SVM, Logistic Regression, SGD Classifier, KNN, DEcision Tree, GaussianNB and RandomForest using 10-fold cross validation and went with the models which were giving us the better balanced accuracy.

B. Train/Test/Validation split

We split the data into training, validation and testing with test set 25 , validation set 25 and training set 50 for second method , however for bag of words method we used 30 data for testing and without validation set.[6]

We transformed the test and validation data also as per training data.

For selecting the model however we used the training data only and using 10-fold cross validation we determined the model to be used as SVM and Logistic Regression out of 8 models.

C. Hyper Parameter Tuning

On the selected model SVM and logistic regression, we gridsearch the hyper parameters by using 10 fold cross validation again .

However while we plotted the learning and validation graphs for the same on hold out validation set it was clearly over fitting [5]. Hence we decided to change it using the validation set.

SVM learning curve after grid search and tested on validation set

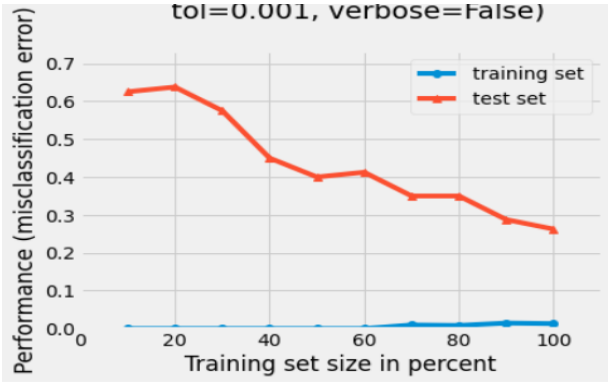

Fig: SVM Learning Curve on Validation set

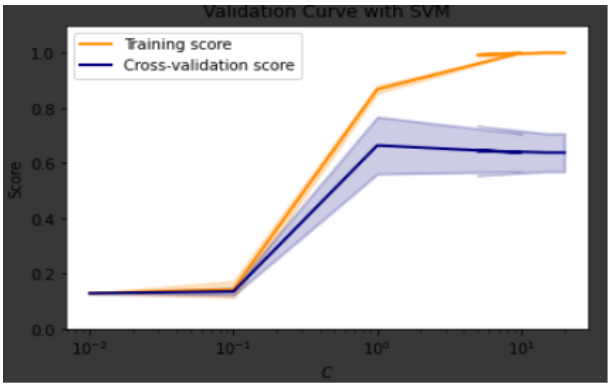

Fig: SVM Validation Curve for "C" on Validation set

D. Evaluation Metrics

To assess the chosen model we combined the training data and used the test data which was kept separately to train the model and test on completely unseen data.

Here after using the classification and plotting the learning curve [7] it was mostly under fitting.

We used accuracy, balanced accuracy , confusion metrics and the classification report to evaluate the model.

SVM after testing on independent test set after fixing overfit as per above-

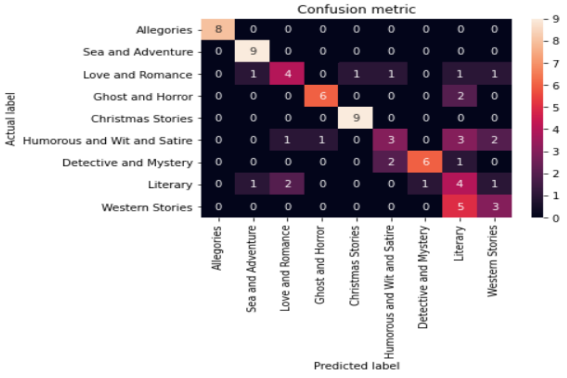

Fig: SVM confusion metric on independent test set

	precision	recall	f1-score	support
Allegories	1.00	1.00	1.00	8
Christmas Stories	0.82	1.00	0.90	9
Detective and Mystery	0.57	0.44	0.50	9
Ghost and Horror	0.86	0.75	0.80	8
Humorous and Wit and Satire	0.90	1.00	0.95	9
Literary	0.50	0.30	0.37	10
Love and Romance	0.86	0.67	0.75	9
Sea and Adventure	0.25	0.44	0.32	9
Western Stories	0.43	0.38	0.40	8
accuracy			0.66	79
macro avg	0.69	0.66	0.67	79
weighted avg	0.68	0.66	0.66	79

Fig: SVM classification report on independent test set

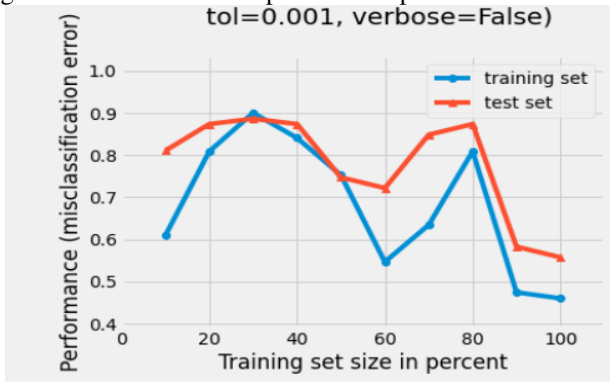

Fig: SVM learning curve on independent test set

	precision	recall	f1-score	support
Allegories	1.00	1.00	1.00	11
Christmas Stories	1.00	1.00	1.00	11
Detective and Mystery	0.00	0.00	0.00	10
Ghost and Horror	1.00	1.00	1.00	10
Humorous and Wit and Satire	1.00	1.00	1.00	10
Literary	0.36	1.00	0.53	12
Love and Romance	1.00	0.70	0.82	10
Sea and Adventure	1.00	0.36	0.53	11
Western Stories	1.00	0.90	0.95	10
accuracy			0.78	95
macro avg	0.82	0.77	0.76	95
weighted avg	0.81	0.78	0.76	95

Fig: BOW-SVM classification report on independent test set

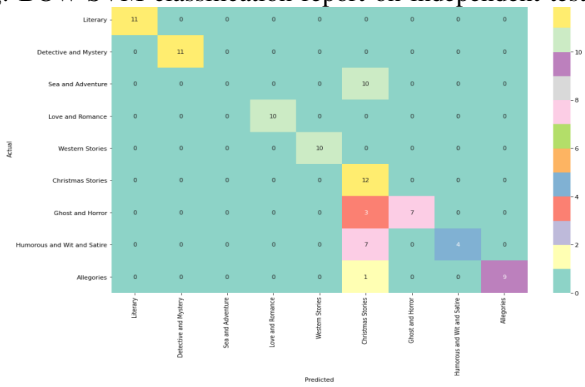

Fig: BOW-SVM confusion metric on independent test set

E. Results

As per the classification report and confusion metric, the results were weird. Compared to BOW methods and using Meta data and text sentiment/emotion analysis features, the accuracy of the model is little better though.

Logistic regression and SVM both with weights in bag of word model did very good with around 80 accuracy however as per classification report one could see that it clearly over fit and performed poor on literary, Sea and Adventure and Western Stories.

On the other hand the meta data features and sentiment and emotion analysis combined did better on these classes , however as per the graph we can see that it clearly over fit for the classes where we over sampled.Overall it shows the accuracy of 66 only which seems lower than the BOW method however clearly in the classes which were under sampled only and not over-sampled it worked better than the BOW approach.

Since all the features were quantitative for second methods the performance was faster than the BOW model.

VI. Conclusion

None of the methods alone were sufficient to make a better predictive model, and it looked like the combination of both the methods would have been a better choice.

More than extracting the relevant features for fiction like, context etc. and handling the huge data file, the class imbalance problem posed a great issue in handling and fitting the models for better performance.

We would have tried number of things, such as collecting more data from minority classes . Focusing on one particular element of the genre and combined methods of Bag of words and context analysis to capture one elements better .

We could run text summarising on the books to capture the plot etc for better results. Also better assumptions while implementing with the experts input would have been helpful in achieving better insults.

If we had the the complete books available may be running sentiment analysis on the conclusion or in the beginning of the book would have added more weight.

On this note one can say that without the class imbalance problem the second method would have worked faster and better anyways.as we earlier thought that having key words of each genre should result in a better model , this second quantitative methods proved it wrong.

We later decided not to pursue to identify the relation between author name and book name to genre , however we can utilise it in future to explore more.

The detailed implementation with the data and comments is available on-https://github.com/BlackCurrantDS/ATiML-Project

References

[1] https://gist.github.com/sebleier/554280
[2] https://scikit-learn.org
[3] https://en.wikipedia.org/wiki/Genre
[4] https://web.csulb.edu/ yamadaty/EleFic.html
[5] http://i.giwebb.com/wp-content/papercite-data/pdf/webbconilione04.pdf
[6] https://imbalanced-learn.readthedocs.io/en/stable/api.html
[7] http://rasbt.github.io/mlxtend
[8] https://saifmohammad.com/WebPages/NRC-Emotion- Lexicon.htm
[9] https://www.nltk.org/howto/sentiment.html

[10] https://github.com/sayantanpolley/fiction/blob/master/SIMFIC$_L NCS.pdf https$: $//github.com/sayantanpolley/fiction/blob/master/ICHMS2020_paper_42.pdf$
[11] https://www.aclweb.org/anthology/C18-1167
[12] https://pypi.org/project/vader-sentiment/