Journal of Theoretical and Applied Information Technology
10th March 2016. Vol.85. No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AUTHORSHIP ATTRIBUTION OF TELUGU TEXTS BASED ON SYNTACTIC FEATURES AND MACHINE LEARNING TECHNIQUES

[1]N V GANAPATHI RAJU, [2]Dr V VIJAY KUMAR, [3]Dr O SRINIVASA RAO

[1]Associate Professor, GRIET, Hyderabad, Research Scholar, JNTU Kakinada, India
[2] Dean, Professor of Computer Sciences, Anurag Group of Institutions, Hyderabad, India
[3]Associate Professor, Dept. of CSE, JNTUK, Kakinada.

E-mail: [1]nvgraju@griet.ac.in, [2]drvvk_dean@cvsr.ac.in, [3]osr_phd@yahoo.com

ABSTRACT

The automatic recognition of an author of a document on the basis of linguistic features of the text is known as authorship attribution and the present paper performs this on one of the very popular and largely spoken languages of India "Telugu". The present paper strongly believes that each author has got his own unique style of writing pattern, which is the signature of that author. The author attribution is similar to text categorization based on stylistic properties that deals with properties of the form of linguistic expression as opposed to the content of a text. The present paper is based on "shallow" features such as function words frequencies and part of speech (POS). The present paper experimented with a corpus that consists editorial articles of Telugu language by different journalists. The token and lexical based features are not considered because all the documents are in a similar genre and roughly constant over the different authors. The present paper focused on the use of syntax-based (shallow) features of an author's style, and evaluated most frequently used syntactic N-gram (unigram, bi-gram and tri-gram with and without overlapping) POS tagging features after performing the preprocessing step. The present paper also computed authorship attribution by considering Avyayas (similar to stop words in English language) of Telugu language. Further the present paper integrated the above two cases (POS tagging with Avyayas) in finding authorship attribution. Modern supervised machine learning algorithms are used by the present paper to explore large feature vectors to achieve high attribution accuracy. We have achieved an average of above 85% attribution rate on all classifiers with different feature vectors.
Keywords: *N-Gram, POS Tagging, Function Words, Shallow Features, Lexical; Stop Words*

1. INTRODUCTION

Telugu is a South-Central Dravidian language with the third largest number of native speakers in India (75 million). It is one of the twenty-two scheduled languages of the Republic of India and primarily spoken in the state of Andhra Pradesh and Telangana. It is also spoken in some neighboring states as well as in the town of Yanam where it is also an official language. It is also spoken by significant minorities in the Andaman and Nicobar Islands, Chhattisgarh, Karnataka, Odisha, Tamil Nadu, and Puducherry, and by the Sri Lankan Gypsy people. It is one of six languages designated as a classical language of India by the Government of India.

Authorship Attribution has been an area of active research of late. The huge number of various social networks had dramatically increased the availability of online digital media. A need to know authors behind tweets, blogs and Facebook feeds had become an increasing interest for many researchers in vast range of applications, in stylometry, forensics, intelligence, criminal law etc. The nascent scientific areas of Information Retrieval (IR) and Natural Language Processing (NLP) have boosted the authorship attribution research by a great deal.

Authorship attribution is the science of deducing characteristics of an author from the characteristics of documents written of the same author [2, 4, 14]. Its roots are from a linguistic research area called stylometry. Stylometry is statistical analysis of variations in literary style of a document. It makes the basic assumption that an author has distinctive writing habits that are displayed in features such as the author's core vocabulary usage, sentence complexity and the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

phraseology that is used. A further assumption is that these habits are unconscious and deeply ingrained, meaning that even if one were to make a conscious effort to disguise one's style this would be difficult to achieve. Stylometry attempts to define the features of an author's style and to determine statistical methods to measure these features so that the similarity between two or more pieces of text can be analysed.

Many researchers are worked on authorship attribution to quantify the writing styles of the authors by considering various stylometric features. The features can be classified as lexical, character, syntactic, semantic and function words. Lexical features include word length, sentence length, word frequencies, vocabulary richness functions, word n-grams etc. Character features include frequency of character types, frequency of character n-grams. Most lexical features are highly author and language dependent. Hence, the rules deduced by Machine Learning classifiers cannot be applied to other authors or other languages [12]. Syntactic features needs the support of some type of Natural Language Processing tool, like a Part-of-Speech Tagger or a Shallow Parser [18]. The recent contributions in authorship attribution are based on words and their occurrence frequencies. But the frequencies of occurrence of POS tags in a text seems to be a new route for authorship attribution that still needs to be explored [18]. Semantic features include synonyms, semantic dependencies etc.

For authorship attribution, the most frequent words have contributed as the most utilitarian feature. The most common words (articles, prepositions, pronouns, etc.) are the best features to distinguish between authors [1, 20, 21]. They carry no semantic information and they are usually called 'function' words. The selection of the function words is based on arbitrary criteria which is generally language-dependent [1]. Various authors worked on functional and significant words of English language for author attribution [1, 20]. Due to their high frequency in the language and highly grammaticalized roles, function words are questionable to be subject to conscious control by the author. Also to be considered is that the frequencies of different function words vary extensively across different authors and genres of text – hence the hope that modeling the interdependence of different function word frequencies with style will result in effective attribution [15].

This paper is an attempt at finding a good method to perform authorship attribution on the Telugu texts. In this paper we compare N-gram POS tagging feature with function word feature together with the supervised machine learning methods for finding correct author of an unknown document. The present paper is organized as follows. The literature is presented in section two. The section 3 and 4 describes the methodology and results and discussion. The conclusions are presented in section 5.

2. LITERATURE

The common approach to determining authorship is to use stylistic analysis that proceeds in two steps: first, specific style markers are extracted, and second, a classification procedure is applied to the resulting description [17]. To extract the style markers of an author we considered a syntactic feature known as N-gram based POS tagging features, where "N-gram" is the term for any sequence of *n* words/n characters. In natural language processing the presence of one-, two-, and three-word sequences is known as unigrams, bigrams, and trigrams, respectively. Part-of-Speech (POS) tags can be subdivided into open (new words can be added) and closed class words (a fixed set of words). The open class consists of nouns, verbs, adjectives and verbs, and the closed class contains prepositions, determiners, pronouns, conjunctions, auxiliary verbs, particles and numerals. A POS Tagger automatically assigns a POS tag to every word in a text.

Similar syntactic pattern are inadvertently a common occurrence. Therefore, they are more reliable than lexical patterns. There has been success of with function words in representing style. This indicates the usefulness of syntactic information as we usually come across them in certain syntactic structures. This necessitates resilient and precise NLP tools able to perform syntactic analysis of texts. This implies that the syntactic measure uprooting is a language-dependent procedure. This is because it depends fully on the availability of a parser able to analyse a particular natural language with high accuracy [1].

Some authors suggests that the frequencies with which syntactic rewrite rules are put to use, provide at least as good cue to authorship as word usage[6]. Argamon et.al. tried author attribution problem with 500 function words and 685 POS trigrams on newspaper articles and magazine

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

articles[7]. Kukushkina et.al. professed that the frequencies of usage of letter pairs and pairs of grammatical classes are steady characteristics of the author [8]. Koppel et.al. considered 59 POS bi-grams features for the author identification and suggested that syntax detected using automated means would certainly help improve accuracy even more [9]. Diederich et. al. investigated on German newspapers by ignoring nouns, verbs and adjectives and replaced them by grammatical tags and bigrams which resulted in slightly reduced performance in author identification[10]. Gamon et.al. affirmed that the authorship attribution using "shallow" features such as function word frequencies and part of speech trigrams results high classification accuracy in style-based task [11]. Luyckx et.al. credits that the results clearly open up new perspectives for further research on combining automatically extracted syntax-based features and Machine Learning techniques for authorship attribution[12]. Zhao et.al. remarked that giving small training samples, simpler style markers such as the function words were generally better [13]. With larger numbers of training samples, and harder tasks, richer style markers can achieve better performance, such as Function words/POS. [16] Considered 732 POS bigrams and the 1,000 most frequent POS trigrams for the authorship identification.

3. METHODOLOGY

The present paper initially carried pre-processing on the given Telugu document. Annotation of POS tagging on every word token is performed and they are divided in to various n-gram features with and without overlapping. Later the frequencies of POS-N-gram features are represented as feature vectors and a supervised machine learning algorithms is applied for a robust and accurate identification of the correct author.

The detailed explanation is given below.

Step one: In step one preprocessing is performed on Telugu editorial documents. In preprocessing normalization is performed to increases the quality of extracted features. In the process of normalization we converted the documents collected from various leading Telugu newspapers in to Unicode. The Unicode Consortium has allotted 0C00-0C7F codes to represent Telugu characters. The present paper considered only the characters between OC00-OC7F, thus it eliminated all the characters of the other languages and special characters. For this we have developed a tool in Python. The preprocessed documents are then segmented into word tokens.

Second step: This POS tagging is applied in second step. That is we assigned Part-of-Speech (POS) Tagger to each of the word tokens of the step one. POS include nouns, verbs, adverbs, adjectives, pronouns, conjunction and their sub-categories. We annotated, every word token using a shallow parser tool. It is a piece of software that reads text in some language and assigns POS to each word. Bureau of Indian Standards (BIS) - POS defined 11 POS categories for Telugu language and each of them has some sub categories and they are shown in table 1. The total numbers of subcategories of POS are 22. The present paper considered all of them. The BIS prepared POS tag set for the other Indian Languages also. Table 2 illustrates the POS tagging for one sentence of Telugu language.

Table 1: POS Tagset for Telugu language

S.No.	Category		Label	Example
	Type	Sub Type		
1	Noun		N	
		Common	NN	ఢిల్లీలో, రాష్ట్రాలకు
		Proper	NNP	భారత, కమలం
		Nloc	NST	ఇటీవల, కింద
2	Pronoun		PR	
		Personal	PRP	ఆయన, దీనితో
		Reflexive	PRF (PSP)	కూడా, నుంచి
3	Demonstrative		DM	ఈ, ఆ
4	Verb		V	
		Main	VM	వెచ్చదనానికి, లొంగిపోతున్న
		Auxiliary	VAUX	కాదు, ఉండవచ్చు
5	Adjective		JJ	సరైన, ఇతర
6	Adverb		RB	అప్రమత్తంగా, రాజకీయంగా
7	Postposition		PSP	వల్ల, కూడా
8	Conjunction		CC	కానీ, అంతేకాక

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

		Quotative	UT	అనే, అంటే
9	Particles		RP	అంతే, గాను
		Default	RPD (RDP)	చిన్న, కీ
		Classifier	CL	మంది, ఎంతోమంది
		Interjection	INJ	అయ్యో, పాపం
		Intensifier	INTF	మరింత, బాగా
10	Quantifiers		QT	
		General	QTF (QF)	మరో, కొద్ది
		Cardinals	QTC (QC)	రెండు, పది
		Ordinals	QTO (QO)	రెండో, ఒక్కో
11	Residuals		RD	
		Symbol	SYM	. , , ?
		Punctuation	PUNC	“ , ‘
		Unknown	UNK	
		Question words	WQ	ఏమంటే, ఎవరైనా

Step three: We converted annotated POS word tokens in to n-gram POS (NgPOS) features. The present paper considers five NgPOS namely 1) non-overlapped POS unigrams 2) non-overlapped POS bigrams 3) non-overlapped POS trigrams 4) overlapped POS bigrams 5) overlapped POS trigrams. All the 22 sub categories of POS tags are applied on all the above five NgPOS features.

Table 2: POS Tagging for a Telugu document

Telugu Editorial document	నాయకుడన్న వ్యక్తికి విశ్వసనీయత ఉండాలి, ఆకర్షణ ఉండాలి. ఆత్మవిమర్శ చేసుకునేందుకు ధైర్యసాహసాలు ఉండాలి. విచిత్రమేమంటే మన దేశంలో ప్రజల పక్షం వహిస్తున్నామన్న నేతలకెవరికీ ఇలాంటి లక్షణాలు లేకుండానే పదవుల్లో వేళ్లాడుతున్నారు. అందుకే నరేంద్రమోదీ ఇవాళ అధికారంలోకి రాగలిగారు. ఆయనను కదిలించడం అంత సులభం కాదు
Preprocessig and Lexical (word) representation	నాయకుడన్నవ్యక్తికి విశ్వసనీయత ఉండాలి ఆకర్షణ ఉండాలి ఆత్మవిమర్శ చేసుకునేందుకు ధైర్యసాహసాలు ఉండాలి విచిత్రమేమంటే మన దేశంలో ప్రజల పక్షం వహిస్తున్నామన్న నేతలకెవరికీ ఇలాంటి లక్షణాలు లేకుండానే పదవుల్లో వేళ్లాడుతున్నారు అందుకే నరేంద్రమోదీ ఇవాళ అధికారంలోకి రాగలిగారు ఆయనను కదిలించడం అంత సులభం కాదు
Partial output of Shallow parser tool	<Sentence id="1"> 1 ((NP<fs af='నాయకుడన్నవ్యక్తికి,unk,,,,,,' head="నాయకుడన్నవ్యక్తికి" poslcat="NM"> 1.1 నాయకుడన్నవ్యక్తికి NN <fs a f = 'నాయకుడన్నవ్యక్తికి ,unk,,,,,,' name="నాయకుడన్నవ్యక్తికి" poslcat="NM">)) 2 ((NP<fs af='విశ్వసనీయత,unk,,,,,,' head="విశ్వసనీయత" poslcat="NM"> 2.1 విశ్వసనీయత NN<fs af='విశ్వసనీయత,unk,,,,,,' name="విశ్వసనీయత" poslcat="NM">)) 3 ((VGNF <fsaf='ఉండు,v,any,any,any,,ఆలి,Ali'head="ఉండాలి_3">3.1 ఉండాలి VM <fs af='ఉండు,v,any,any,any,,ఆలి,Ali' name="ఉండాలి_3">)) 4 ((NP <fs af = 'ఆకర్షణ , n , , sg , ,d,0,0'head="ఆకర్షణ">4.1ఆకర్షణNN<fs af='ఆకర్షణ,n,,sg,,d,0,0' name="ఆకర్షణ">)) 5 ((VGNF <fs af = 'ఉండు, v ,any ,any, any,,ఆలి,Ali' head="ఉండాలి_2"> 5.1ఉండాలిVM<fs af =' ఉండు ,v, any,any,any,,ఆలి,Ali' name="ఉండాలి_2">)) </Sentence>
POS tagging	<Sentence id="1"> 1.నాయకుడుNN,2.అన్నUT,3.వ్యక్తికిNN,4.విశ్వసనీయత NN, 5. ఉండాలిVM, 6.ఆకర్షణNN,7.ఉండాలి VM, 8.ఆత్మవిమర్శNN,9.చేసుVM,10.కునేందుకు NN, 11.ధైర్యసాహసాలుNN,12.ఉండాలి VM, 13.విచిత్రముNN ,14.ఏమంటే WQ, 15.మన PRP, 16.దేశంలోNN, 17.ప్రజల NN, 18.పక్షంNN ,19.వహిస్తున్నాము VM, 20.అన్న UT, 21.నేతలకెవరికీ NN, 22.ఇలాంటి JJ, 23.లక్షణాలుNN, 24.లేకుండానే VM, 25.పదవులలో NN, 26.వేళ్లాడుతున్నారు VM, 27.అందుకే RP,28.నరేంద్రమోదీ NN,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

	29.ఇవాళ NST, 30 . అధికారంలోకి NN, 31.రాగలిగారు VM, 32.ఆయనను PRP, 33.కదిలించడం VM, 34.అంత QF, 35.సులభం NN, 36.కాదు VM </Sentence>

Step four: This step evaluates the term frequencies of five different NgPOS features on all documents of each and every author using our Python tool.

Step five: This step computes most frequently used above five NgPOS features from the above step. These mostly used NgPOS features are used for future reference. The most frequently used NgPOS features are selected based on a random threshold and in our case it is the above average. This reduces overall complexity.

Step six: various machine learning classifiers are used to predict the author of unknown document.

4. RESULTS AND DISCUSSION

The corpus used for this paper is collected from the editorial columns of leading Telugu newspaper i.e. Eenadu(ఈనాడు), Andhra Jyothi(ఆంధ్రజ్యోతి), Namaste Telangana(నమస్తేతెలంగాణ) of six authors namely A Krishna Rao (AKR) (ఎ.కృష్ణారావు), Allam Narayana(AN) (అల్లంనారాయణ), Ananda Sai Swamy (ASS) (ఆనందసాయిస్వామి), BharathJanjanwala (BJ) (భరత్‌ఝున్‌ఝున్‌వాలా), ChakkilamVijaya Lakshmi (CVL) (చక్కిలంవిజయలక్ష్మి), KattaShekar Reddy(KSR)(కట్టా శేఖర్ రెడ్డి). We have collected around 40 documents of the each author and this leads a total of 240 editorial documents. Out of these we have chosen randomly 25 documents per author (leads to a total of 150 documents) as training data base and the remaining 90 documents as testing database. Table 3 represents the author names - documents and other attributes.

Table 3: The Telugu language corpus (editorial) attributes

Names of the Author (Author labels)	No of documents	Size in kb	Type of articles
A Krishna Rao (AKR)	40	1704	Political
Allam Narayana (AN)	40	1374	Political
Ananda Sai Swamy (ASS)	40	734	Spiritual
BharathJ anjanwala (BJ)	40	822	Political
ChakkilamVijaya Lakshmi (CVL)	40	798	Spiritual
KattaShekar Reddy (KSR)	40	1002	Political
Total documents	240	6434	

For an accurate author attribution rate the present paper used four different machine learning classifiers 1) Naïve Bayes classifier (NB) (Lewis 1998), 2) Support-Vector Machines (SVM) using Sequential Minimal Optimization (Platt, 1998) with a linear kernel and default settings, 3) J4.8 decision tree method (Quinlan 1986) with no pruning Decision trees classifier (DT), and 4) Multilayer Perception algorithms (MP) with varied parameters and five-fold cross-validations. To predict unknown author we used Weka (Waikato Environment for Knowledge Analysis) software package (Witten and Frank, 1999). Weka is applied on the derived most frequently used NgPOS shallow features of training data set and query document using CSV format. We have used Version 3.7 implementation of Weka for the identification of an unknown author on the above four machine learning algorithms.

The novelty of the present paper is, it evaluated author attribution based on three different modes as explained below.

Mode 1: The present paper initially performed preprocessing on the Telugu documents and extracted NgPOS shallow tags on five different n-grams. The most frequently used five different NgPOS shallow features are evaluated. Table 4, 5 and 6 shows the attribution rate using above NgPOS features without overlapping and with overlapping respectively.

Table 4: Attribution rate based on unigram POS shallow features for mode 1

Names of the authors	Unigram-POS classifiers			
	NB	SVM	MP	DT
AKR	76.47	73.94	77.3	71.4
AN	75.63	73.1	75.6	73.1
ASS	75.63	73.1	75.6	73.1
BJ	74.78	72.26	72.3	71.4
CVL	71.42	73.94	71.4	69.7
KSR	73.1	74.78	78.2	73.9

Journal of Theoretical and Applied Information Technology
10th March 2016. Vol.85. No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table5: Attribution rate based on NgPOS shallow features for mode 1 without overlapping

Names of the authors	Bigram-POS Non overlapping				Trigram- POS Non overlapping			
	classifiers				classifiers			
	NB	SVM	MP	DT	NB	SVM	MP	DT
AKR	74.5	80.39	76.47	73.52	68.9	76.5	66.38	57.98
AN	76.5	78.43	77.45	77.45	72.3	77.3	68.06	52.94
ASS	75.5	77.45	73.52	75.49	73.9	77.3	66.38	63.86
BJ	74.5	81.37	73.52	80.32	68.1	73.9	68.06	51.26
CVL	73.5	80.39	74.5	77.51	69.7	76.5	68.06	52.94
KSR	68.1	75.63	73.1	65.55	68.9	73.9	73.94	52.1

Table 6: Attribution rate based on NgPOS shallow features for mode 1 with overlapping

Names of the authors	Bi-gram-POS with overlapping				Tri-gram- POS with overlapping			
	classifiers				classifiers			
	NB	SVM	MP	DT	NB	SVM	MP	DT
AKR	78.99	78.15	78.15	70.58	84.87	80.67	78.15	62.2
AN	79.83	78.99	82.35	66.38	85.71	83.19	83.19	50.7
AS	81.52	78.15	79.83	71.42	87.39	84.03	78.15	65.5
BJ	78.15	79.83	76.47	70.58	84.87	82.35	79.83	58.8
CVL	76.47	77.31	73.94	66.38	81.52	78.99	79.83	61.3
KSR	78.99	79.83	81.51	70.58	83.19	81.51	78.15	67.2

From the above table, it's clearly observed that the performance of all classifier's on an average is 75%. It's also observed that overlapping Trigram – POS shallow feature out performs all other features and NB and SVM classifiers are performing better in authorship attribution.

Mode 2: In mode 2 the present paper considered functional (Stop) words of Telugu language known as Avyayas for the authorship attribution. The Telugu language Avyayas alone does not have any meaning like stop words in English. The CALTS Lab of University of Hyderabad, Hyderabad, India derived 2600 Avyayas tokens in Telugu language. For our experiment sake we have considered Avyayas of Telugu language defined by CALTS Lab. The novelty of the present method is, we have not evaluated all the above 2600 avyaya tokens. We have evaluated most frequently used avyaya tokens from our training corpus document ranging from 50 to 800 and carried out our experiment. Table 6 shows the attribution rate of different classifiers based on most frequently used 500 and 250 avyayas. This allows testing the relevance of selecting function words as clues for authorship attribution. The graph of figure 1 shows the average author attribution rate of all six authors on the different classifiers with increasing number of most frequently used Telugu avyayas ranging from 50 to 800. The graph clearly shows, the attribution rate increases from 50 most frequently used avyayas to around 250. After 250 the graphs are constant with a slight increase.

Table 7: The author attribution rate on different classifiers using 250 and 500 Telugu avyayas.

Names of the authors	500 Avyaya words				250 Avyaya words			
	classifier				classifier			
	NB	SVM	MP	DT	NB	SVM	MP	DT
AKR	82.78	82.78	89.34	57.37	82.78	85.24	87.7	63.11
AN	81.14	81.14	92.62	67.21	81.96	81.96	88.52	59.01
AS	82.78	85.24	90.16	66.39	83.6	85.24	86.88	62.29
BJW	78.68	81.14	90.16	68.65	77.86	85.24	90.16	61.47
CVL	81.96	78.68	88.52	57.37	81.96	81.96	85.24	56.55
KSR	82.78	81.14	90.16	60.65	82.78	84.42	89.34	52.45

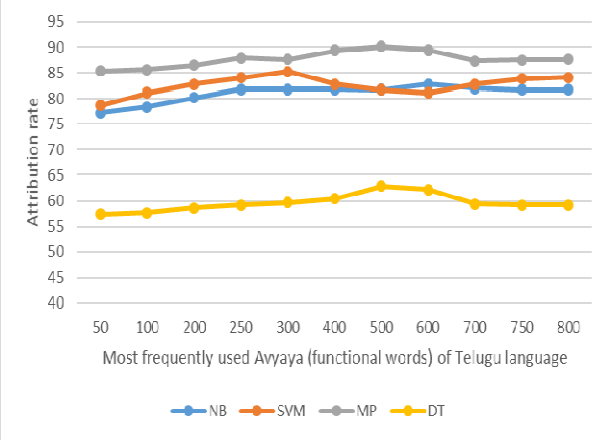


Figure 1: author attribution rate verses most frequently used avyayas

Table 7 and figure 1 shows better performance than mode 1. The performance is more or less similar either by considering most frequently used 250 or 500 avyayas. This indicates one need not necessarily to test on all 2600 avyayas for author attribution. The above table illustrates that the average accuracy is between 80% and 85% for all classifiers and especially and MP classifier shown an accuracy of 90.16.

Mode 3: integrated scheme: In mode 3 we have integrated mode 1 and mode 2 with only 250

Journal of Theoretical and Applied Information Technology
10th March 2016. Vol.85. No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

avyayas. Table 7, 8 and 9 represents the attribution rates with different NgPOS features combined with 250 avyayas, with overlapping and non-overlapping methods respectively.

Table 7: Attribution rate on different classifiers based on 250 avyayas and unigrramPOS

Names of the authors	Unigram POS and 250 avyayas (non- overlapped)			
	classifiers			
	NB	SVM	MP	DT
AKR	82.78	88.52	91.06	63.93
AN	81.14	88.52	91.8	64.75
AS	84.12	90.16	93.44	69.67
BJ	78.68	91.8	92.62	65.57
CVL	81.14	86.06	88.52	63.93
KSR	83.6	90.16	90.98	75.4

Table 8: Attribution rate on different classifiers based on 250 avyayas and NgPOS non-overlapping features.

Names of the authors	Bigram POS and 250 avyayas (non- overlapped)				Trigarm POS and 250 avyayas (non- overlapped)			
	Classifiers				classifiers			
	NB	SVM	MP	DT	NB	SVM	MP	DT
AKR	80.32	93.44	90.98	62.29	82.64	95.04	93.98	71.07
AN	79.5	92.63	90.16	68.-03	83.47	95.86	94.16	66.11
AS	80.32	93.44	93.44	70.49	81.01	95.04	93.44	66.46
BJ	77.86	92.62	90.98	68.03	79.38	91.73	90.98	74.38
CVL	81.14	91.8	89.34	63.93	80.16	94.21	93.34	71.07
KSR	81.14	92.63	91.8	65.75	81.14	94.26	93.8	63.11

Table 9: Attribution rate on different classifiers based on 250 avyayas and NgPOS with overlapping features.

Names of the authors	Overlapped bigram POS and 250 avyayas				Overlapped Trigram POS and 250 avyayas			
	classifiers				classifiers			
	NB	SVM	MP	DT	NB	SVM	MP	DT
AKR	75.4	91.8	90.98	68.03	82.78	90.98	90.98	71.31
AN	75.4	90.16	89.34	72.13	83.6	89.34	89.34	65.57
AS	77.04	92.62	93.44	74.59	84.42	91.8	93.44	71.31
BJ	77.04	90.98	90.93	65.57	80.32	90.98	90.93	68.35
CVL	78.68	88.52	89.16	63.93	83.6	90.16	90.16	64.75
KSR	79.5	86.06	91.8	77.13	86.88	90.16	93.44	71.31

Mode 3 on average exhibited more than 90% of attribution rate. It can also be perceived that Trigram- POS feature without overlapping and with 250 avyayas out-performed all other features with an average accuracy of 94.26% for SVM classifier. The mode 3 outperformed mode 1 and mode 2 and this is visible from the graphs of figures from 2 to 6.

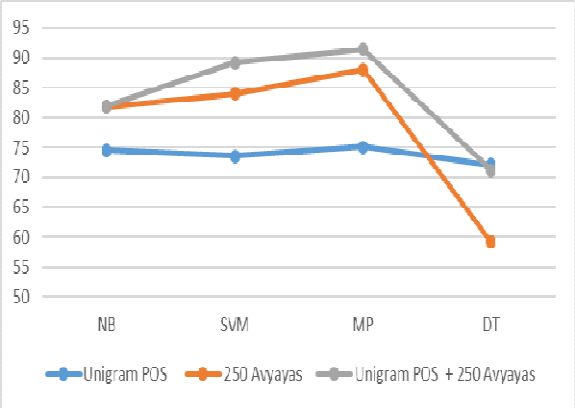


Fig 2: analysis non-overlapped case (Unigram): Unigram POS (mode1), 250 Avyayas (mode2), Unigram POS and 250 Avyayas (mode3) features on classifiers.

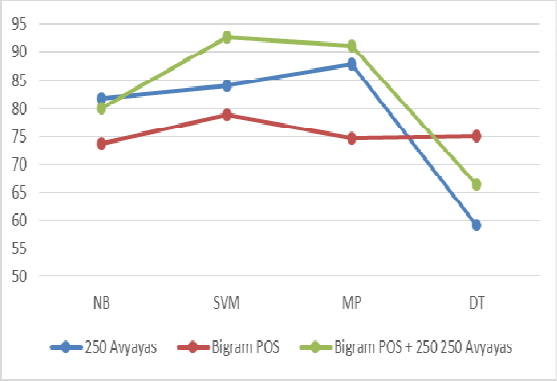


Fig 3: analysis of non-overlapped case (Bi-grams): Bigram POS (mode1), 250 Avyayas (mode2), Bigram POS + 250 Avyayas (mode3) features on classifiers

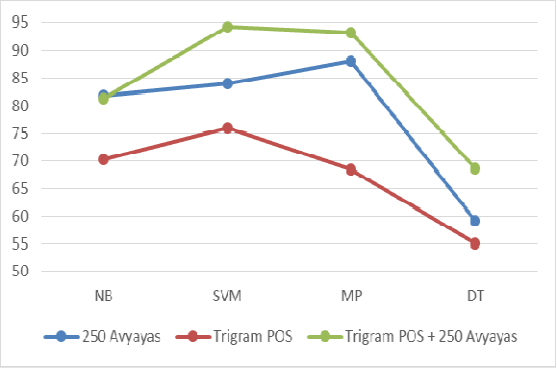
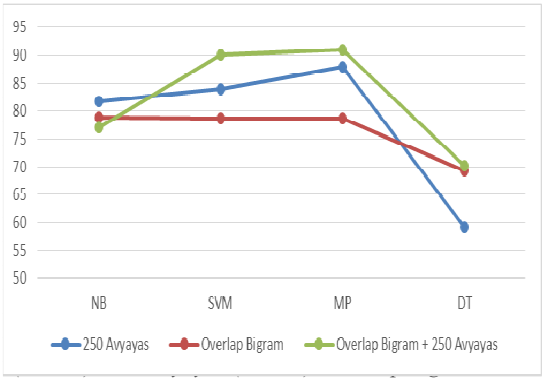


Fig 4: analysis of non-overlapped case (Tri-grams): Trigram POS (mode1), 250 Avyayas (mode2), Trigram POS and 250 Avyayas (mode 3) features on classifiers

Journal of Theoretical and Applied Information Technology
10th March 2016. Vol.85. No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195



and 250 Avyayas (mode 3) features on classifiers

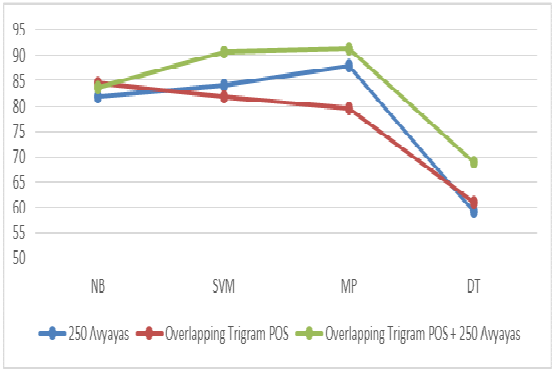


Fig 6: analysis of Overlap Tri-gram: Tri-gram POS (mode1), 250 Avyayas (mode 2), Overlap Trigram POS and 250 Avyayas (mode 3) features on classifiers

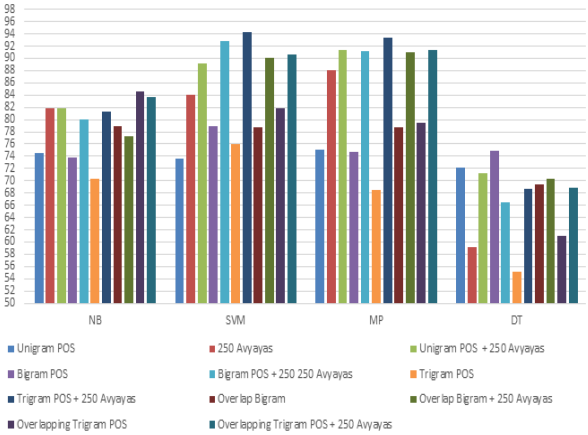


Fig 7: comparison of various N gram POS and 250 Avyaya features on machine learning classifiers

The graph shown in Fig 7 illustrates the attribution rate of all four four classifiers based on three different modes. SVM and MP classifiers performed better then rest.

5. CONCLUSION& FUTURE WORK

We have a presented a new method to automated authorship attribution based on 1) most frequently used word n-gram POS features and 2) most frequently used avyayas in the document 3) integration of 1 and 2. We have illustrated the feasibility of our approach on a corpus consisting of newspaper articles of one of the popular and official languages of two states of India 'Telugu'. We have obtained a state of the art performance. The present method obtained syntax-based features of the documents by considering bi-gram and tri-gram POS tags. The best syntax-based feature sets are based on the distribution of parts-of-speech (POS). We have used various machine learning classifiers and compared the performance. Out of the three modes the integration method with trigram-POS features with 250 ayavyas achieved 94.36% of attribution rate and it outperformed the other modes. The SVM and MP classifiers achieved higher attribution rate in all three cases with the derived features than other classifiers.

ACKNOWLEDGMENT

The authors would like to express their gratitude to Management, Anurag Group of Institutions (AGOI), Hyderabad for providing necessary infrastructure for Centre for Advanced Computational Research (CACR) at AGOI, which is bringing various research scholars across the nation to work under one roof. The CACR is providing a research platform for exchanging and discussing various views on different research topics related to computer science. Authors extended their gratitude to Management, Gokaraju Rangaraju Institute of Engineering & Technology, Bachupally, Kukatpally, Hyderabad, India. This research has been supported by UGC under minor research project grant MRP-4590/14 (SERO/UGC) in March 2014.

REFERENCES

[1]. EfstathiosStamatatos, "A Survey of Modern Authorship Attribution Methods", Journal of the American Society for Information Science and Technology, Volume 60 Issue 3, Pages 538-556, March 2009.

[2]. P. Juola, "Authorship Attribution", Journal of Foundations and Trends in Information Retrieval, Vol 1, Issue 3, 2006, pp 233-334, 7 March 2008.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[3] VladoKeˇselj, Fuchun Peng, Nick Cercone, Calvin Thomas,"N-Gram-Based Author Profiles For Authorship Attribution", Pacific Association for Computational Linguistics, 2003.

[4] Jonathan Doyle,VladoKeˇselj, "Automatic Categorization of Author Gender via N-Gram Analysis", The 6th Symposium on Natural Language Processing,2005

[5] EfstathiosStamatatos, "On the robustness of authorship attribution based on character n-gram features", Journal of Law and Policy, 2013

[6] H. Baayen, H. Van Halteren, and F. Tweedie, "Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution," Literary and Linguistic Computing, vol. 11, pp. 121-132, 1996.

[7] Argamon-Engelson, S., Koppel, M., &Avneri, G., "Style-based text categorization: What newspaper am I reading?", In Proceedings of AAAI Workshop on Learning for Text Categorization, Pages 1-4,1998.

[8] Kukushkina, O.V., Polikarpov, A.A., &Khmelev, D.V., "Using literal and grammatical statistics for authorship attribution", Problems of Information Transmission, Volume 37(2), Pages 172-184.2001

[9] M. Koppel and J. Schler, "Exploiting stylistic idiosyncrasies for authorship attribution." IJCAI'03 Workshop on Computational Approaches to StyleAnalysis and Synthesis, pp. 69-72, 2003.

[10] Diederich, J., Kindermann, J., Leopold, E., Paass, G., "Authorship attribution with support vector machines", Journal of Applied Intelligence, volume 19(1/2), pages 109-123.2003

[11] M. Gamon, "Linguistic correlates of style: authorship classification with deep linguistic analysis features," in Proceedings of the 20th international conference on Computational Linguistics, 2004, p. 611.

[12] Kim Luyckx,Walter Daelemans, "Shallow Text Analysis and Machine Learning for Authorship Attribution", 2005.

[13] Ying Zhao, Justin Zobel, "Searching with Style: Authorship Attribution in classic literature", Proceeding ACSC, Volume 62, Pages 59-68, 2007.

[14] Marcin Opacki,"Stylometry and Authorship attribution"

[15] Argamon, S.,Levitan, S., "Measuring the usefulness of function words for authorship attribution", In Proceedings of the Joint Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing,2005

[16] Ludovic Tanguy, AssafUrieli, BasilioCalderone, Nabil Hathout, Franck Sajous, "A multitude of linguistically-rich features for authorship attribution", 2012.

[17] Fuchun Peng, Dale Schuurmans, VladoKeselj, Shaojun Wang, "Language Independent Authorship Attribution using Character Level Language Models", 2003.

[18] Kim Luyckx, "Syntax-Based Features and Machine Learning techniques for Authorship Attribution", Ph.D. Thesis, 2003-2004.

[19] Ramyaa, Congzhou He, Khaled Rasheed, "Using Machine Learning Techniques for Stylometry", Proceeding of the IC-AI, 2004.

[20] Burrows, J.F. "Word patterns and story Shapes: The statistical analysis of narrative style". Literary and Linguistic Computing, 2, 61-70, 1987.

[21] Argamon, S., Levitan, S, "Measuring the usefulness of function words for authorship attribution". In Proceedings of the Joint Conference of the Association for Computers and the humanities and the Association for Literary and Linguistic Computing, 2005.