International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org, editorijettcs@gmail.com
Volume 8, Issue 3, May - June 2019 ISSN 2278-6856

Authorship Identification: Naïve Bayes with XGBoost Approach

Dr. B.S. Daga[1], Jason Dsouza[2], Ryan Furtado[3], Manupendra Tiwari[4]

[1]Associate Professor, Department of Computer Engineering, Fr. Conceicao Rodrigues College of Engineering, Fr. Agnel Ashram, Bandstand, Bandra (West), Mumbai, India - 400050.

[2]Student, Department of Computer Engineering, Fr. Conceicao Rodrigues College of Engineering, Fr. Agnel Ashram, Bandstand, Bandra (West), Mumbai, India - 400050.

[3]Student, Department of Computer Engineering, Fr. Conceicao Rodrigues College of Engineering, Fr. Agnel Ashram, Bandstand, Bandra (West), Mumbai, India - 400050.

[4]Student, Department of Computer Engineering, Fr. Conceicao Rodrigues College of Engineering, Fr. Agnel Ashram, Bandstand, Bandra (West), Mumbai, India - 400050.

Abstract: ***In today's world, electronic text is used for communication on a large scale. Most of this content is provided anonymously or under unverified names. For forensic applications, it is important to segregate text into groups of text that may be written by the same individual under a different alias. There are many copyright dispute cases, where multiple people claim the ownership of some content. Authorship identification along with mathematical or statistical analysis of texts could be the key to solve this problem. When an individual writes, they subconsciously use a certain array of words or writing patterns and sentiments, and we could use this to determine their writing style. The fundamental assumption of authorship identification is that each individual has a habit of subconsciously using certain words, patterns and emotions that make their writing style unique. Extraction of these individual features from text could be used to distinguish one author from another. The problem statement for our system is as follows: Building a system that can be trained to recognize a certain individual based on his writing style i.e. the set of words (features) used frequently by the individual. This is also known as generating a writeprint (similar to a fingerprint). With the help of this writeprint the system will be able to identify any other documents or texts which have been written by the same individual. This should help reduce plagiarism in case of authors and can also be used in forensics to identify criminals based on their writing.***

Keywords: Authorship Identification, Handwriting Analysis, Plagiarism Detection, Writeprint, Feature Extraction.

1. INTRODUCTION

Nowadays, text is used for every basic form of electronic communication. Many times it is also used for malpractices like copyright issues, plagiarism, terrorist communication, cyber harassment, etc. In most of these situations, the texts are sent anonymously. Thus, it becomes important to be able to identify the actual author of such malicious texts to counter these malpractices. Authorship Identification could provide a simple solution for the same.

There has been previous research in this field. Sadia Afroz [1] has studied deception in authorship attribution and derived features for the same. Mubin Tamboli and Rajesh Prasad [2] have studied the various techniques for authorship identification. Yitao Li [3] and Efstathios Stamatatos [4] have studied the different machine learning approaches for authorship identification and have compared the results for various algorithms. Siddharth Swain, Gaurav Mishra and C. Sindhu [6] have studied recent experiments and compared between their results, based on the approaches they used. [7]- [23] include various researches in the fields of online text attribution, nature recognition from text, and text analysis techniques.

This paper studies the need of Authorship Identification, and thereafter, proposes an architecture for it. It focuses on using N-gram with certain features and feeding them into a Naïve Bayes classifier along with XGBoost for classification.

The paper also tests the accuracy for various sub approaches like tfidf, word count and their combination with XGBoost to determine the best sub approach. Finally, a dataset is generated and the system is coded in Python using Anaconda3 and Jupyter Notebook along with WordCloud for display.

2. ARCHITECTURE AND WORKING

2.1 Proposed System Architecture

After studying the previous works in the field of Authorship Identification [1], [2], [3], and studying the different components involved in Authorship Identification, we have broken down the entire process of Authorship Identification into four parts: Document Collection, Feature Extraction, Classifier Construction, and Authorship Identification [2].

Document Collection refers to collection of input in the form of text from various documents and sources and generation of dataset.

Feature Extraction is the process of identifying various stylometric features. It also uses N-gram up to 4-grams.

This is followed by the generation of a Writeprint for testing and classification.
Classifier Construction is the comparison and classification of the set. Using the input and database of feature sets, the classification is done with the help of Naïve Bayes classifier along with XGBoost.
Authorship Identification uses the Writeprint obtained in Classifier Construction to validate the author and generate the final result.

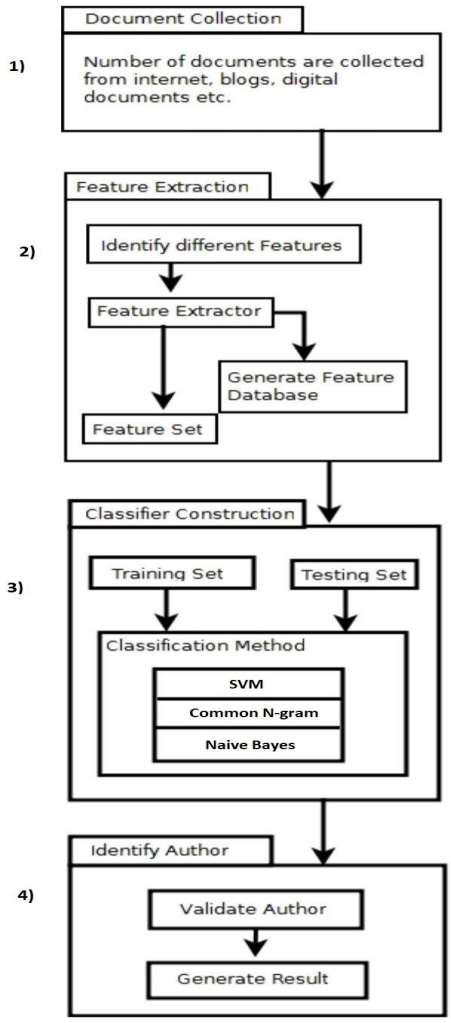

Figure 1 Proposed System Architecture

2.2 Proposed Tool for Feature Extraction

Natural Language Tool Kit (NLTK) is a platform for building Python programs to execute on, with data as human language. It provides simple standard interfaces to over 50 text databases, patterns, semantic and lexical resources like WordNet, along with a collection of text processing libraries for preprocessing, parsing, classification, stemming, tagging, tokenization, and semantic reasoning,.
Using the libraries found in NLTK the system first "learns" the style of known candidate authors based on documents of those authors, and the style of a given set of anonymous documents by extracting up to 8 meta-features and selecting the top 3 most used meta-features. It then estimates the ownership of the anonymous documents to one of the known authors or individuals.

2.3 Proposed Algorithm for Classifier

After studying the different approaches for Classifier Construction like Support Vector Machines (SVM), K Nearest Neighbors (KNN), Random Forest in previous researches [3], [4] and comparing the results of previously built systems [6], we selected Naïve Bayes algorithm for classification.
Naive Bayes classifier has the advantage of being simple and efficient in both the training and testing phases. Also unlike the SVM classifier and the CNG classifier, the Naive Bayes classifier does not depend on any model with tuneable parameters, hence does not require any search in parameter space. Another advantage of the Naive Bayes classifier is that its output can be probability values, while the outputs of the SVM, KNN, and the random forest classifiers can only be discrete class labels.

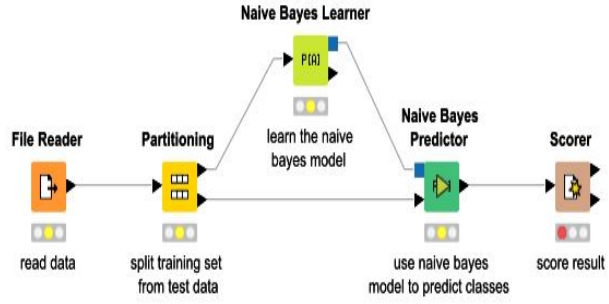

Figure 2 Naïve Bayes Classifier

2.4 Working

The libraries used for all the natural language processing are imported from the 'Natural Language Tool Kit' i.e. NLTK. The data set used is in the form of an excel sheet (.csv le) where each sentence of the book is stored in a row associated with author name and line ID.
The first step involves pre-processing the collection of sentences, which involves tokenization, stemming and lemmatizing the training data set From this the frequency of each word used is calculated followed by determining the occurrence (in which context) of each word using N-gram technique. Eight key meta-features are then extracted from the pre-processed text. Out of these the top 3 features are chosen for vectorization.
The text is then made to undergo vectorization, which is then fed into the Naive Bayes classifier. The unknown author's text is then imported from the testing dataset and the probabilities for each author is calculated using XGBoost Model. The output is displayed in the form of a statement along a WordCloud representation to represent the most frequently used words by that particular author.
The entire setup has been on the Anaconda IDE used for Python programming, along with Jupyter Notebook, which allows programmers to execute code step by step in order to debug with greater ease.

3. TESTING AND RESULTS

3.1 Testing various Sub Approaches

3.1.1 Naïve Bayes on Tfidf Vectorizer Loss with Confusion Matrix

TFIDF is short for 'term frequency–inverse document frequency'. It is a computational statistic that intends to put forth how significant a word is to a text sample in a collection or database. It is generally used as a weighting or determining factor in retrieval of information and data mining. The TFIDF value grows in proportion to the frequency or occurrence of the word in the text sample, and is offset by the number of text samples in the database that contain that particular word. This establishes the claim that some words appear more frequently than others. TFIDF is one of the most popular schemes for estimating the weightage of a term. A majority of the text-based systems today use TFIDF.

Based on the confusion matrix obtained, we calculated an accuracy of approximately **66.87%.**

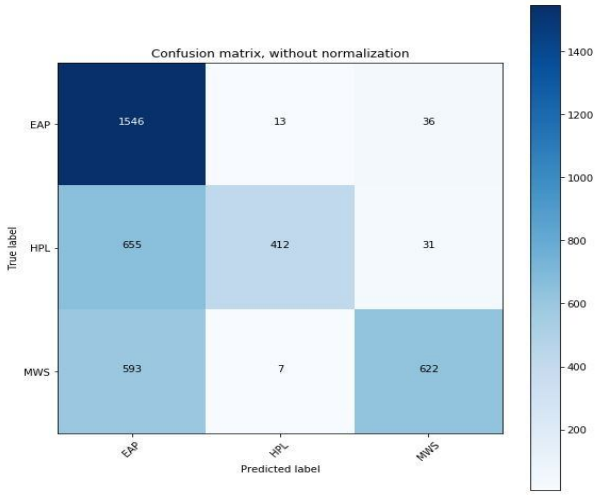

Figure 3 Confusion Matrix for Naïve Bayes on Tfidf Vectorizer

3.1.2 Naïve Bayes on Word Count Loss with Confusion Matrix

The word count is the number of words in a document or passage of text.

Based on the confusion matrix obtained, we calculated an accuracy of approximately **86.79%.**

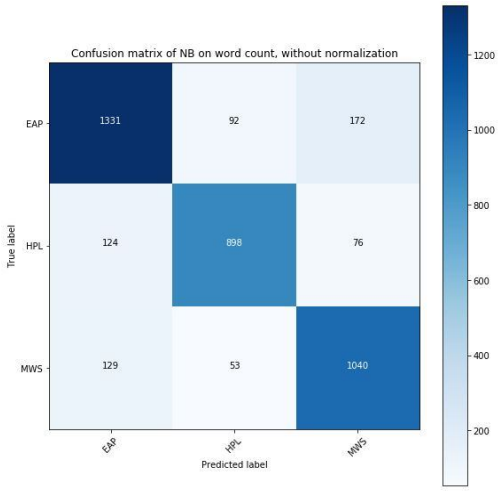

Figure 4 Confusion Matrix for Naïve Bayes on Word Count

3.1.3 Naïve Bayes on Combination of Word Count and XGBoost

XGBoost is a highly efficient, flexible and portable gradient boosting library that can be used to boost the performance of a learning algorithm. It is highly optimized and distributed. It makes use of a Gradient Boosting framework to bolster machine learning algorithms. XGBoost provides boosting known as GBDT or GBM, which operates in a parallel tree fashion. This solves any machine learning problem in a quick and efficient manner. The exact same code can run on various different distributed environments like Hadoop, SGE, MPI, etc.; and can solve any given problem.

Based on the confusion matrix obtained, we calculated an accuracy of approximately **86.95%.**

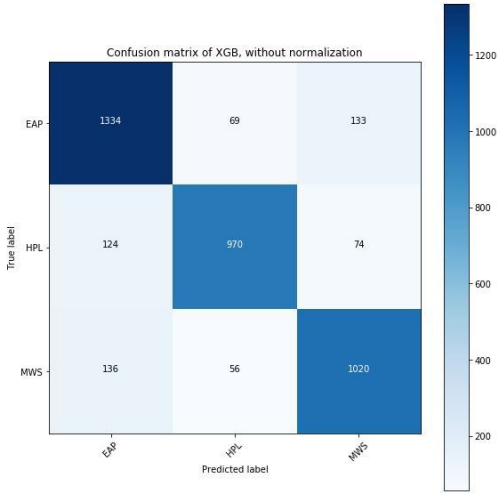

Figure 5 Confusion Matrix for Naïve Bayes on Combination of Word Count and XGBoost

Thus, Naïve Bayes on Combination of Word Count and XGBoost gives the best possible result with an accuracy of **86.95%.**

3.2 Training and Testing Data Set

The dataset contains text from works of fiction written by famous authors of the public domain such as Edgar Allan Poe (EAP), H. P. Lovecraft (HPL) and Mary Wollstonecraft Shelley (MWS).

The training dataset contains an ID, the text sample and its corresponding author. The testing dataset contains only an ID and the text sample.

id	text	author
id26305	This process, however, afforded me no means of ascertaining the dimensions of my dungeon; as I might make its circuit, and return to the poi	EAP
id17569	It never once occurred to me that the fumbling might be a mere mistake.	HPL
id11008	In his left hand was a gold snuff box, from which, as he capered down the hill, cutting all manner of fantastic steps, he took snuff incessantly v	EAP
id27763	How lovely is spring As we looked from Windsor Terrace on the sixteen fertile counties spread beneath, speckled by happy cottages and weal	MWS
id12958	Finding nothing else, not even gold, the Superintendent abandoned his attempts; but a perplexed look occasionally steals over his countenan	HPL
id22965	A youth passed in solitude, my best years spent under your gentle and feminine fosterage, has so refined the groundwork of my character tha	MWS
id09674	The astronomer, perhaps, at this point, took refuge in the suggestion of non luminosity; and here analogy was suddenly let fall.	EAP
id13515	The surcingle hung in ribands from my body.	EAP
id19322	I knew that you could not say to yourself 'stereotomy' without being brought to think of atomies, and thus of the theories of Epicurus; and sin	EAP
id00912	I confess that neither the structure of languages, nor the code of governments, nor the politics of various states possessed attractions for me.	MWS
id16737	He shall find that I can feel my injuries; he shall learn to dread my revenge" A few days after he arrived.	MWS
id16607	Here we barricaded ourselves, and, for the present were secure.	EAP
id19764	Herbert West needed fresh bodies because his life work was the reanimation of the dead.	HPL
id18886	The farm like grounds extended back very deeply up the hill, almost to Wheaton Street.	HPL
id17189	But a glance will show the fallacy of this idea.	EAP
id12799	He had escaped me, and I must commence a destructive and almost endless journey across the mountainous ices of the ocean, amidst cold th	MWS
id08441	To these speeches they gave, of course, their own interpretation; fancying, no doubt, that at all events I should come into possession of vast c	EAP
id13117	Her native sprightliness needed no undue excitement, and her placid heart reposed contented on my love, the well being of her children, and	MWS
id14862	I even went so far as to speak of a slightly hectic cough with which, at one time, I had been troubled of a chronic rheumatism of a twinge of he	EAP
id20836	His facial aspect, too, was remarkable for its maturity; for though he shared his mother's and grandfather's chinlessness, his firm and precocio	HPL
id11411	Now the net work was not permanently fastened to the hoop, but attached by a series of running loops or nooses.	EAP
id08075	It was not that the sounds were hideous, for they were not; but that they held vibrations suggesting nothing on this globe of earth, and that a	HPL
id18925	On every hand was a wilderness of balconies, of verandas, of minarets, of shrines, and fantastically carved oriels.	EAP
id19925	With how deep a spirit of wonder and perplexity was I wont to regard him from our remote pew in the gallery, as, with step solemn and slow,	EAP
id01704	These bizarre attempts at explanation were followed by others equally bizarre.	EAP
id10125	For many prodigies and signs had taken place, and far and wide, over sea and land, the black wings of the Pestilence were spread abroad.	EAP
id02448	All that as yet can fairly be said to be known is, that 'Pure gold can be made at will, and very readily from lead in connection with certain other	EAP
id23451	I seemed to be upon the verge of comprehension without power to comprehend men, at times, find themselves upon the brink of remembra	EAP
id27907	Our compasses, depth gauges, and other delicate instruments were ruined; so that henceforth our only reckoning would be guesswork, basec	HPL
id08121	This the young warriors took back with them to Sarnath as a symbol of conquest over the old gods and beings of Ib, and a sign of leadership in	HPL
id15222	Meantime the whole Paradise of Arnheim bursts upon the view.	EAP
id00764	I was rich and young, and had a guardian appointed for me; and all about me would act as if I were one of their great society, while I must keej	MWS
id00683	We could make out little by the dim light, but they seemed to contain prophecies, detailed relations of events but lately passed; names, now	MWS
id11733	Even now They talked in Their tombs.	HPL
id03205	Sheehan especially did they ply with inquiries, yet without eliciting any information of value concerning Old Bugs.	HPL
id01948	He cried aloud once, and a little later gave a gasp that was more terrible than a cry.	HPL
id22412	The old tracks crossed River Street at grade, and at once veered off into a region increasingly rural and with less and less of Innsmouth's abhor	HPL

Figure 6 Training Data Set

id	text
id02310	Still, as I urged our leaving Ireland with such inquietude and impatience, my father thought it best to yield.
id24541	If a fire wanted fanning, it could readily be fanned with a newspaper, and as the government grew weaker, I have no doubt that leather and iron acquired durability in proportion, for, in a very sh
id00134	And when they had broken down the frail door they found only this: two cleanly picked human skeletons on the earthen floor, and a number of singular beetles crawling in the shadowy corners.
id27757	While I was thinking how I should possibly manage without them, one actually tumbled out of my head, and, rolling down the steep side of the steeple, lodged in the rain gutter which ran along t
id04081	I am not sure to what limit his knowledge may extend.
id27337	"The thick and peculiar mist, or smoke, which distinguishes the Indian Summer, and which now hung heavily over all objects, served, no doubt, to deepen the vague impressions which these obj
id24265	That which is not matter, is not at all unless qualities are things.
id25917	I sought for repose although I did not hope for forgetfulness; I knew I should be pursued by dreams, but did not dread the frightful one that I really had.
id04951	Upon the fourth day of the assassination, a party of the police came, very unexpectedly, into the house, and proceeded again to make rigorous investigation of the premises.
id14549	"The tone metaphysical is also a good one.
id22505	These, the offspring of a later period, stood erect and seemed ready to advance fearlessly into coming time; while those out worn stragglers, blasted and broke, clung to each other, their weak b
id24002	What kept him from going with her and Brown Jenkin and the other to the throne of Chaos where the thin flutes pipe mindlessly was the fact that he had seen the name "Azathoth" in the Necron
id18982	Persuading the widow that my connexion with her husband's "technical matters" was sufficient to entitle me to his manuscript, I bore the document away and began to read it on the London boa
id15181	When I arose trembling, I know not how much later, I staggered into the house and made shocking obeisances before the enshrined amulet of green jade.
id21888	And by the shores of the river Zaire there is neither quiet nor silence.
id12035	Idris heard of her mother's return with pleasure.
id17991	I say this proudly, but with tears in my eyes for the firm proved themselves the basest of ingrates.
id10707	But let us glance at the treatise Ah "Ability or inability to conceive," says Mr. Mill, very properly, "is in no case to be received as a criterion of axiomatic truth."
id07101	"What a place is this that you inhabit, my son" said he, looking mournfully at the barred windows and wretched appearance of the room.
id00345	At his nod I took one of the latter and seated myself upon an aged, discoloured gravestone close by the newly uncovered aperture.
id05912	No one doubted now that the mystery of this murder would be immediately brought to light.
id13443	But although, in one or two instances, arrests were made which promised elucidation, yet nothing was elicited which could implicate the parties suspected; and they were discharged forthwith.
id09248	Festivity, and even libertinism, became the order of the day.
id17542	For I am Iranon, who was a Prince in Aira."
id06995	"Gaze not on the star, dear, generous friend," I cried, "read not love in its trembling rays; look not upon distant worlds; speak not of the mere imagination of a sentiment.
id25159	I am serious in asserting that my breath was entirely gone.
id25729	The thing will haunt me, for who can say the extermination is complete, and that analogous phenomena do not exist all over the world?
id26949	Before each of the party lay a portion of a skull, which was used as a drinking cup.
id27191	If she had been bred in that sphere of life to which by inheritance the delicate framework of her mind and person was adapted, she would have been the object almost of adoration, for her virtu
id07668	Or, if this mode of speech offend you, let me say, that my mother, the proud queen, instilled early into me a love of distinction, and all that, if the weakness of my physical nature and my peculia
id02230	I have promised that someone should watch for him and give him instant notice if any new object should appear in sight.
id15553	The rope drags, either on land or sea, while the balloon is free; the latter, consequently, is always in advance, when any progress whatever is made: a comparison, therefore, by means of the con
id25688	In his frenzied strains I could almost see shadowy satyrs and Bacchanals dancing and whirling insanely through seething abysses of clouds and smoke and lightning.
id17545	The cat and pigeons seemed to suffer no inconvenience whatsoever.
id13929	Did he think of this as we journeyed up to town?
id12880	All other art objects I had ever seen either belonged to some known racial or national stream, or else were consciously modernistic defiances of every recognised stream.

Figure 7 Testing Data Set

3.3 Results

3.3.1 Word Cloud

Word Cloud is used to display most frequently used words by the authors.

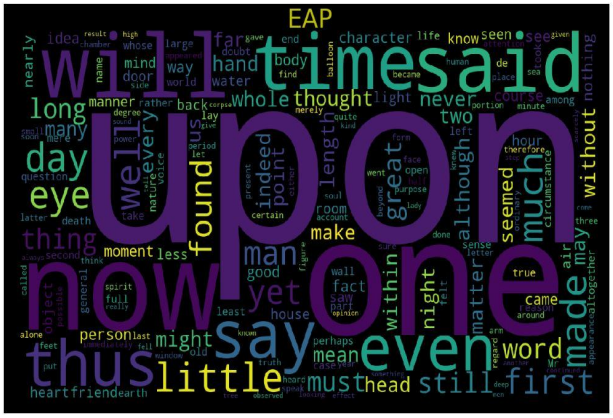

Figure 8 Word Cloud

The figure above shows the Word Cloud for author EAP. Similarly, the Word Clouds for the other two authors were generated.

3.3.2 N-Gram

We determine 4-Gram for each author. This gives the relation and context in which the word was used, by grouping it with other words it was used along with, and checking the frequency of this group up to a total of 4 words in a group.

	1-Gram	Occurence	2-Gram	Occurence	3-Gram	Occurence	4-Gram	Occurence
0	one	516	old man	59	terribl old man	9	eric moreland clapham lee	4
1	could	480	could see	31	heh heh heh	9	oonai citi lute danc	4
2	thing	433	one night	23	charl le sorcier	8	mad arab abdul alhazr	4
3	old	392	old woman	22	small pane window	8	necronomicon mad arab abdul	4
4	would	357	one might	19	great great great	6	aira citi marbl beryl	4
5	seem	317	one could	18	cap n obe	5	villag plain edg bog	3
6	like	296	shun hous	17	citi marbl beryl	5	sir eric moreland clapham	3
7	time	281	next day	17	mad arab abdul	4	heh heh heh heh	3
8	man	280	heh heh	17	oonai citi lute	4	nithra fall tini kra	3
9	night	265	new england	16	basalt pillar west	4	region sea meet sky	3
10	hous	257	one thing	16	arab abdul alhazr	4	great great great grandfath	3
11	saw	236	could tell	16	small furri thing	4	overlook sea gaili paint	2
12	even	234	look like	16	horn wane moon	4	sea gaili paint galley	2
13	though	234	year ago	15	necronomicon mad arab	4	gaili paint galley sail	2
14	came	216	one anoth	14	aira citi marbl	4	shun hous benefit street	2
15	look	213	tempest mountain	14	great old one	4	peak overlook sea gaili	2
16	strang	204	small hour	14	new church green	4	paint galley sail harbour	2
17	men	204	look upon	14	eric moreland clapham	4	galley sail harbour toward	2
18	dream	199	herbert west	14	thraa ilarnek kadatheron	4	strang olden ruin far	2
19	day	197	brown jenkin	13	moreland clapham lee	4	snowi peak overlook sea	2

Figure 9 N-Gram

The figure above shows the N-Gram for author HPL. Similarly, the N-Grams for the other two authors were generated.

3.3.3 Feature Graphs

This plots a graphs of each of the 8 Meta features for each author.

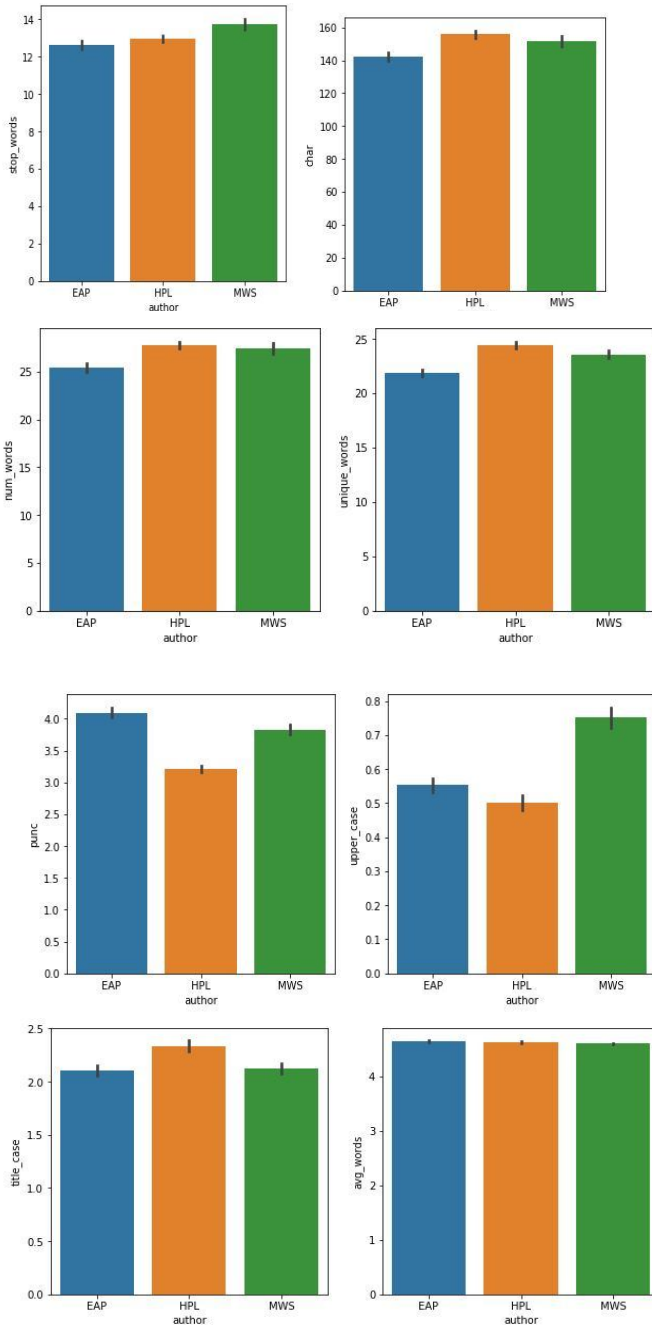

Figure 10 Feature Graphs

3.3.4 Top 3 Meta Features

As seen here, the following are the top 3 meta-features chosen for further processing: Number of characters used, Number of average words used, Number of unique words used.

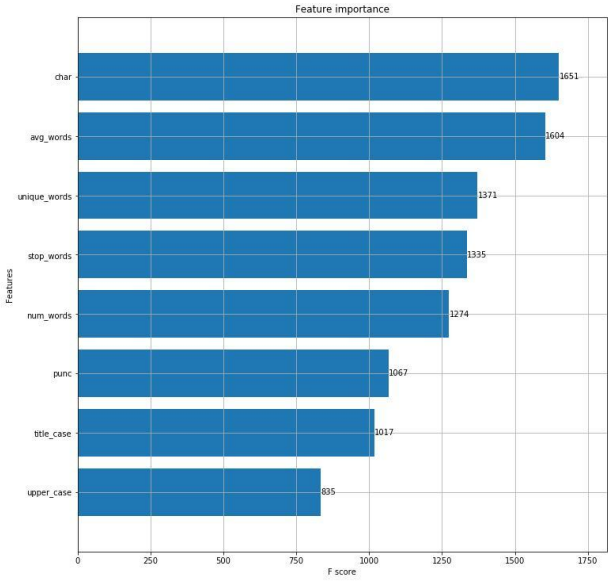

Figure 11 Top 3 Meta Features Graph

3.3.5 Sentimental Analysis

Sentimental Analysis provides information about the nature and mindset of an individual. The nature of the authors have been categorized into 3 types:

1. Positive
2. Negative
3. Neutral

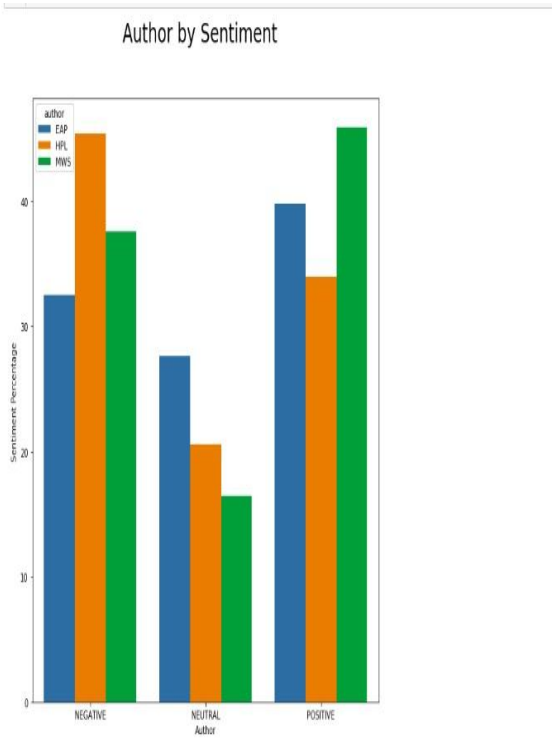

Figure 12 Sentimental Analysis

3.3.6 Final Output

The final output is the set of testing text samples with the probabilities for each author.

id	EAP	HPL	MWS
id02310	0.10884753	0.013279164	0.8778733
id24541	0.9990132	0.000470766	0.000516034
id00134	0.082437724	0.91595507	0.001607204
id27757	0.6347978	0.36032006	0.004882119
id04081	0.96809924	0.017208721	0.014692075
id27337	0.8680448	0.11736581	0.014589395
id24265	0.7265841	0.21585634	0.057559595
id25917	0.022266673	0.08215693	0.8955764
id04951	0.991418	0.007698841	0.000883175
id14549	0.7968022	0.08696985	0.116227865
id22505	0.12946755	0.07253701	0.79799545
id24002	0.000570433	0.99930394	0.000125665
id18982	0.18507627	0.70038754	0.11453622
id15181	0.001556115	0.99809307	0.000350759
id21888	0.657936	0.16411138	0.17795265
id12035	0.004578187	0.000721971	0.99469984
id17991	0.25157344	0.021241935	0.72718465
id10707	0.9995952	0.000251543	0.000153274
id07101	0.30747294	0.18672259	0.5058045
id00345	0.025031079	0.96373504	0.011233942
id05912	0.9233926	0.057256445	0.019350922
id13443	0.962705	0.020440133	0.016854823
id09248	0.32645735	0.09793721	0.57560545
id17542	0.03881715	0.9522454	0.008937417
id06995	0.021959744	0.000254763	0.9777855
id25159	0.7215763	0.02817608	0.25024766
id25729	0.346314	0.5744545	0.079231545
id26949	0.9199588	0.066146836	0.013894446
id27191	0.6026896	0.06730664	0.33000368
id07668	0.2227586	0.001355732	0.77588564
id02230	0.31790027	0.16124044	0.5208593
id15553	0.9985763	0.000695454	0.000728208
id25688	0.000912652	0.99857295	0.000514422
id17545	0.98383164	0.006970119	0.009198289
id13929	0.4942487	0.4084561	0.0972953
id12880	0.092972435	0.82534635	0.08168122
id19282	0.6411031	0.071913056	0.28698385

Figure 13 Final Output

4. CONCLUSION AND FUTURE SCOPE

A system for Authorship Identification was successfully created using Naïve Bayes with XGBoost approach. The system performs at an accuracy of 86.95%.

In the future, we plan to expand the current system to a system that can dynamically take inputs and create a dataset that can be used for training of the system. We plan to implement an OCR component within the system to enable the application of the system over pictures, PDFs, etc. We will improve the accuracy of the system and enhance its performance. We also plan to implement our system in real world applications like authorship identification, plagiarism detection, handwriting analysis, text matching, etc.

REFERENCES

[1.] Sadia Afroz, "Deception in Authorship Attribution", Drexel University, December 2013.

[2.] Mubin Shaukat Tamboli and Rajesh S. Prasad, "Authorship Analysis and Identification Techniques: A Review", International Journal of Computer Applications (0975 – 8887) Volume 77 – No.16, September 2013.

[3.] Yitao Li, "Application of Machine Learning Techniques to Paper-Author Identification Problem", Project Report for Course TCSS 702, Institute of Technology University of Washington, Tacoma, 2013.

[4.] Efstathios Stamatatos, "A Survey of Modern Authorship Attribution Methods", Department of Information and Communication Systems Engineering, University of the Aegean, 2008.

[5.] Richmond Hong Rui Tan, Flora S. Tsai, "Authorship Identification for Online Text", Nanyang Technological University, International Conference on Cyberworlds, 2010.

[6.] Siddharth Swain, Gaurav Mishra, C. Sindhu, "Recent Approaches on Authorship Attribution Techniques – An Overview", SRM University, International Conference on Electronics, Communication and Aerospace Technology, ICECA 2017.

[7.] Azah Kamilah Muda, Siti Mariyam Shamsuddin, Maslina Darus, "Mining Generalized Features for Writer Identification", 2009 2nd Conference on Data Mining and Optimization, 27-28 October 2009.

[8.] Jose Hurtado, Napat Taweewitchakreeya, Xingquan Zhu, "Who Wrote This Paper? Learning for Authorship De-identification Using Stylometric Features", Florida Atlantic University, IEEE IRI 2014, August 13-15, 2014.

[9.] Ioannis Kourtis, Efstathios Stamatatos, "Author Identification Using Semi-supervised Learning", Notebook for PAN at CLEF 2011, University of the Aegean.

[10.] B. Rama Krishna, J. Ramesh, "An Efficient Self Constructing Algorithm for Text Categorization" International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 7, 2012, ISSN: 2278-0181.

[11.] Na Cheng, R. Chandramouli, K.P. Subbalakshm, "Author gender identification from text", Eslevier Digital Investigation (2011), pp 78-88.

[12.] Abdur Rahman, Haroon A. Babri, Mehreen Saeed, "Feature Extraction Algorithms for Classification of Text Documents", ICCIT 2012, pp. 231 -236.

[13.] Daniel Pavelec, Edson Justino, Leonardo V. Batista, and Luiz S. Oliveira, "Author Identification using Writer-Dependent and Writer-Independent Strategies", SAC'08 March 16-20, 2008, ACM 978-1 -59593-753-7/08/0003, pp.414-418.

[14.] Abbasi, A. and Chen, H. "Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace" ACM Trans. Inf. Syst. 26, 2, Article 7 (March 2008), pp. 1 -29.

[15.] Rosen-Zvi, M., Chemudugunta, C., Griffiths, T., Smyth P., and Steyvers, M. "Learning author topic models from text corpora" ACM Trans. Inform. Syst. 28(1), Article 4 January 2010, pp. 1 -38.

[16.] Giacomo Inches, Fabio Crestani, "Online Conversation Mining for Author Characterization and Topic Identification" PIKM'11, October 2011, ACM978-1-4503-0953-0/11/10.

[17.] Farkhund Iqbal, HamadBinsalleeh, Benjamin C.M. Fung, Mourad Debbabi, "A unified data mining solution for authorship analysis in anonymous textual communications", Elseveir Pub., Information Sciences, 231 (2013) pp. 98–112.

[18.] Jacques Savoy, "Authorship attribution based on a probabilistic topic model," Information Processing and Management 49 (2013) Elsevier Pub. pp. 341 –354.

[19.] ShlomoArgamon, Marin Sari, Sterling S. Stein, "Style Mining of Electronic Messages for Multiple Authorship: First Results", SIGKDD'03, August 2003 pp. 24-27, Washington, DC, USA, ACM 1 -58113-737-0/03/0008.

[20.] Rong Zheng, "A Framework for Authorship Identification of Online Messages: Writing-Style Features and Classification Techniques," Wiley Periodicals, Inc., Published online 21 December 2005 (www. interscience. wiley.com).
[21.] Jiexun Li, RongZheng, and Hisinchun Chen, "From Fingerprint to Writeprint," Communication of ACM, April 2006 Vol. 49 No. 4 pp. 76-82.
[22.] Prasad, R.S., U.V. Kulkarni, "Implementation and Evaluation of Evolutionary Connectionist Approaches to Automated Text Summarization," Journal of Computer Science 6 (1 1) 2010, pp. 1366-1376, ISSN 1549-3636.
[23.] Jacques Savoy, "Authorship attribution based on a probabilistic topic model," Information Processing and Management 49 (2013) Elsevier Pub. pp. 341 –354.

AUTHORS

Dr. B. S. Daga has received the Bachelor in Computer Engineering from Government Engineering College, Amravati in 1990 and Master in Computer Science & Engineering from National Institute of Technology, Allahabad. He has completed the Ph.D. degree with Department of Computer Engineering, SGB, Amravati University. He has also worked as Coordinator in Entrepreneurship Development Programs with Department of Science & Technology (Government of India). His research interest includes Multimedia Systems, Data Mining, Artificial Intelligence and Machine Learning.

Jason Dsouza has completed his B.E. in Computer Engineering from Fr, Conceicao Rodrigues College of Engineering, Mumbai. He was also the Club Service Director (2016-17) and Treasurer (2017-18) of The Rotaract Club of Fr. CRCE, a student body that conducts various social and charity events across the city. He has strong communication skills and writes novels as a hobby. With respect to academics, he is fluent in various areas in the field of Computer Science. His areas of interest include Machine Learning, Artificial Intelligence and Databases.

Ryan Furtado acquired his degree in engineering from Fr. Conceicao Rodrigues College of Engineering in the field of Computers. He is fluent in various programming languages and has worked especially with designing websites and databases. He has exceptional speaking skills and writes really well too. He was also the President of the Students' Council of his college and led his team to organize numerous events throughout the year.

Manupendra Tiwari has completed his Bachelor's degree in Computer Engineering from Fr. Conceicao Rodrigues College of Engineering, which is one of the best institutes in Mumbai, India. He is currently working at Morgan Stanley Capital International as an Analyst in the Quality Assurance Team. He is passionate about solving problems related to Machine Learning and Artificial Intelligence.