Chapter 1

Introduction

Authors have different styles of writing for the purpose of sentence construction. Authors follow either Active Voice or Passive Voice while expressing the views on a concept. Concepts can be presented elaborately with more number of sentences or the same concept can be presented with precise and less number of sentences. It is difficult to understand the narration of particular author where as it easy to understand the same concept presented by another author. It shows the different writing styles of authors. The writing style of an author may expose his personality, ideas and opinions on a concept based on his write-ups. The writing style may be influenced by his educational background and based on his experience in the real life in different situations.

In general, the reader’s familiar with novels can easily distinguish the writing styles of novelists. They can easily classify a novel with unknown author to the predefined set of authors by comparing with author writing styles. Stylometry is the Statistical Analysis which studies the differences among writing styles of a set of writers. Stylometry can be used to identify the real author of disputed documents or to identify the author of a document where authorship of a particular document is missing. Stylometry is feasible and more appropriate to address the problem of Authorship Attribution which has many applications in legal issues, academics in terms of plagiarism and literary for unknown authorship.

Focus of research presented in this thesis is to identify the author of a Telugu Text whose authorship is unknown. Firstly, in order to identify the authorship of a given Telugu document various features such as Lexical, Syntactic and Structural features and their combinations are considered. Various Machine Learning Techniques such as Naive Bayes, K- Nearest Neighbor, Support Vector Machines and Decision Trees are also examined to identify the best suitable techniques for Automatic Authorship Attribution as second experiments. The influence of data set
size and authorship set size on different features such as Lexical, Syntactic, and Structural features on Authorship Attribution for Telugu Text is also studied in this thesis. As a part fourth evaluation, the impact of Data Compression Techniques on Authorship Attribution for Telugu Text is also thoroughly evaluated.

This chapter gives an Introduction to the problem of Authorship Attribution, significance of attempting the Authorship Attribution problem in Indian context. The importance of Text Mining in terms of Authorship Attribution is presented. The theoretical introduction about various features, Classification Techniques and Data Compression Techniques are also presented. The various applications of Automatic Authorship Attribution, motivation towards attempting the problem and organization of thesis are presented at the end of the chapter.

1.1 Information Retrieval and Machine Learning:

As the number of Internet user’s increases, the volume of information available on the Web is also increasing exponentially. According to the Internet World Statistics, the number of Internet users exceeds three billion by the end of 2015. It has resulted to an explosion in the availability of electronic information in the web. As the availability of information increases, the availability documents with unknown authorship or miscellaneous authorship documents are also enormously increasing which necessitate attempting the problem of Authorship Attribution.

The branch of Computer Science that deals with facilitating access to large collections of data is called Information Retrieval (IR). Information Retrieval is the science of searching for information within relational databases, documents, text, multimedia files, and the World Wide Web [56]. The applications of Information Retrieval are diverse, they include but not limited to extraction of information from large documents, searching in Digital Libraries, Information Filtering, Spam Filtering, Object Extraction from Images, Automatic Summarization, Document Classification and Clustering, Web Searching, Language Identification and Authorship Attribution.

Machine Learning Algorithms have proven to be very successful in solving many problems such as Text Categorization, Text Summarization and Language
Identification. Machine Learning Algorithms can be divided into two types such as Supervised Algorithms and Unsupervised Learning Algorithms. Supervised learning Algorithms reduce objective models from a set of training examples and then use the learned models on the target set to solve the addressed problem. Unsupervised learning operates to find useful relations among the objects of the target set. The goal of this thesis is to perform the research towards Authorship Attribution on Telugu Text in line of Text Categorization using Supervised Learning Algorithms. Text Categorization and Authorship Attribution are considered as a discipline which at the crossroads of Machine Learning and Information Retrieval as it shares a number of characteristics with these two fields. Since Authorship Attribution is not only a content-based document management task but also need to consider the stylometry of the document.

### 1.2 Text Mining:

Natural Language Processing (NLP) is a combination of techniques from Computer Science, Artificial Intelligence, and Linguistics. Natural Language Processing deals with the human and computer interaction. Natural Language Processing applications are to extract knowledge from text. Text Mining is a process of extracting meaningful information from huge text set by performing linguistic analysis [57].

Text Categorization is an important step in Text Mining which is also called as Text Classification.

Text Categorization learns the models by classifies the training document set into predefined set of categories. For a test document dj, a Boolean value is assigned by using previously learned models. If a value is true for a pair <dj, ci> which indicates that dj is labeled under category ci. A value of false means that the document dj is not belongs to category ci [57].

There are many Text Categorization applications. Some of the applications are document indexing and document filtering, Word Sense Disambiguation (WSD), web pages hierarchical categorization and language identification [57]. Authorship Attribution is one kind of classification problem. Stylometry is one of the techniques
which are very often used in Authorship Attribution. In Authorship Attribution, the task is to automatically extract author's information from the given text document such as linguistic style which is not under control of author's consciousness [58]. Based on the extracted information from the text document, it can be assigned to one of predefined categories.

1.3 Natural Language Processing in Indian Languages:

Humans have the capacity of understanding complex mechanisms which makes the language growing. Making the computer to learn such mechanisms to explore the language acquisition theory and also to learn the characteristics of the natural language is a primary task. Knowledge of human language can be attained by developing software products that acts more interactively with human knowledge. Machine Learning deals with training computers to understand the logic and the mathematical ability that deals with the language generation.

Linguistic techniques used to learn what constituents are to be retained and what to be removed for Machine Learning. Natural Language Processing approaches are well established in the European languages such as on Chinese, Japanese and Korean scripts. India scripts are still lagging far behind in learning through statistical approaches. Machine Learning research is to understand complex patterns of script features to make intelligent decisions. In Natural Language Processing, the importance of large volume of corpus is being pointed out the need for technology of analyzing the linguistic data. Corpora have proved their value in knowledge acquisition of language technology. Statistical approaches rely on corpora and on few linguistic clues.

Telugu is one of the most commonly used languages in India. It has nature of complex Agglutinative Morphology. Agglutinative Morphology deals with various new words which can be derived by adding suffixes and prefixes to root word. Automatic Authorship Attribution involves some form of Machine Learning and author's writing style rather than an algorithmic approach only. Working with a more complex Agglutinative Morphological language such as Telugu is a real and important research issue in the context of Automatic Authorship Attribution. In
contrast to the other languages, there is no research done on Telugu Text in terms of Automatic Authorship Attribution.

1.4 Authorship Attribution:

Authorship Attribution is a task of assigning a document to an author from a given set of authors. It means, Authorship Attribution identifies the author that who wrote a particular document by characterizing the writing style of a particular author. Authorship Attribution can be performed by identifying the characteristics of author's writing style which are not under the control of author’s consciousness. Based on the number of authors involved in the authorship set, Authorship Attribution task can be divided into five categories. They are presented below.

Binary Authorship Attribution is of simplest type of Authorship Attributions. In Binary Authorship Attribution there are only two authors are considered in the authorship set. In this attribution, it is assumed that the anonymous test document is written by one of those two authors which are considered in the authorship set.

Multi-class Authorship Attribution is also termed as N-class Authorship Attribution, where the number of authors in the authorship set is more than two. In Binary Authorship Attribution the probability of attribution is 50% where as in Multi-class Authorship Attribution with n authors the probability is 1/n. It shows that the more number of authors in the authorship set, the level of difficulty of assigning the document to the correct author more.

Authorship verification is also known as one-class Authorship Attribution problem. In this there is only one author exists in the authorship set. The task of authorship verification is to verify whether the anonymous document is written by the author presented in the authorship set or not. This task can be performed based on the prior knowledge author writing style.

Authorship Identification is to identify the author of an anonymous text when there is no prior knowledge about the author. There is no prior authorship set. The
Identification of an author depends on subject experts and linguistic experts. This is highly difficult to address this problem with an automated process.

Authorship Collaboration or Authorship Sharing deals sharing of a document write up by a group of authors. In this identifying the fragments of a write-up written by which author and also percentage of sharing of document of each author is an important concern. This problem is not widely addressed by the researchers as the lack of availability sufficient number of such documents.

Authorship Attribution can also be visualized in two categories. They are Traditional Authorship Attribution and Modern Authorship Attribution. Traditional Authorship Attribution explores both Internal and External evidences in order to identify the author of a given text. Internal evidence means collecting the authorship proofs of a particular author from the given anonymous text itself. External evidence depends on the domain experts’ expertise in that particular area of domain. Hence Traditional Authorship Attribution involves both human effort and usage of stylometry.

Non-traditional Authorship Attribution also referred as Modern Authorship Attribution which has also referred as Automated Authorship Attribution using Machine Learning techniques to identify author's writing style. Authors writing style can be measured by using set of features which are automatically extracted from the documents. Writing style of an author can be visualized in two ways. They are conscious impressions on the document and unconscious impressions on the document. The conscious writing style can be controlled by authors where as the unconscious writing style is independent of authors interest. The assumption in Modern Authorship Attribution is that each author has a unique style of writing.

An Automated Authorship Attribution approach uses a set of training documents with known authorship. Automated Authorship Attribution follows the concept of Text Categorization. Hence it consists of two main steps as in Automatic Text Categorization such as Feature Extraction and Classification based on the extracted features.
In Traditional Authorship Attribution it requires both human effort and Machine Learning. Traditional Authorship Attribution depends more on external proofs to establish attribute authorship which may take years together to a decision. In Modern Authorship Attribution which uses computers, statistical tools and Machine Learning techniques to assign the Authorship Attribution without no time. Modern Authorship Attribution deals with internal evidences only. Non-traditional Authorship Attribution deals mainly with authorship verification and authorship identification of documents.

1.5 Significance:

The influence of number of parameters on Authorship Attribution approaches will be addressed in the proposed research. Still today Authorship Attribution to Telugu Text is not attempted by the researchers. The influence of attribution methods and features which are applied to text in various languages may not be suitable for Telugu Text for adaptation. Hence, in this thesis the existing statistical approaches, Machine Learning techniques, Data Compression techniques and various features such as Lexical, Syntactic and Structural features which are thoroughly tested on text of various languages need to be tested for its most likely adaptability for text in the Telugu language.

Another important issue addressed in this thesis is the identification of influence of number of authors in the authorship set and also the impact of number of documents in the training data set. The proposed research will explore the affect of features, Machine Learning techniques, Data Compression techniques, data set size and authorship set size on Authorship Attribution for Telugu Text in terms of Natural Language Processing. The combinations of various approaches are empirically evaluated for Authorship Attribution on Telugu Text and effective subsets are identified. For each of these combinations, the changes in their performance based on number of documents in the training set and number of authors in the authorship set are also addressed.

In this thesis, the combinations of various features for Authorship Attribution that have not been used in the context of Indian languages such that Telugu Texts are
identified. The proposed researches apply commonly used approaches in Authorship Analysis such as Decision Tree, Naive Bayes, Nearest Neighbour and Support Vector Machine. It also tested the different Data Compression Algorithms such as Zip, BZip, GZip, LZW, PPM and PPMd in combination with three different compression distance measures such as Normalized Compressor Distance (NCD), Compression Dissimilarity Measure (CDM) and Conditional Complexity of Compression (CCC) for Telugu Text. The influence of feature selection methods on Authorship Attribution is also addressed in this thesis.

1.6 Features:

One of the main tasks in Authorship Attribution is the identification of features or combination of features which can differentiate among the authors. The features are broadly divided into four categories [60, 61] such as Token Level Features, Syntax Based Features, Vocabulary Richness Features and Common Word Frequencies as a feature. They are briefly described below.

1.6.1 Token-Level Features:

Tokenization is a process of dividing the text into set of tokens. The text in a document is considered as a set of tokens. The tokens may be paragraphs, sentences, and a set of words, individual words, and part of a word or even each character syllable. Based on the characteristics of the language on which the text has developed, various types of tokens are suitable for Authorship Attribution. In the literature most of the research in Authorship Attribution deals with tokens as features. This approach is basic and quantifiable and more effective for Authorship Attribution. As the processing power of the computers is low in 1950's, token level features are more powerful for identification of an author of an anonymous text [61]. Even today a token-level feature seems to be simple but tend to work well for Authorship Attribution.

Experiments conducted by Mendenhall [62] in 1880's resulted that word length is a good characteristic for Author Identification. In [63], sentence length was
used as a feature Authorship Attribution. Average number of syllables per word, the
distribution of n-gram of syllables and frequency of n-gram syllable words in a text
are able to capture the style of a specific author. An n-gram of characters or syllables
is a sequence of n characters or syllables in a word. The n-grams could be either
overlapped or non-overlapped n-grams. Probability of occurring n-grams of syllables
and frequencies of occurrence of n-grams of characters [62, 63, 64] are useful features
for identification of style of an author.

1.6.2 Syntax Based Features:

Syntax Based Features are depends on the grammar rules of the language. To
identify the syntax of the language used in the text, there is a need of Natural
Language Processing tools such as Part-Of-Speech identifiers, Morphological
Analyzers, and Grammar Checking Tools. The phrase-level features like number of
noun phrases in the text, number of prepositional phrases in the text are used in [65].
Number of Passive Voice Sentences and number of Active Voice Sentences are also
considered as syntax based features. The number of nouns, pronouns, verbs,
adjectives, adverbs and other types of Part-Of-Speech tags in a text are useful to
identify the stylistic characteristics of an author for Authorship Attribution [61].
Rewrite rules frequency count can also distinguish among the authors [66].

1.6.3 Vocabulary Richness Features:

The writing style of authors differs in syntactic or structural features used in
the text written and also varies in the vocabulary used in the text. There is lot research
work done on Authorship Identification in terms measuring the vocabulary richness
measure of the content in the document. Type-token ratio is a basic measure of
identifying the vocabulary richness of an author. It deals with the frequency
distribution of the sample set. If size of the vocabulary is S in the sample text and T is
the number of tokens in the sample text then the type-token ratio R is calculated as
S/T [67]. Simpson’s Index (D) is another type of vocabulary richness measure and it
is the percentage of two members of an arbitrarily chosen pair of words belongs to the
same type. Yule’s Characteristic (K) assumes that the occurrence of a given word is
based on choice of Poisson distribution [61]. Hapax Legomena means that the percentage of words that occur only once in a given text. Zipf law assumes that there are only few words that occur very often in a text where as there is more number of words that occur only once in a text. Hapax Dislegomena is number of a word that occurs twice in the text [61].

### 1.6.4 Content-Based Word Frequencies:

There are many studies in Authorship Attribution to distinguish among the authors by using frequency count of individual words in the sample text. Functional words or common words which are not under the conscious control of the author [61] are also good features for identification author style. Hoover in [68] experimented with content and functional words to identify the impact on classification of texts of various authors. Stamatelos in [69] uses the frequencies of most occurring words that exist in the training data set for Authorship Attribution which is also proved as more reliable stylistic discriminator for author's writing style.

### 1.7 Classification Techniques:

Authorship Attribution can be viewed as a problem of Text Classification or Document Categorization problem. Hence it could have two phases such as Feature Extraction and Classification. In There is a need of addressing the classification techniques which are successfully applied on Text Classification in order to use for Authorship Attribution. The Classification Techniques can be categorized into three types such as Machine Learning Techniques, Statistical Techniques and Data Compression Techniques.

### 1.7.1 Machine Learning Techniques:

Machine Learning is a process of learning model automatically from the existing system and improves the system with the experience. A Machine Learning Algorithm discovers a set rule from the existing system in order to predict the results for a new system. [70]. Machine Learning Algorithms can be divided into two categories. They are Eager Learning Algorithms and Lazy Learning Algorithms. An
Eager Learning Algorithm learns the rules from the examples and deletes the original examples. A Lazy Learning Algorithm carries all the examples and generates the rules from these examples whenever it is needed. The major focus of Lazy Learning Algorithms is on classification other than on learning. The best example of Lazy Learning Algorithm is K-Nearest Neighbour Algorithm and most of the other algorithms are eager learning algorithms which are Naive Bayes, Decision Tree, Artificial Intelligence, Neural Networks and Support Vector Machines. These algorithms are suitable for the problems of Text Classification, Language Identification and Authorship Attribution.

Naive Bayes Algorithm is based on the joint probabilities of words and categories. Using these probabilities, the probabilities of categories for a given document can be calculated. The words in a training document set are considered as the features to identify the author of an anonymous document [71]. Decision Tree can be applied for Authorship Attribution task by representing discriminatory features as nodes or tests and test conditions or thresholds on branches from the nodes. The leaf nodes represent the authors in the authorship set [72]. When a test document is processed through the Decision Tree, by reaching the leaf node author of the document will be identified. Diederich in [73] uses Support Vector Machines for Authorship Attribution.

1.7.2 Statistical Techniques:

As in [61] by Holmes, the various Statistical Techniques such as Discriminate Analysis (DA), Cluster Analysis (CA) and Principal Components Analysis (PCA) [61] can be adoptable to the Authorship Attribution.

Discriminate Analysis predicts the members of a group on the basis of a Predictor Analysis set. Discriminate Analysis maximizes the differences among the clusters and minimizes the differences within the clusters. Stamatelos and Baayen in [66, 69] used the Discriminate Analysis for Authorship Attribution. There are similarities between cluster analysis discriminate analyses that is both the techniques makes the clusters or groups as internally homogenous and externally heterogeneous.
Hoover in [68] experimented with high-frequency words for grouping texts by the same author and distinguishing texts by different authors. Principal Components Analysis (PCA) is a technique which reduces the dimensionality of the input. In the vector space, Principal Components Analysis rotates the data set and scales the data set to attain the conceptual correlation among the features in the data set and also reduces the dimensionality of the data set. Baayen in [68] uses the Principal Components Analysis to compare the performance of word-based features with syntactic features for Authorship Attribution. Multiple Regression techniques provide the relationship between several predictor variables and target variable.

1.7.3 Data Compression Algorithms:

For the task of Text Categorization, Data Compression Algorithms are best alternative approaches when compared with Machine Learning approaches and Statistical Approaches. According to information theory, entropy is in each message received the average amount of information contained in the message. Message may be an event, sample or character from a data stream. Entropy characterizes the uncertainty source of information.

A Data Compression technique tries to find repeated strings in the text and considers the longest matching strings from the text. A text zipper encodes frequent sequences of words with few bytes where as frequent sequences of words will be encode with more bytes [75]. There are several advantages in using Data Compression techniques for Authorship Attribution such as considering the overall judgments of the document as a whole so that it avoids the problem of word ambiguities, take into account of phrasal effects, uniform way of dealing with different document types so that it minimizes arbitrary decisions [76].

Data Compression Algorithms are used to compress the test document to derive character-based author model. The representative training texts are used to derive training author models for various authors. For each author a model will be generated. The author of the test document is the one that has the best compression performance with the author specific generated models [74].
1.8 Applications:

Authorship Attribution can be applied to different domains of applications. One of the domains is forensic linguistics. The various possible issues in forensic linguistics are for example a statement allegedly made by an accused person, a term paper that could be plagiarized, stealing the content of content in research document by other researchers. Forensic linguistics uses the various features like Linguistic Features, Structural Features, Syntactic Features and Statistical Features to determine the attribution of actual author for different possible issues.

The other type of research in Authorship Attribution is disputed authorship. Using the techniques developed in Authorship Attribution it is possible to detect the Age, Document Written Era, Authors Nationality or Gender [59].

1.9 Motivation:

Authorship Attribution depends on three basic factors. Good knowledge in linguistics gives better choice of selecting style markers for author stylistic analysis. Statistics are the good measures to quantify these style markers. Classification Techniques are used to discriminate the obtained style markers correctly among the authors in the author set.

The research on Authorship Attribution is evaluated on multiple data sets with various corpus sizes makes it difficult to compare the methods used in different approaches. The attribution methods which were successful for one type of data set may not be successful for other data sets. Hence it is not clear whether these methods are scalable, reliably effective, or robust. The problem of Authorship Attribution is not yet attempted on the text of Indian languages such as Telugu. Telugu is complex Morphological Language. There are many research papers published on Authorship Attribution for various languages like English, Arabic, Dutch, Chinese and Greek with various corpus sizes. Many features such Lexical, Syntactic and Structural features and their combinations are also experimented for feature extraction for various languages text. The features which are more suitable for one language text
may not be suitable for other languages text. Feature selection measures are also explored for dimensionality reduction of the training and test set. Various classifiers and their combinations are also tested for different texts developed on various languages.

Data Compression Techniques are also thoroughly tested for Authorship Attribution on various languages text. It is required to identify the influence of Machine Learning techniques on Authorship Attribution for Telugu Text. It is required to evaluate the influence of character and word n-grams features for Telugu Text. The influence of Lexical, Syntactic and Structural features and their combination for Telugu Text and also there is a need to study the impact of Data Compression techniques and Data Compression distance measures on Telugu Text.

1.10 Structure of the Thesis:

The thesis is organized in Eight Chapters:

- Chapter 1 deals with the problem of Authorship Attribution and its significance in Indian context. The concept of Text Mining with respect to Authorship Attribution. Various features of Text Classification and Data Compression techniques towards Authorship Attribution are presented. The applications, motivation to attempt the problem and organization of thesis are also presented.

- In chapter 2, the detailed discuss is presented on Text Categorization techniques, Authorship Attribution techniques, Features, Classification Techniques such as Machine Learning techniques, Statistical Techniques and Data Compression techniques.

- In chapter 3, the characteristics of Indian languages in the context of Telugu language, the process of data collection, flow of preprocessing of a Telugu text such as Normalization, Stemming, Tagging and Stop Word Elimination are described with a suitable example.
In Chapter 4 describes about influence of different Machine Learning techniques such as Decision Trees, K-Nearest Neighbors, Naive Bayes and Support Vector Machines on Authorship Attribution for Telugu Text. The different evaluation measures are also addressed in this chapter.

In Chapter 5 Influence of number of authors in the data set, influence of number of documents in the data set and feature extraction and feature selection measure such chi-square also addressed.

In chapter 6 Influence of Lexical, Syntactic and Structural features and their combination on Authorship Attribution for Telugu Text are empirically evaluated. The evaluation measures such precision, recall; F1 measure and accuracy are used to measure the impact of features on Authorship Attribution.

Chapter 7 explains about the influence of Data Compression Techniques on Authorship Attribution for Telugu Text, various Compression Techniques, various Compression Distance measures and micro and macro F1 measures on Telugu Text are evaluated.

Finally, Chapter 8 gives salient features of the work, the Conclusions of the proposed work from the results obtained from the previous chapters and also possible further extensions that can be done on the proposed work.