Using Machine Learning for Text Analysis: Author Profiling and Classification for Better Identification of Age, Gender and Author Identity

Team Members:

19MAI0033 BHASWATI SENGUPTA

19MAIOOO3 INDRANIL RAY

19MAI0031 SOMTIRTHA MUKHERJEE

Report submitted for the
THREE Project Review of

Course Code: CSE6019 –Text, Web and Social Media Analytic

Slot: G1/TG1/G2/TG2

Professor: Dr. Jagalingam Pushparaj

Deadline: 30th June 2020

ABSTRACT:

Text analysis is one of the most important applications of machine learning. Author profiling is a sub domain of text analysis which deals with a bunch of classification problem related to the author of the text. This includes gender and age classification and identity classification of textual authors as well. The analysis can be done using several machine learning algorithms. In this paper, we have deployed many such techniques to deploy the classification. We distribute the work of this paper into three main stages. In the first stage, we start by pre-processing the data. Next, feature vectors were created and on the final stage, we classified the text dataset by several machine learning algorithms. We have demonstrated K-Nearest Neighbour method, Logistic Regression, Logistic Regression with Cross Validation, ensemble learning, specifically the Random Forrest method, Support Vector Machine and Decision Trees. A well-documented comparison report is also provided with this paper.

KEYWORDS: *Author Profiling, KNN, SVM, Decision trees, Logistic regression, Random Forrest, Gender and Age Classification from texts, Machine Learning, Text Analysis.*

I. INTRODUCTION:

Though our world consists of 8.7 million species of animals, only one among them is capable of expressing their feelings by conversational methods clearly and properly – us, the human beings. The evolution from cavemen to the modern-day homo sapiens has crafted a natural habit of clarified voice conversation and even written texts. The first examples of literature are believed to be found in the ancient Mesopotamia, around 3400 BC, known as cuneiform. Of course, the literature and linguistic domain itself has gone through tremendous evolutionary process and has cultivated many languages and literatures as well, in a global aspect. Modern day civilisation has come very far not only in terms of literature, butalsoand importantly so in terms of art, science, andtechnology, impacting uponour lifestyle immensely. Whether the society that we live in are better than that of our forefathers, is up for debate, but, the thing most certainly out of the debate is the amount of technical advancements we have made over the centuries and ages, making our lives easier, is an exaggerating figure in itself. Today, the modern 21st century people are still drooling over the idea of making our lives even for advanced, and thus, calling upon technical analytics into the picture more and more.Since the advent of machine learning, machines are not mere assistants, they now do possess learning and decision-making ability, most often, used for analytics.This is century is called and rightly so that century of big data as people all over the world generate huge chunks of data every second, with most contextually appropriate ones being those generated in the form of texts.Natural language processing has given the machine learning enthusiasts the proper wings to explore about linguistic analysis domain. In thispaper, we mostly focus on the aspect of a three-fold author profiling problem. Authors can produce different type of texts to express their feeling.The real challenge is to identify the author from the text. This certainly has a conventional application domain in forensics, sales, security and also many other fields as well. The text contains keywords and other features, to identify and classify them into categories will help identify the author. It is also possible to determine the gender and the age group as well from text datasets. This computational task of predicting the demographics related to the author is accomplished with the help of several machine learning algorithm. In this paper, we will be demonstrating six different methods for the classification purpose and will briefly discuss about them too. The comparison table is provided at the end of this paper to clearly indicate the most efficient method to deploy in real time, of course, taking certain conditions into account such as the type of text and dataset size. Nonetheless, the machine learning algorithms considered here were tested on the same dataset for getting an accurate comparison, contextually.

II. LITERATURE REVIEW:-

- In the paper **Celebrity Profiling on Twitter using Sociolinguistic Features[1]** published in the year **2019** by authors Luis Gabriel Moreno-Sandoval1,3 , Edwin Puertas2,1,3 , Flor Miriam Plaza-del-Arco4 , Alexandra Pomares-Quimbaya1,3 , Jorge Andres Alvarado-Valencia1,3 , and L. Alfonso Ureña-López4 described a strategy for characterizing the profile of celebrities on Twitter is presented based on the generation of socio-linguistic features from their posts which serve as inputs for a collection of classifiers. In particular, they created four classifiers defining the degree of fame, the gender, the date of birth and the possible occupation of a celebrity. They obtained the training and test data sets from PAN 2019 at CLEF. Future work includes analyzing the models with real samples with a similar or greater volume of messages and reviewing the posts and background data to provide models that socially respond to the variables representing actual network phenomena.

- **A Straightforward Multimodal Approach for Author Profiling[2]** by Mario Ezra Aragón1 and A. Pastor López-Monroy published in **August 2018** identified the gender of different users by using tweets and images posted. For representation of the text the following strategies were evaluated: 1) Bag of Terms (BoT), 2) Second Order Attributes (SOA) representation, 3) Convolutional Neural Network (CNN) models and 4) an Ensemble of n-grams at word and character level was done.For representation of images they used a Convolutional Neural Network (CNN). It was observed that the n-grams Ensemble presented the highest

performance.

- In another paper **Author Profiling based on Text and Images Notebook for PAN at CLEF[3]** in the year 2018 by Luka Stout, Robert Musters, and Chris Pool. They defined the writers gender based on written text and shared images. Their approaches to the text-based, image-based, and combined task were identified. The existence of three different languages raises the question of whether a single model architecture which works well in all three languages can be developed. They also suggested a way of integrating multiple predictions about shared content into a single user-level prediction. The final text structure is an ensemble of a script from Naive Bayes, and an RNN. The image classification is achieved by using CNNs to find the selfies and predict the person's gender on those pictures. Future work includes:-tuning each model (textual and visual) separately as well as tuning the value of alpha to find the best combined model & developing a new combined model performing multi-modal fusion using the textual and visual modalities.Also applying various deep neural models can be done as improvement in the model.

- In the research work **Author Profiling and Aggressiveness Detection in Spanish Tweets: MEX-A3T[4]** in the year 2018 by Mario Ezra Aragon´ 1 and A. Pastor Lopez-Monroy ´ performed different evaluation strategies on Spanish classification tasks that have been successfully for English classification tasks. The strategies evaluated are : 1) a Bag of Terms (BoT), 2) a Second Order Attributes (SOA) representation, 3) a Convolutional Neural Network (CNN) models and 4) a huge ensemble of n-grams at word and character level. Datasets used are from PAN2013-2017.The results obtained are in agreement with previous studies that showed high performance for words and n-grams and represent a very first attempt to bring these and other ideas to the Spanish classification tasks, such as distributional representations and neural networks.Even though n-gram model performed better the future work involves improving the performance of SOA and CNN as these models did not perform well in this task.

- In **June,2017** in the paper **Author profiling: A machine learning approach towards detecting gender, age and native language of users in social media** by Elias Lundeqvist & Maria Svensson they compared two different approaches to author profiling, one in which features are extracted manually and the author's gender , age and native language are known from an unknown text using the learning algorithms Support Vector Machine and Feed-Forward Neural Networks and the other in which features are learned from raw data using Convolutional Neural Networks learning algorithms. Further works include using Pretrained word embedding vectors instead of training both the word embeddings and training the neural network simultaneously.The SVM features can be extended .

- In the **year 2016**,in the thesis **Predicting Author Traits Through Topic Modeling of Multilingual Social Media Text** Caitlin McCollister, 22 teams implemented software to predict the age, gender, and certain personality traits of Twitter users based on the content of their posts to the website. Training dataset of tweets in various languages like English, Spanish, Dutch, and Italian was given. They then trained a random forest classifier based on the labeled training dataset to predict the age, gender and personality traits for authors of tweets in the test set.

- The paper **Segmenting Target Audiences: Automatic Author Profiling Using Tweets** published in the year **2015** by Maite Giménez, Delia Irazú Hernández, and Ferran Pla describes a methodology proposed for author profiling using natural language processing and machine learning techniques. They used lexical information in the learning process. For those languages without lexicons, they automatically translated them, in order to be able to use the info This paper explores how to define user profiles using classic techniques of NLP. Corpora have been created compiling tweets in different languages.

- In the year **2014** the paper **Examining Multiple Features for Author Profiling** by Edson R. D. Weren, Anderson U. Kauer, Lucas Mizusaki, Viviane P. Moreira, J. Palazzo M. de Oliveira, Leandro K. Wives in their work,focused on discovering age and gender from blog authors. With this goal in mind, they analyzed a large number of features – ranging from Information Retrieval to Sentiment Analysis. This paper reports on the usefulness of these features. Experiments on a corpus of over 236K blogs show that a classifier using the features explored here have outperformed the state-of-the art. More importantly, the experiments show that the Information Retrieval features proposed in their work are the most discriminative and yield the best class predictions .

Comparison Table of past work:-

SL NO	TITLE	ADVANTAGE	DISADVANTAGE
1.	Celebrity Profiling on Twitter using Sociolinguistic Features	Characterizing the profile of celebrities on Twitter is presented based on the generation of socio-linguistic features from their posts which serve as inputs for a collection of classifiers.	The models are not analyzed with real samples with a similar or greater volume of messages .
2.	A Straightforward Multimodal Approach for Author Profiling	identified the gender of different users by using tweets and images posted	SOA & CNN strategy needs improvement.
3.	Author Profiling based on Text and Images Notebook for PAN at CLEF 2018	The writers gender based on written text and shared images.Image classification is done using CNN.	Model and alpha value not properly tuned.
4.	Author Profiling and Aggressiveness Detection in Spanish Tweets	Spanish text classification in done using various strategies and comparing them .Conclusion is n-gram model is better than the rest.	SOA & CNN strategy is poor.
5.	Author profiling: A machine learning approach towards detecting gender, age and native language of users in social media	Performance is done following two different methods One is features are extracted manually and the author's gender , age and native language are known from an unknown text using the learning algorithms Support Vector Machine and Feed-Forward Neural Networks and the other in which features are learned from raw data using Convolutional Neural Networks learning algorithms.	Did not use Pretrained word embedding vectors instead training of both the word embeddings and training the neural network simultaneously is done.
6.	Predicting Author Traits Through Topic Modeling of Multilingual Social Media Text	Implemented software to predict the age, gender, and certain personality traits of Twitter users based on the content of their posts to the website for english,Spanish,dutch and Italian language	There was some problem formatting the data in WEKA
7.	Segmenting Target Audiences: Automatic Author Profiling Using Tweets	This paper define user profiles using classic techniques of NLP (lexical information)	Gender identification results were poor.

III. METHOLOGY

The entire work is divided into 2 parts ,the first part is author identification where we used spooky author identification dataset and the second part is age and gender identification of the authors of a text here we used different hotel review , blogs and tweets as the dataset . Our work is divided into 4 parts :

1. Pre Processing of data
2. Creation of Feature Vectors
3. Developing a Machine learning Algorithm
4. Accuracy Testing

Pre-Processing of Dataset

The data needed to be cleaned before we feed this data to the machine learning models . Without proper data cleaning the machine learning wont be able to give accurate results as there would be bias . The data cleaning included few tasks which are as follows :

- Modifying URL's and numbers
- Removing HTML tags
- To decrease the biasness removing the duplicate tweets
- Removing non English charecters

Creation of Feature Vectors

The next step was creating the feature vectors for the data . The vectorization was done on general features .It is very important on the type of data we are working on and this feature vectorization will give us the basic idea of that . Few of the feature vectors of the dataset is as follows :

- Number of Characters per item
- Number of words per item
- The number of different parts of speech tag
- Number of words starting with capital letters
- Percentage of capital tokens per sentence
- Ending of a sentence with punctuation
- Percentage of punctuation in a sentence

Developing a Machine Learning Algorithm

We have used different libraries like NLTK and pyEnchant for feature extraction .Then used these features for using different machine learning algorithms . The machine learning algorithm was chosen based on the size of the input . We have used the following machine learning algorithms in our project :

1. K-Nearest Neighbours
2. Naïve Bayes
3. Logistic Regression
4. Support Vector Machines
5. Decision Tree

6. Ensemble (using - KNN , Logistic Regression and SVM)

Accuracy Testing
After deploying the machine learning models the accuracy of each model was noted and compared with different algorithms . The detailed comparison is done in the result section .

a. DATASET USED:-

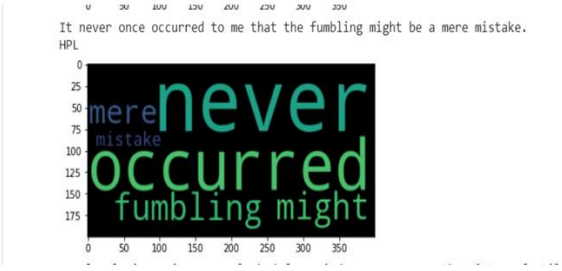
We have used two datasets in this work . The 1st dataset is used for author identification and the dataset is spooky author identification . This dataset was available in GitHub . The dataset comprised of 3 columns which are , id , text and author . The 2nd dataset is a collection of various hotel reviews , blogs and tweets . This dataset is used for age and gender identification .
In previous papers no work was done on author identification given a text of some writing style . Then for age and gender identification only one of the machine learning algorithms had been used previously . But in this paper different machine learning algorithms were used and the results were compared with each other ,in order to find the best machine learning algorithm for this scenario .

b. TOOLS:-

- **Softwares Used :** Jupyter Notebook , Google Colab
- **Language Used :** Python 3.7
- **Machine Learning Tools :** Sckit Learn ,Numpy , Pandas , matplotlib etc.

Word Relation Tools : NLTK pyEnchant

c. SCREENSHOTS:-




```
model = model.fit(text_bow_train, y_train)

[ ] model.score(text_bow_train, y_train)

    0.9074889867841409

[ ] model.score(text_bow_test, y_test)

    0.8414198161389173
```

Author Identifcation

Feature Vectorization

```
[1, 12, 17, 18, 3, 4, 5, 6, 7, 8, 9, 10, 11, 16, 14, 19, 21]
['FEMALE' 'MALE']
['18-24' '23-50' '50-XX']
(5452,)
(5452,)
(5452, 17)
```

```
5 nearest neighbor for age:
Accuracy: 0.45646196150320806
5 nearest neighbor for gender:
Accuracy: 0.5664527956003667
Logistic Regression for gender:
Accuracy: 0.5747021081576535
```

Accuracy for K NN and Logistic Regression

```
In [18]: words_to_fet()

Busy but a good quality hotel. Would stay again.
0
Great Location, Great room We just returned from three nights at The London NYC. We had a very nice suite and our only complain
t was that the sofa was not comfortable. The bedding was wonderful, the bathroom was large and beautiful. We had a tub with a s
hower though most rooms do not. Our room had no closed closet but had a nice separate open dressing area with enough storage sp
ace. Service was attentive from checking in to departure.Ordering coffee from room service is very expensive but there are lots
of places near the hotel including Starbucks and Au Bon Pain. The location is wonderful for walking to Central Park, Fifth Aven
ue shopping, and theater. We had a summer weekend rate which was very reasonable.
5000
```

Word To Feature

IV. RESULTS:-

The different models on which the datasets were tested has been compared in the following table .Different models gave different accuracies and it can understood that all models are not suitable for this dataset .

Machine Learning Algorithm	Accuracy of Age	Accuracy of Gender
K -Nearest Neighbour	0.45	0.56
Logistic Regression	0.51	0.57
Logistic Regression with Cross Validation	0.51	0.58
SVM	0.52	0.56
Decision Tree	0.46	0.54
Ensemble	0.60	0.57

As it can be seen from the above table that ensemble learning works best for this dataset .
The accuracy for author identification using spooky author identification is 0.85 . Here we used Naïve Bayes Classifier as the machine learning algorithm .

V. CONCLUSION:

This paper clearly reflects the fact that Ensemble models produced the best results in terms of age classification while almost every algorithm more or less produced similar results in case of gender classification. We believe, the situation might be more clarified with the help of more training data for better training purpose of the models. However, a significantly better performance was observed in case of Author identification where Naïve Bayes method were implemented. The future work of this paper should be directed towards the proper evaluation of the algorithms on different dataset and more importantly real time analysis.

VI. REFERENCES:-

[1] Luis Gabriel Moreno-Sandoval1,3 , Edwin Puertas2,1,3 , Flor Miriam Plaza-del-Arco4 , Alexandra Pomares-Quimbaya1,3 , Jorge Andres Alvarado-Valencia1,3 , and L. Alfonso Ureña-López4 "Celebrity Profiling on Twitter using Sociolinguistic Features" "*CLEF 2019, 9-12 September 2019, Lugano, Switzerland."*
[2] Mario Ezra Aragón1 and A. Pastor López-Monroy "A Straightforward Multimodal Approach for Author Profiling", *Notebook for PAN at CLEF 2018*
[3] Luka Stout, Robert Musters, and Chris Pool Anchormen, The Netherlands "Author Profiling based on Text and Images *Notebook for PAN at CLEF 2018*".
[4] Mario Ezra Aragon´ 1 and A. Pastor Lopez-Monroy ´ "Author Profiling and Aggressiveness Detection in Spanish Tweets: MEX-A3T 2018" "*Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) 135*"
[5] V. N. Vapnik, "An overview of statistical learning theory," IEEE Transactions on Neural Networks, vol. 10, pp. 988–999, Sep 1999.
[6] A. Sharma, "Handwritten digit recognition using support vector machine," CoRR, vol. abs/1203.3847, 2012.
[7] J. ying Wang, "Application of support vector machines in bioinformatics," Brief Bioinform., vol. 5, pp. 328–338, 2004.
[8] K. bo Duan and S. S. Keerthi, "Which is the best multiclass svm method? an empirical study," in Proceedings of the Sixth International Workshop on Multiple Classifier Systems, pp. 278–285, 2005.
[9] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in Proceedings of the Fifth Annual Workshop on Computational Learning Theory, COLT '92, (New York, NY, USA), pp. 144–152, ACM, 1992. [10
] C. Clapham and J. Nicholson, The concise Oxford dictionary of mathematics. Oxford University Press, 2009.