International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 07 | July 2020 www.irjet.net p-ISSN: 2395-0072

Author Identification and Sentiment Analysis for novels using Natural Language Processing

Pratiksha Karpe[1], Abhishek Agarwal[2], Rakhi Marathe[3], Prachi Kolekar[4], Prof. Namrata Wasatkar[5]

[1-4]Student, Computer Engineering, VIIT, Pune, Maharashtra, India
[5]Assistant Professor, Dept. of Computer Engineering, VIIT, Pune, Maharashtra, India

---

Abstract - *Author identification of novels is performed using Natural Language Processing (NLP) which is used to find the author of the novel or book by checking the text given by user in such way that it will check its language, vocabulary, use of different words according to their tone, etc. The given input text is much smaller than the previous ones which were used initially to determine the author. Different classifiers like SVM, Naïve Bayes are used to perform these actions. Sentiment analysis of given input text is performed for knowing genre. Different methods like tokenization, filtering, stemming are performed for further results. This project is useful for the people who read Novel related articles online and wish to know the details of authors.*

Key Words: **Natural Language Processing, Feature Selection, Author Identification, Sentiments, SVM, Naive Bayes classifier, Thayer's model, Deep Learning**

1. INTRODUCTION

Many novels are been written, but among them some acquire cult status over the years and are remembered for ages. The novels are of several genres and cross genres The cross genre is a mixture of several genres. Every author has their own style of writing which includes their signature fashion of using certain words, making their literature unique and recognizable. We will use this fact to identify the author from text snippets or quotes drawn from their novels. Machine learning powered by Natural Language Processing (NLP) is an excellent solution to the problem.

Author identification is the task of identifying the person who wrote a given piece of text from a given set of candidate authors. The importance of author identification lies on its wide applications. It can be used to find the original author of widely re-printed novel articles. It provides a new way to recommend the authors who have a high similarity of writing style with anonymous writers to a reader. The given piece of text will be pre-processed using NLP techniques.

Sentiment analysis is widely used in day to day life. For using any product, review of that product is done. In the same manner, in novels, it is checked whether that novel belongs to any different genre. From this project, advice is given to end user in such a way that user will read different novels according to their interests. By using Natural Language Processing, it will be easier to understand user's interests by analyzing their taste in books. Sentiment analysis is done in such a way that the input will be classified according to its genre.

2. RELATED WORK

We can find the author from sentences by many ways. In sentiment analysis constructing lexical chains, machine learning and many more are very useful approaches for the purpose. Others could be statistical approaches, domain knowledge driven analysis. Such approaches proved to be very advantageous in the task of sentiment analysis.

Work has been accomplished by many researchers in Automated Analysis of Bangla Poetry for Classification and Poet Identification[3] the various techniques Geetanjali Rakshit, Anupam Ghosh, Pushpak Bhattacharyya and Gholamreza Haffari had used are Shallow Parser, SVM classifier, Alliteration and Reduplication, Rhyme Scheme Detector, Document Statistics, Naive Bayes Classifier. Only words are not always sufficient for classifying poems into categories, because of poets often resorting to symbolism. It would be fascinating to additional investigate if this downside may well be helped with Word Sense Disambiguation (WSD). They were able to determine the poet correctly 92.3% of the the time using the SVM classifier.

Author Identification using Deep Learning[4] in this Ahmed M. Mohsen, Nagwa M. El-Makky and Nagia Ghanemhad used different methodologies like Deep learning, Auto Encoder, Denoising AutoEncoder (DAE),

Stacked Denoising Auto Encoder (SDAE)-Unsupervised pre-training and Supervised fine-tuning. They said the performance using the features extracted by the SDAE has been compared with the performance of the state-of-the art author identification techniques using the same corpus. The proposed system was able to outperform their results in terms of classification accuracy using a 10-fold cross validation settings. The system was able to reach classification accuracy up to 95.12%.

Comparing Frequency and Style-Based features for Twitter Author Identification[1] in this Rachel M. Green and JohnW. Sheppard used minimalistic text for search operation, Twitter dataset having threshold value for any text, SVM for classification purpose, Use of Bag-of-Words and style marker for extraction.To know the author of any text which is present on Twitter, style marker is effective method. This will require Threshold value to minimize size of the text to get meaning from them. This experiment uses 12.5 words (81.2 characters) as a average text length of tweets and classification accuracies on these datasets were consistently far above random chance, even with small set sizes. The results of these experiments counsel that author identification via style marker feature sets could also be more practical than ancient strategies of linguistics or word analysis for brief text.

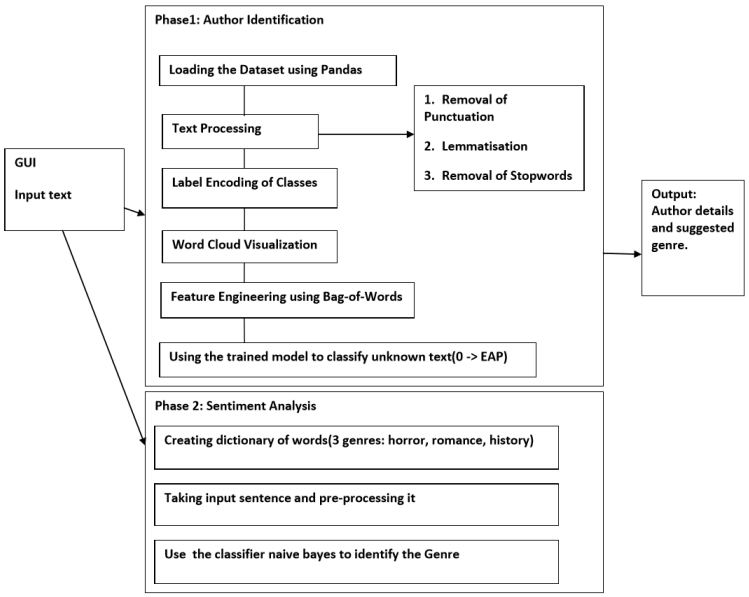
Author Identification on Literature in Different Languages: A Systematic Survey[2] In this Dr. Rajesh S. Prasad, Kale Sunil Digamber rao did Author identification on different languages such as English, Japanese, Mongolian, Persian, Indian, Albanian, Brazilian, etc. The Feature selection is main important task for identifying writing style of author. This depends on language which is chosen and size of text also. If size of text of author is large, then it will be easy for identification while for small data or text, it is challenging work to perform. They observed there is necessity for development of data set with respect to Languages. Most of the work is on English language very few work on other languages so there is scope for research on other Languages.

Two-Dimensional Sentiment Analysis of Text[6] In this Rahul Tejwani has discussed about Thayer's model for getting emotions from the given input text. Firstly, polarities are checked and then after intensity of input text is checked. This model uses lexicon features for training purpose. Here, predefined sets are used for checking which are generic. As dataset is not created based on the training dataset, it can be used to any domain. This model gives results about 81.60% which can be compared with current techniques or methods.

Sentiment Content Analysis and Knowledge Extraction from News Article[7] in this Mohammad Kamel, Neda Keyvani, Hadi Sadoghi-Yazi have suggested the use of sentiment analysis for news articles from different countries. It differentiates between different categories from each countries news articles and shows the results. Natural Language Techniques and some fusion methods are used to extract sentiment from the news to make rankings of all the countries according to their news interest. The news data is reduced to some content for getting hottest news topics using instance selection algorithms. This will be useful for different analysts, sociologists to get knowledge from politic, entertainment, sports, technology, business conditions.

3. PROPOSED WORK

In this system there are two phases which are working together to get final output. In first phase, user will give a text input then NLP will be performed on the text and author will be identified for the given text and details of the author will be visible to the users and in the second same input will be used to perform sentimental analysis and genre of the input text will be identified. The architecture is given below:



In Phase1: Author identification

Step1: To load the required datasets using the library called Pandas.

Step2: Applying Natural Language Processing for preprocessing:

- Punctuation Removal:-Punctuation marks shall and will be removed from all the sentences from the datasets.
- Lemmatisation of words:-Multiple forms of a word are called as lemma. For example, (running, ran) are inflected forms or lemma of the word run which is the root word.
- Removal of Stopwords:- Stop-words also called as vague words that are not useful in forming the sentences are removed in this step.

Step3: Label Encoding of Classes
For our dataset the number of classes shall be equal to the number of unique authors. But the name of authors i.e. labels are/maybe non-numeric. These shall be label encoded to make them numeric, that starts from 0 depicting each label in an order (shall be alphabetically encoded)

Step 4: Word Cloud Visualization
Each and every author has their own unique way or style of writing i.e. some of the words are frequently used by them and some are not. Hence by using word cloud visualization a chart of most used words and that of least used words can be formed, helping us in understanding the authors' pattern of writing.

Step 5: Feature Engineering using Bag-of-Words
The algorithms of Machine Learning work only on numeric data. But in the dataset that we shall use contain sentences that is present in the form of text only. Hence there is a need to convert this textual data into numeric form. One such approach of doing this, is Feature Engineering. Using this approach, the numeric features are extracted/engineered or found from textual data. A number of Feature Engineering Techniques exist that we can use. However, we shall use Bag-of-Words Technique of Feature Engineering in our approach.

Step 6: Training our Model
Any classifier can be used as the Classification Machine Learning Algorithm. Multiple classifiers needs to be tested and the best among them must be selected.

Step 7: Once the model is trained the training accuracy and validation accuracy must be found. (This step is useful in finalizing our classifier i.e. the one that gives maximum accuracy).

Step 8: This trained model shall be used to classify unknown text i.e. finding the author of random text. Bag of words shall also be applied on this random text and then shall be tested.

In Phase2: Sentiment analysis

Step 1: Creating individual vocabulary

Step 2: Creating dictionary of words
A dictionary is a collection which is unordered, changeable and indexed and turning words into respective features

Step 3: Creating training data set and training a classifier
Naive Bayes: The Naive Bayes classifiers are the probability based classifiers which are based on the Bayes theorem to get strong independence assumptions between features in machine learning technique.

train_set = horror_features + romantic_features + historic_features

They are among the simplest Bayesian network models. Taking input sentence and processing it .
Processing steps: Tokenization, remove Punctuation and non-alphabetic tokens, filter out stop words, stemming

Step 4: Using the classifier maintain the count of words in the particular sentence.

classifier = NaiveBayesClassifier.train(train_set)

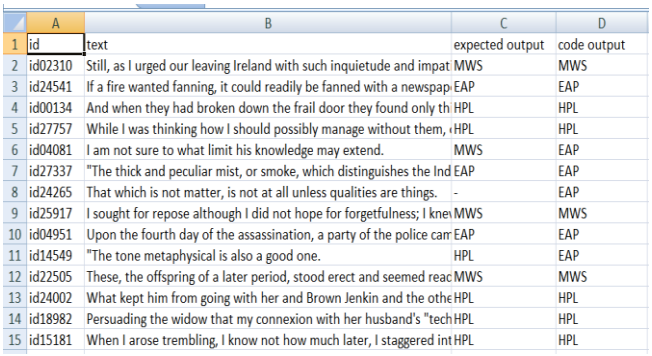
Step 5: Use the greater count to identify the Genre.

4. RESULTS AND DISCUSSION

The implemented model gives the result for the phase1 which is shown in the below picture that some of the sentences are getting their very own author. So in these we taken an observation table of 50 sentences and from them 34 are showing correct result. The accuracy score is 68%.

Next for the phase 2 if we are giving more number of vocabulary words to the dictionary it will give you more accurate genre. In this we have given 65-70 words for each genre i.e. horror, romance, history and

for 50 sentences of each genre it is identifying the correct genre for upto 35 sentences.



	A	B	C	D
1	id	text	expected output	code output
2	id02310	Still, as I urged our leaving Ireland with such inquietude and impat	MWS	MWS
3	id24541	If a fire wanted fanning, it could readily be fanned with a newspap	EAP	EAP
4	id00134	And when they had broken down the frail door they found only th	HPL	HPL
5	id27757	While I was thinking how I should possibly manage without them,	HPL	HPL
6	id04081	I am not sure to what limit his knowledge may extend.	MWS	EAP
7	id27337	"The thick and peculiar mist, or smoke, which distinguishes the Ind	EAP	EAP
8	id24265	That which is not matter, is not at all unless qualities are things.	-	EAP
9	id25917	I sought for repose although I did not hope for forgetfulness; I kne	MWS	MWS
10	id04951	Upon the fourth day of the assassination, a party of the police car	EAP	EAP
11	id14549	"The tone metaphysical is also a good one.	HPL	EAP
12	id22505	These, the offspring of a later period, stood erect and seemed reac	MWS	MWS
13	id24002	What kept him from going with her and Brown Jenkin and the othe	HPL	HPL
14	id18982	Persuading the widow that my connexion with her husband's "tech	HPL	HPL
15	id15181	When I arose trembling, I know not how much later, I staggered int	HPL	HPL

So again the accuracy for the phase2 model is 70% which is implemented over the random sentences to identify its genre.

5. CONCLUSIONS

There have been various tasks done on the topic authorization and sentiment analysis too. But these both techniques never applied together.

The proposed model gives both in one model to identify the author of particular sentence and to identify its genre also using sentiment analysis. The model used naïve bayes classifier to classify features in sentiment analysis where in author identification we have used wordcloud visualization.

This will definitely help people who frequently read novel related articles online and wishes to identify the author of the novel and by that way it will improve the reading culture.

6. FUTURE SCOPE

In future we can make the system work in a distributed architecture manner. To train the system in order to identify more authors. To make the system identify authors of novels written in languages other than English (Hindi/Marathi). As well as can provide e-book links to user of the searched novel.

REFERENCES

[1] Rachel M. Green, John W. Sheppard, "Comparing Frequency and Style-Based Features for Twitter Author Identification" [2013]

[2] Dr. Rajesh S. Prasad, Kale Sunil Digamber rao, "Author Identification on Literature in Different Languages: A Systematic Survey" [2018]

[3] Geetanjali Rakshit, Anupam Ghosh, Pushpak Bhattacharyya, Gholamreza Haffari, "Automated Analysis of Bangla Poetry for Classification and Poet Identification" [2015]

[4] Ahmed M. Mohsen, Nagwa M. El-Makky, Nagia Ghanem, "Author Identification using Deep Learning" [2016]

[5] Tan Thongtan, Tanasanee Phienthrakul, "Sentiment Classification using Document Embeddings trained with Cosine Similarity" [2016].

[6] Rahul Tejwani, "Two dimensional Sentiment Analysis of text" [2014].

[7] Mohammad Kamel, Neda Keyvani, Hadi Sadoghi-Yazi, "Sentimental Content Analysis and Knowledge Extraction from News Articles" [2018].

[8] Navoneel Chakrabarty, Machine Learning Approach to "Author Identification of Horror Novels from Text Snippets" [2019].