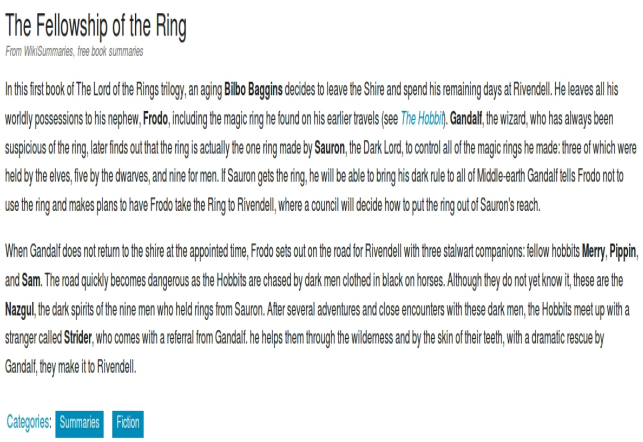
What is this Book's Genre?

Hakan AKYUREK [1] **Sefa YURTSEVEN** [1]

Abstract

In this paper we discuss various methods for predicting a book's genre from it's summary. We experiment with CMU Book Summary dataset with 16000 book summaries along with their respective authors and multiple genres. Right now we evaluate our results considering if predicted genre is one of book's labels, but we aim to evaluate our results with multi-label evaluation metrics such as Hamming Scoring.

1. Introduction



The Fellowship of the Ring

From WikiSummaries, free book summaries

In this first book of The Lord of the Rings trilogy, an aging **Bilbo Baggins** decides to leave the Shire and spend his remaining days at Rivendell. He leaves all his worldly possessions to his nephew, **Frodo**, including the magic ring he found on his earlier travels (see *The Hobbit*). **Gandalf**, the wizard, who has always been suspicious of the ring, later finds out that the ring is actually the one ring made by **Sauron**, the Dark Lord, to control all of the magic rings he made: three of which were held by the elves, five by the dwarves, and nine for men. If Sauron gets the ring, he will be able to bring his dark rule to all of Middle-earth Gandalf tells Frodo not to use the ring and makes plans to have Frodo take the Ring to Rivendell, where a council will decide how to put the ring out of Sauron's reach.

When Gandalf does not return to the shire at the appointed time, Frodo sets out on the road for Rivendell with three stalwart companions: fellow hobbits **Merry**, **Pippin**, and **Sam**. The road quickly becomes dangerous as the Hobbits are chased by dark men clothed in black on horses. Although they do not yet know it, these are the **Nazgul**, the dark spirits of the nine men who held rings from Sauron. After several adventures and close encounters with these dark men, the Hobbits meet up with a stranger called **Strider**, who comes with a referral from Gandalf. he helps them through the wilderness and by the skin of their teeth, with a dramatic rescue by Gandalf, they make it to Rivendell.

Categories: Summaries Fiction

Finding book genres from their summaries is a different topic that challenges us with some of the common NLP and Machine Learning problems. The problem in question is quite challenging and as much as interesting as it requires working with a highly imbalanced multi-labelled dataset, which requires data balancing techniques and multi-label algorithms, which we couldn't find a chance to work with before.

Book genre prediction is a rather unique problem, while being a variation of a common text classification problem. It is important in cases such as: library documentation or book store database construction. Recently, many researches worked on the problem of both imbalanced datasets and multi-label text classification or multi-label classification in general. So, few algorithms have been devised for multi-label classification. Common approaches include Bayesian approach and binary approach.

In this study we came up with our own approach to multi-label classification and analysed it and are going to work with other approaches in literature about the problem in the future. We use vectorised text documents to feed our ANN and other models. The vectorised documents go through some preprocessing before being fed to a model.

It is important to note that, we started this study considering it will be a multi-class classification problem. Accordingly, we researched methods about multi-class text classification problems, so the initial results are results of our own approach.

Main contributions can be considered as such:

Classification of multi-label data in a single label and it's evaluation.

Analysis of classification of multi-label data with a few models is literature.

The rest of the paper is as follows. We briefly review the related studies in Sec. 2. In Sec. 3 we describe our approach in detail. Results of experiments are discussed in Sec. 4. We discuss our conclusions in Sec. 5.

2. Related Works

This study (Imbalanced dataset with NB, 2003) discusses handling skewed data while working with Multinomial-Naive-Bayes(MNB for text classification, 2004). But Naive-Bayes is used for comparing results by us, so we mainly studied text classification using neural networks(ANN for text classification, 2003)(ANN for text classification, 2015).

According to imbalanced dataset problem, one can over-sample(Resampling), under-sample(Resampling). Under-sampling has various methods like NearMiss, Random under sampling, Tomek link removal, and they are briefly discussed in this paper(Resampling to improve performance,

[*]Equal contribution [1]Hacettepe University, Department of Computer Engineering. Correspondence to: <>.

BBM406 Fundamentals of Machine Learning, Fall 2018.

2016). Oversampling methods' performance are discussed in this report(Oversampling Comparison, 2017).

We also researched about multi-label text classification in literature. This paper(Predicting Movie Genres, 2018) discusses predicting movie genres from their plot summaries using various methods and analyses their performance. Of course there is still data imbalance with our dataset(Dataset, 2013)(Multi-label imbalance, 2015) and some researches are worked on this problem.

3. Our Approach

The aim of the study is to analyse various methods' and approaches' performance on classifying book genres using their summaries. Accordingly, the goal is to predict book genres as accurate as possible, again, using only their summaries. Obviously, this problem is no different than a standard text classification problem. However, working on a imbalanced dataset is what makes it different.

We believe that evaluating our results in a different manner would benefit us with more correct information about the things we do. Since many books belong to multiple genres, classifying a book to only one genre is somewhat wrong, because a book often belongs to multiple genres. We come up with two kinds of evaluations about this:

> Pick the class with highest probability and search it if it is one of books genres.
>
> Add a threshold to the outputs and pick classes with probability above that threshold instead of picking the highest one.

In this study we use a dataset[7] created from wikipedia book database. The dataset contains 16000 book information, which include publication date, author, summary, genres, name. Most books have multiple, unequal number of genres. From the dataset one can extract around 200 different book genre types. But number of books assigned to each genre is really uneven. For example a genre type has only 1 book assigned to it while another has 4000. To avoid the problems this imbalance causes we filtered our dataset to have only the top 27 genres and their books, thus leaving us with around 12000 books. In this filtered form our dataset has at least 100 books assigned to each book.

We chose ANN as our main method in the project. Simply because working with neural networks beats many other approaches in terms of performance and it is an interest of ours. If we were to compare it with a common method like NB, we could say that NN can handle word sequences and relationship between them much better than NB. We also learned that NN will perform better at relatively huger dataset. Although learning process for NN's are longer, they make it up with their better accuracy, since our aim is to get better accuracy we can say that NN's are a better choice for us.

Before our data goes into our model we do some preprocessing: removing stopwords, punctuations, lowering all characters, stemming or lemmatizing. After our data is prepared we first vectorize each input in bag of words manner, since we need to feed numerical data to our ANN model. We also use tfidf techniques to get more distinct input data. Then, we feed our input matrix to our NN. We are using keras' sequential neural network model with 'relu' activation functions. We are currently in the state of tuning hyper-parameters , but we got 63% accuracy so far with our model.

In our ANN model we used categorical crossentropy as loss function.

4. Experimental Results

We run some experiments on Naive Bayes. The accuracy scores are represented in the table below.

Table 1. Naive Bayes Accuracies with different methods

EXPERIMENT	ACCURACY
BASE	61.5%
TFIDF	62.7%
EQUAL CLASS PRIORS	63.8%
EQUAL CLASS PRIORS-TFIDF	63.7%
REMOVE STOPWORDS-TFIDF	61.8%
STEMMED DATA	60.2%
LEMMATIZED DATA	60.0%
SMOOTHING ALPHA=0.03	56.7%
1-3 GRAM	56.5%

Looking at the experiments, we can can quickly analyse a few key points. Firstly, equal class priors increase the accuracy most. From this we can easily say that our dataset is imbalanced without even checking the dataset manually. Interestingly, common preprocessing techniques of textual data do not increase and even decrease the accuracy score. We can also see that playing with some hyper-parameters in our Vectorizer and Naive Bayes classifier does not help at all and makes the model overfit the training data. Playing around with these **most accuracy we got is 64.1%**. Currently we haven't applied any re-sampling techniques on our dataset, and we aim to try a few before switching to multi-label classification.

Unlike Naive Bayes stemming and lemmatizing increase model's performance, as they should. When we look to the activation functions sigmoid performs better than relu in terms of performance, it seems sigmoid handles text classification better than relu. With neuron count increasing

in each layer model's performance seems to increase a little bit. But with a single layer of 512 neurons model directly overfits the training data, reducing accuracy to 47%. Other than that adding a third hidden layer again made model to overfit the training data, maybe adding extra layers with less neurons can help ANN perform better than MNB. **The best accuracy we got at this point is 62.8%**.

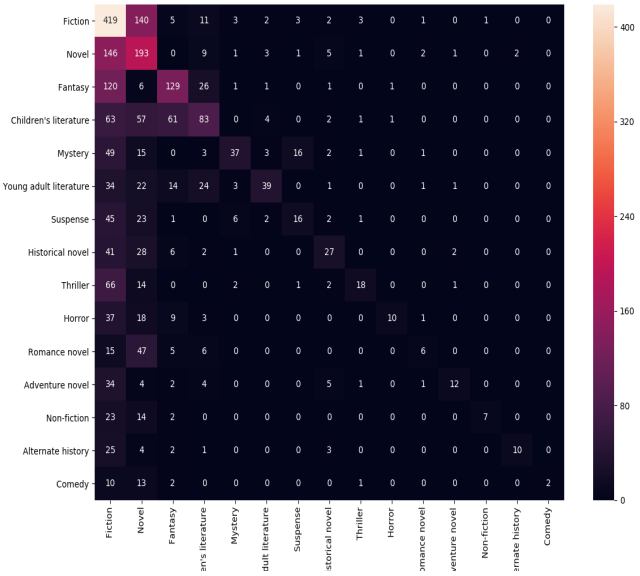
Table 2. ANN Accuracies with different methods

EXPERIMENT	ACCURACY
BASE-SIG-SIG(128N, 128N)	59.2%
BASE-RELU-RELU(128N, 128N)	55.4%
BASE-SIG-SIG(256N, 256N)	62.8%
BASE-RELU-RELU(256N, 256N)	56.9%
BASE-SIG-SIG-SIG(128EACH)	55.5%
BASE-RELU(512N)	47.1%
STEMMED DATA	61.5%
LEMMATIZED DATA	60.7%
STEMMED AND LEMMATIZED DATA	61.3%

5. Conclusions

Before switching to multi-label classification for sure, we approached this classification problem from a different angle based on our knowledge. We approached this as a multi-class classification problem. However, we aim to work on multi-label classification from here on out.

As we mentioned this is a multi-label classification problem in origin. We came to a conclusion about that after studying the confusion matrix below.



From looking to this matrix, we can see that there is something seriously wrong with the top most 4 classes. It can be seen that around half of the documents are misclassified in these classes. After looking to dataset more closely we noticed that some of the misclassified documents actually also belong to the classes our model classified. This was the trigger that made us research more on this topic and learn about multi-label classification.

We run many experiments on both Naive Bayes and ANN models. Both seemed to perform similar at best. We believe that it is because we chose to work with ANN's instead of RNN's or CNN's and our ANN model needs more hyper-parameter tuning. Highest accuracies of both models are represented in the table below.

Table 3. Best classification accuracies for Naive Bayes and ANN.

METHOD	BEST ACCURACY
NAIVE BAYES	63.8%
ANN	62.8%

Nevertheless, this isn't what we hoped for. We though artificial neural could easily perform batter than Naive Bayes. Maybe ANN is not the correct NN for text classification, we read RNN and CNN perform really well with text classification. Switching to them doesn't seem that illogical at this point.

References

Jason D. M. Rennie, Lawrence Shih, Jaime Teevan, David R. Karger Tackling the Poor Assumptions of Naive Bayes Text Classifiers, 2003 http://people.csail.mit.edu/jrennie/papers/icml03-nb.pdf

Rodrigo Fernandes de Mello, Luciano Jose Senger, Laurence Tianruo Yang AUTOMATIC TEXT CLASSIFICATION USING AN ARTIFICIAL NEURAL NETWORK, 2003 https://link.springer.com/content/pdf/10.1007%2F0-387-24049-7_12.pdf

Ashraf M. Kibriya, Eibe Frank, Bernhard Pfahringer, and Geoffrey Holmes Multinomial Naive Bayes for Text Categorization Revisited, 2004 http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.473.5982&rep=rep1&type=pdf

Oversampling and undersampling in data analysis https://en.wikipedia.org/wiki/Oversampling_and_undersampling_in_data_analysis

Ajinkya More Survey of resampling techniques for improving classification performance in unbalanced dataset, 2016 https://arxiv.org/pdf/1608.06048.pdf

Samrat Jayanta Dattagupta A PERFORMANCE COMPARISON OF OVERSAMPLING METHODS FOR DATA GENERATION IN IMBALANCED LEARNING TASKS, 2017 https://run.unl.pt/bitstream/10362/31307/1/TEGI0396.pdf

David Bamman Book dataset, 2013 http://www.cs.cmu.edu/~dbamman/booksummaries.html

Quan Hoang Predicting Movie Genres Based on Plot Summaries, 2018 https://www.researchgate.net/publication/322517980_Predicting_Movie_Genres_Based_on_Plot_Summaries

Francisco Charte, Antonio J. Rivera, Mara J. del Jesus, Francisco Herrera Addressing imbalance in multilabel classification: Measures and random resampling algorithms, 2015 https://sci2s.ugr.es/sites/default/files/ficherosPublicaciones/1790_2015-Neuro-Charte-MultiLabel_Imbalanced.pdf

Fraser Murray Text Classification using Artificial Neural Networks, 2015 https://minerva.leeds.ac.uk/bbcswebdav/orgs/SCH_Computing/FYProj/reports/1415/MURRAY.pdf