# Gender Identification From SMS Text Message Using Machine Learning

Pavithra C P
M.Tech Computational Linguistics

## Acknowledgement

First and foremost I wish to express my wholehearted indebtedness to God Almighty for his gracious constant care and blessings showered over me for the successful completion of the Mini Project.

I am extremely grateful to Mrs. Shibily Joseph, My Guide and Asst. Professor, Department of Computer Science and Engineering, Govt Engineering College Palakkad, for her sincere guidance, inspiration and support throughout the Mini Project.

I am thankful to Dr. Rafeeq P C, Mini Project Co-ordinator, Head of the Department,, Department of Computer Science and Engineering, Govt Engineering College Palakkad, for his support and inspiration through out the Mini Project.

Gratitude is extended to all teaching and non teaching staffs of Department of Computer Science and Engineering, Govt. Engineering College Palakkad for the sincere directions imparted and the cooperation in connection with the Mini Project.

I am also thankful to my parents for the support given in connection with the Mini Project. Gratitude is extended to all well-wishers and my friends who supported me to complete the Mini Project in time.

# Table of Contents

# List of Figures

# List of Tables

# Abstract

Short message service (SMS) has become a very popular medium for communication due to its convenience and low cost. SMS messages can also be used as a communication channel for gangs, terrorists, and drug dealers. The pervasive use of SMS is increasing the amount of digital evidence available on cellular phones. SMS text messages have unusual characteristics making it hard or impossible to apply traditional stylometric techniques, such as frequency counts and word similarity. Here proposes a new combined method for authorship classification of gender of SMS text messages, which combines machine learning algorithms with text processing features to increase the prediction accuracy of the author gender classification. It mainly consist of 4 modules: data collection, text processing, validation and gender identification. For which, SMS corpus was downloaded from the National University of Singapore website the corpus contained 55,585 SMS text messages written in English. Each of the data subsets were evaluated using the Naive Bayes and SVM algorithms. In these approach SVM performs better than Naive bayes algorithm.

**Keywords:** Text Messages, Gender Identification, SMS Corpus, Support Vector Machine(SVM), Naive Bayes.

# CHAPTER 1

# Introduction

Text is still the most prevalent Internet media type. Examples of this include popular social networking applications such as Twitter, Craigslist, Facebook, etc. Other web applications such as e-mail, blog, chat rooms, etc. are also mostly text based. A question addressed in this paper that deals with text based Internet forensics is the following: given a short text document, can identify if the author is a man or a woman? This question is motivated by recent events where people faked their gender on the Internet. Note that this is different from the authorship attribution problem.

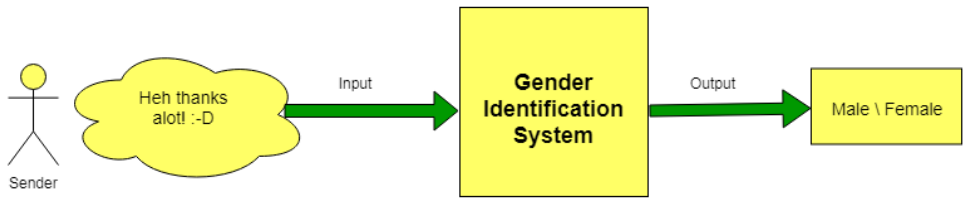


Figure 1.1: Input Output Flow Of Proposed System

The rapid growth of the Internet has created myriad ways to share information across time and space. Online social networking (such as Twitter, Myspace, Facebook), e-commerce (such as eBay, Craiglist), usenet newsgroups, etc. are gaining more prominence. As of October 2009, the number of the Internet users is estimated to be 1.69 billion according to statistics evaluated by AMD (50x50). However, this growth also encourages various kinds of misuses. Online communities are vulnerable to deceptive attacks, receiving false information, etc. The 2008 Annual Report of Internet Crime Complaint Center (IC3) states that there was a 33.1mitigate this

situation, homeland security and law enforcement agencies have launched projects to prevent deceptive attacks and track the identities of senders to protect against terrorism, child predators, etc.

In these approach investigation was done to author gender identification for short length, multi-genre, content-free text, such as the ones found in many Internet applications. Fundamental questions asked are: do men and women inherently use different classes of language styles? If this is true, what are good linguistic features that indicate gender?
The problem addresseed in this paper is author gender identification from short Internet text is different from other types of authorship identification/attribution problems, due to the following:

- gender identification is a higher level of abstraction; unlike authorship attribution the candidate set of authors is unavailable a priori.
- the length of Internet text messages is usually small compared to traditional text documents such as books for which authorship attribution is mostly studied
- unlike traditional text documents special linguistic elements such as emoticons often appear in Internet texts
- the format or the structure of Internet texts may vary among different users and situations due to real-time constraints such as Internet chat, instant messaging, etc.

SMS text messages often contain words that are abbreviated or written representations of the sounds and compressions of what the intended reader should perceive of the message (e.g. kt instead of Katie). Emoticons, such as :-( (representing a frown) and :-) (representing a smile), are a representation of facial expressions or body language, which would otherwise be missing from non-face-to-face communication. Senders of SMS text messages also tend to us different phonetic spellings

in order to create different types of verbal effects in their messages, such as hehe for laughter and muaha for an evil laughter. Letters and numbers are also often combined for compression and convenience (e.g. See you can be texted as CU and See You Later can be texted as CUL8R). All of these facts make it difficult to use syntactic means to characterize the authorship of SMS text messages.

In these approach, comparison study was performed based on two methods : First one is based on n-gram feature and the second method is based on feature extraction technique. Both the technique was analysed with two classification algorithms: Naive bayes and Support vector machine.

## 1.1 Motivation

- Gender identification from text is relatively less explored area due to limited length of SMS text.
- Usual methods employed for gender identification doesn't employ text processing causing relatively less accuracy.
- Text processing along with machine learning algorithms can have a significant impact in identifying the gender of SMS text.

## 1.2 Thesis Outline

A small description about various features, ngram, machine learning, Support vector machine and Naive bayes are included in Chapter 2. Chapter 3 contains the literature survey. Chapter 4 describes the tools used in this project. The proposed architecture for gender identification is described in detail in Chapter 5. Chapter 6 presents the results and discussion on the results. Conclusion and future work are included in Chapter 7.

# CHAPTER 2

# Theoretical Background

## 2.1 Feature Set

What are good linguistic features that indicate gender? This is an open research problem. Based on human psychology research and extensive experimentation, five sets of gender-related features are classified, They are:

- Character based
- Word based
- Syntactic
- Structure based
- Function words

**Character-based features** include 29 stylometric features widely adopted in authorship attribution problems such as number of white-space characters, number of special characters (eg., ,), etc.

**Word-based features** include 33 statistical metrics such as vocabulary richness, Yules K measure and entropy measure. Text analysis based on these studies indicated that those individuals who benefit the most from writing tend to use relatively high rates of positive emotional words (such as love, nice, sweet), a moderate number of negative emotional words (like hurt, ugly, nasty), and an increasing number of cognitive words (like cause, know), and switch their use of pronouns from one session to another.

**Syntactic features** capture authors writing style at the sentence level. Syntactic features include regular punctuation (such as comma, colon, etc.) and multiple question/exclamation marks (???, !!!) since it is not uncommon for writers in very informal situations to use several question marks and exclamation marks to express the attitude or mood. The discriminating power of syntactic features is derived from man and womans different habits of using punctuation, for example, women tend to use more question marks.

**Structure based** features represent the way an author organizes the layout of a message. People have different habits when organizing articles. These habits, such as paragraph length and use of greetings, can be strong authorial evidence of personal writing styles. This is more prominent in online documents, which have less content information but more flexible structures or richer stylistic information.

**Function words** (or grammatical words) are words that have little lexical meanings or have ambiguous meanings, but instead serve to express grammatical relationships with other words within a sentence, or specify the attitude or mood of the speaker. Set function words as one particular subset apart from word-based features, because function words play an important role in distinguishing the personal style of different genders.

## 2.2 N-gram

An n-gram model models sequences, notably natural languages, using the statistical properties of n-grams. This idea can be traced to an experiment by Claude Shannon's work in information theory. Shannon posed the question: given a sequence of letters, what is the likelihood of the next letter? From training data, one can derive a probability distribution for the next letter given a history of size n. Some of the advantages to our approach of combining machine learning algorithms with a text processing feature such as N-gram modeling are that

- 1) N-grams are able to find the roots of common words in SMS data

- there is a high tolerance for spelling mistakes since N-gram modeling creates a character based lexicon and does not use a natural language lexicon
- it is not necessary to prepare a word list from the data beforehand.

## 2.3 Machine Learning

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves. Machine learning teaches computers to do what comes naturally to humans and animals: learn from experience. Machine learning algorithms use computational methods to learn information directly from data without relying on a predetermined equation as a model. The algorithms adaptively improve their performance as the number of samples available for learning increases. Machine learning algorithms find natural patterns in data that generate insight and help you make better decisions and predictions. Machine learning uses two types of techniques: supervised learning, which trains a model on known input and output data so that it can predict future outputs, and unsupervised learning, which finds hidden patterns or intrinsic structures in input data.

Machine learning is a core subarea of artificial intelligence. It is very unlikely that it will be able to build any kind of intelligent system capable of any of the facilities that associate with intelligence, such as language or vision, without using learning to get there. These tasks are otherwise simply too difficult to solve. Further, they would not consider a system to be truly intelligent if it were incapable of learning since learning is at the core of intelligence.

Although a subarea of AI, machine learning also intersects broadly with other fields, especially statistics, but also mathematics, physics, theoretical computer science and more

## 2.4 Support Vector Machine(SVM)

In classification tasks a discriminant machine learning technique aims at finding, based on an independent and identically distributed (iid) training dataset, a discriminant function that can correctly predict labels for newly acquired instances. Unlike generative machine learning approaches, which require computations of conditional probability distributions, a discriminant classification function takes a data point x and assigns it to one of the different classes that are a part of the classification task.

SVM is a discriminant technique, and, because it solves the convex optimization problem analytically, it always returns the same optimal hyperplane parameterin contrast to genetic algorithms (GAs) or perceptrons, both of which are widely used for classification in machine learning. For perceptrons, solutions are highly dependent on the initialization and termination criteria. SVM distinguishes itself from ANN in that it does not suffer from the classical multilocal minimathe double curse of dimensionality and overfitting. Overfitting, which happens when the machine learning model strives to achieve a zero error on all training data, is more likely to occur with machine learning approaches whose training metrics depend on variants of the sum of squares error. By minimizing the structural risk rather than the empirical risk, as in the case of ANN, SVM avoids overfitting.

SVM does not control model complexity, as ANN does, by limiting the feature set; instead, it automatically determines the model complexity by selecting the number of support vectors.

**Advantages**

1) SVMs are effective when the number of features is quite large.

2) It works effectively even if the number of features are greater than the number of samples.

3) Non-Linear data can also be classified using customized hyperplanes built by using kernel trick.

4) It is a robust model to solve prediction problems since it maximizes margin.

**Disadvantages**

1) The biggest limitation of Support Vector Machine is the choice of the kernel. The wrong choice of the kernel can lead to an increase in error percentage.

2)With a greater number of samples, it starts giving poor performances.

3)SVMs have good generalization performance but they can be extremely slow in the test phase.

4)SVMs have high algorithmic complexity and extensive memory requirements due to the use of quadratic programming.

## 2.5 Naive Bayes

The Naive Bayes Classifier technique is based on the so-called Bayesian theorem and is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods. Naive Bayes is a classification algorithm for binary (two-class) and multi-class classification problems. The technique is easiest to understand when described using binary or categorical input values. This is a very strong assumption that is most unlikely in real data, i.e. that the attributes do not interact. Nevertheless, the approach performs surprisingly well on data where this assumption does not hold.

Naive Bayes classifiers are a collection of classification algorithms based on Bayes Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other. Bayes theorem named after Rev. Thomas Bayes. It works on conditional probability. Conditional probability is the probability that something will happen, given that something else has already occurred. Using the conditional probability, they can late the probability of an event using its prior knowledge. Below is the formula for calculating the conditional probability.

$$P(H \mid E) = \frac{P(E \mid H) * P(H)}{P(E)}$$

- P(H) is the probability of hypothesis H being true. This is known as the prior

probability.

- P(E) is the probability of the evidence(regardless of the hypothesis).
- P(E |H) is the probability of the evidence given that hypothesis is true.
- P(H |E) is the probability of the hypothesis given that the evidence is there.

Naive Bayes is a kind of classifier which uses the Bayes Theorem. It predicts membership probabilities for each class such as the probability that given record or data point belongs to a particular class. The class with the highest probability is considered as the most likely class. This is also known as Maximum A Posteriori (MAP).

**Advantages**

1) Naive Bayes Algorithm is a fast, highly scalable algorithm.

2) Naive Bayes can be use for Binary and Multiclass classification. It provides different types of Naive Bayes Algorithms like GaussianNB, MultinomialNB, BernoulliNB.

3) It is a simple algorithm that depends on doing a bunch of counts.

4) Great choice for Text Classification problems. Its a popular choice for spam email classification.

5) It can be easily train on small dataset.

**Disadvantages**

1) It considers all the features to be unrelated, so it cannot learn the relationship between features. E.g., Lets say Remo is going to a part. While cloth selection for the party, Remo is looking at his cupboard. Remo likes to wear a white color shirt. In Jeans, he likes to wear a brown Jeans, But Remo doesnt like wearing a white shirt with Brown Jeans. Naive Bayes can learn individual features importance but cant determine the relationship among features.

# CHAPTER 3

# Literature Survey

Cheng et al[3] investigated author gender identification for short length, multi-genre, content-free text, such as the ones found in many Internet applications [1]. In their article, they proposed 545 psycho-linguistic and gender-preferential cues along with stylometric features to build the feature space. Their initial design began with a set of features that remained relatively constant for a large number of messages written by authors of the same gender. From there, a model was built to determine the category of a given message.

One of the datasets used in their study was a collection of all English language stories produced by Reuters journalists between August 20, 1996 and August 19, 1997. For their analysis, they classified five sets of gender-related features: (1) character-based; (2) word-based; (3) syntactic; (4) structure-based; and (5) function words. One of problems with their approach was that the initial categorization of the documents by the gender of the journalists appeared to be based solely on the perceived gender of a persons name. In todays society, there are many women baring a historically male name and vice versa. This could have possibly impacted the results of the study.

Soler and Wanner[7] propose using a small number of features that mainly depend on the structure of the texts as opposed to other approaches that depend mainly on the content of the texts and that use a huge number of features in the process of identifying the gender of an author. Their study included 83 features, which consisted of five different types of features: character-based features, word-based features, sentence-based features, dictionary-based features, and syntactic features. WEKAs Bagging variant was used for classification with REPTree (a fast decision

tree learning algorithm) as base classifier. The dataset used consisted of a collection of postings of a New York (NY) Times opinion blog, which is extremely multi-thematic. In order to randomize the dataset, a temporary column was added containing the Excel formula =RAND(). The dataset was then sorted on the column containing the randomly generated number. Each of the data subsets were evaluated in WEKA using the Nave Bayes, J48, and Multi-layer Perceptron algorithms. The data was run once with each algorithm without any text processing and then run a second time with text processing.

Argamon et al[5] proposed using two types of features for authorship profiling: content-based features and style-based features. They used Systemic Functional Linguistics to represent the taxonomies describing meaningful distinctions among various function words and parts-of-speech. The Bayesian Multinomial Regression learning algorithm was used for learning the weight vector.The dataset used for the gender profiling experiment consisted of full sets of posts by 19,320 blog authors, where the gender of the author was self-reported. A potential problem with their approach was that the gender of the authors was selfreported. It is well known that people often misrepresent themselves online in order to mask their true identity.

Shannon Silessi and Cihan Varol[1] propose SMS corpus was downloaded from the National University of Singapores website. In order to prep the data for WEKA, all double quotation characters were replaced with single quotation characters. In order to randomize the dataset, a temporary column was added containing the Excel formula =RAND().

The dataset was then sorted on the column containing the randomly generated number. Each of the data subsets were evaluated in WEKA using the Nave Bayes, J48, and Multi-layer Perceptron algorithms. The data was run once with each algorithm without any text processing and then run a second time with text processing. Some of the advantages of these approach of combining machine learning algorithms with a text processing feature such as N-gram modeling are that 1) N-grams are able to find the roots of common words in SMS data, 2) there is a high tolerance for

spelling mistakes since N-gram modeling creates a character based lexicon and does not use a natural language lexicon, and 3) it is not necessary to prepare a word list from the data beforehand.

| Sl No | Name | Author | Year | Method |
|---|---|---|---|---|
| 1 | Identifying Gender From SMS Text Messages | Shannon Silessi | 2016 | Dataset: Nus SMS Corpus<br>Feature: N-gram |
| 2 | How to use Less Features & reach better performance in gender identification | J.Soler | 2014 | Dataset: Postings of NY Times opinion blog<br>Features: Structural |
| 3 | Author Gender identification from from text | N.Cheng | 2011 | Dataset: All English stories by reuters<br>Features: Character, Word,Syntactic, Structural,Functional |
| 4 | Automatically Profiling the author of an anonymous text | S.Argamon | 2009 | Dataset: Posts by blog authors<br>Features: Content based and style based. |

Table 3.1: Comparison study- Literature Survey

# CHAPTER 4

# Tools Used

The implementation of the proposed system requires the following toolkits and packages:

## 4.1 Python 3.4

Python is an easy to learn, powerful programming language. It has efficient high-level data structures and a simple but effective approach to object-oriented programming. Pythons elegant syntax and dynamic typing, together with its interpreted nature, make it an ideal language for scripting and rapid application development in many areas on most platforms. The Python interpreter is easily extended with new functions and data types implemented in C or C++ (or other languages callable from C). Python is also suitable as an extension language for customizable applications.

## 4.2 Natural Language Tool Kit (NLTK)

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning and wrappers for industrial-strength NLP libraries. Best of all, NLTK is a free, open source, community-driven project.

### 4.3 Numpy

NumPy is the fundamental package for scientific computing with Python. It contains among other things: a powerful N-dimensional array object, sophisticated (broadcasting) functions, tools for integrating C/C++ and Fortran code and useful linear algebra, Fourier transform, and random number capabilities. Besides its obvious scientific uses, NumPy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined. This allows NumPy to seamlessly and speedily integrate with a wide variety of databases.

### 4.4 Scikit-learn

scikit-learn is a package that enables Machine Learning in Python. It provides Simple and efficient tools for data mining and data analysis. Its major advantage is that it is accessible to everybody, and reusable in various contexts. Scikit-learn is built on NumPy, SciPy, and matplotlib. It is open source and commercially usable. It have implementation of algorithms for various machine learning tasks such as clustering, classification, regression, model selection etc.

### 4.5 Pandas

Pandas is an open-source, BSD-licensed Python library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language. Python with Pandas is used in a wide range of fields including academic and commercial domains including finance, economics, Statistics, analytics, etc.

### 4.6 Pickle

The pickle module implements an algorithm for turning an arbitrary Python object into a series of bytes. This process is also called "serializing" the object. The

byte stream representing the object can then be transmitted or stored, and later reconstructed to create a new object with the same characteristics.

## 4.7 Tkinter

Tkinter is Python's de-facto standard GUI (Graphical User Interface) package. It is a thin object-oriented layer on top of Tcl/Tk. Tkinter is not the only GuiProgramming toolkit for Python. It is however the most commonly used one. Python 2.7 and Python 3.1 incorporate the "themed Tk" ("ttk") functionality of Tk 8.5. This allows Tk widgets to be easily themed to look like the native desktop environment in which the application is running, thereby addressing a long-standing criticism of Tk (and hence of Tkinter).

# CHAPTER 5

# Gender Identification System

## 5.1 Methodology

The proposed gender identification system is based on machine learning approaches. The architecture and the important steps for the building of the proposed system is described here.

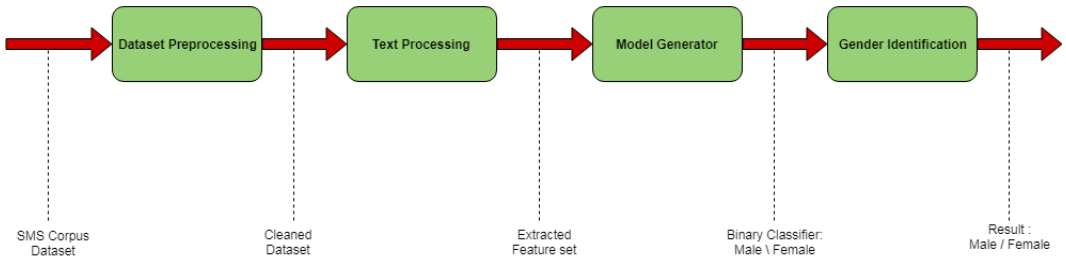


Figure 5.1: Input-Output flow in different stages

### 5.1.1 Architecture

Figure 5.1 shows an overall work flow or architecture of the proposed gender identification system from text messages. The main processes in the proposed system are :

- Dataset Collection
- Text Processing
- Model Generation
- Gender Identification

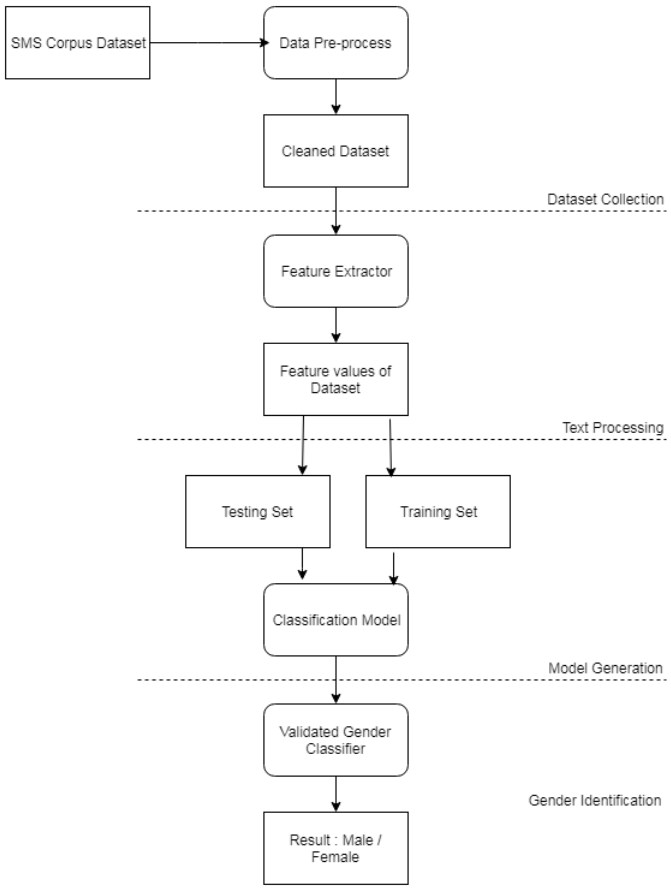


Figure 5.2: Architecture of Gender Identification System

#### 5.1.1.1 Dataset Collection

In which SMS Corpus is used as dataset which is available in National University of Singapore's website. These dataset includes 55,585 text messages. First, in order to clean up the data, all entries where the gender was unknown were removed, leaving 41,241 SMS text messages. Then, all columns except for gender and message were removed. There were several entries containing only VCARDs, so those were removed. Several entries containing a mixture of English and Chinese were removed as well. The final dataset contained 41,002 SMS messages, which consisted of 28,355 SMS text messages written by male authors and 12,647 SMS text messages written by female

authors.

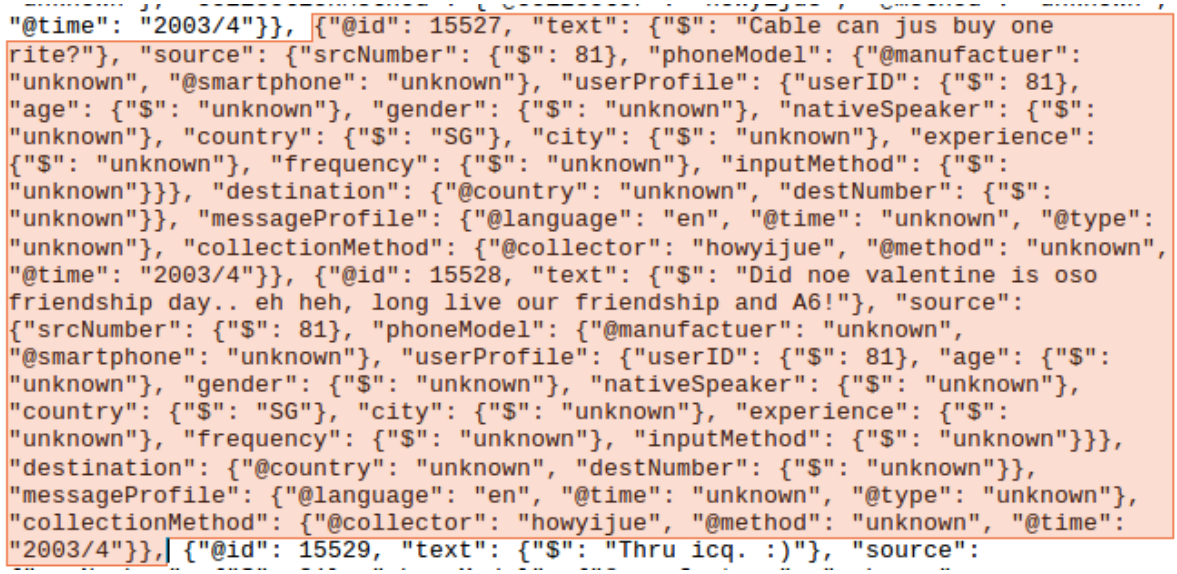


Figure 5.3: Sample of dataset

#### 5.1.1.2 Text Processing

In the second module the cleaned dataset is the input. In these approach 2 techniques are tested. In the first approach 5 types of features are selected. Total 59 features are extracted from the selected 5 types of features and in the second approach n-gram feature is only consider.

5 types of selected features are:

- Character based
- Word Based
- Syntactic based
- Structure based
- Functional Words

In character based 6 features such as *charcount, letters, upper and lower characters count etc..* is consider. In word based 4 features: *words count, long and short words count etc..*, syntactic based 10 features: *quotation, commas, exclamation mark, question mark* , structure based 4 features: *words per sentence, number of lines, number of sentence* and functional wods based 35 features: *35 POS tags* are taken.

In the second approach, n-gram is only consider as the feature. From sklearn feature extraction module TfidfVectorizer is imported. In which, n-gram range is set as 1 and 2. ie, unigram and bigram is considered as the feature.
(eg) vect = TfidfVectorizer(ngram range=(1,2),max features=100)

| gender | charcount | letter | upper | spcl | digts | space | wrds | long | short | avg lngth | qutn | commas |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 18 | 15 | 1 | 15 | 0 | 2 | 3 | 2 | 1 | 5 | 0 | 0 |
| 0 | 3 | 2 | 1 | 2 | 0 | 0 | 1 | 0 | 1 | 3 | 0 | 0 |
| 0 | 13 | 10 | 1 | 10 | 0 | 2 | 3 | 0 | 2 | 3 | 0 | 0 |
| 0 | 16 | 11 | 1 | 11 | 0 | 2 | 3 | 0 | 0 | 4 | 0 | 0 |
| 0 | 14 | 9 | 1 | 9 | 0 | 1 | 2 | 1 | 0 | 6 | 0 | 0 |
| 0 | 27 | 19 | 1 | 19 | 0 | 4 | 5 | 2 | 2 | 4 | 1 | 0 |
| 0 | 19 | 15 | 1 | 15 | 0 | 2 | 3 | 2 | 1 | 5 | 0 | 0 |
| 0 | 12 | 11 | 1 | 11 | 0 | 1 | 2 | 1 | 1 | 5 | 0 | 0 |
| 0 | 16 | 13 | 1 | 13 | 0 | 2 | 3 | 1 | 1 | 4 | 0 | 0 |
| 0 | 8 | 7 | 1 | 7 | 0 | 1 | 2 | 0 | 1 | 3 | 0 | 0 |
| 0 | 14 | 12 | 1 | 12 | 0 | 2 | 3 | 0 | 1 | 4 | 0 | 0 |
| 0 | 14 | 12 | 1 | 12 | 0 | 2 | 3 | 0 | 1 | 4 | 0 | 0 |
| 0 | 9 | 8 | 1 | 8 | 0 | 1 | 2 | 0 | 1 | 4 | 0 | 0 |
| 0 | 14 | 12 | 1 | 12 | 0 | 2 | 3 | 0 | 1 | 4 | 0 | 0 |
| 0 | 11 | 10 | 1 | 10 | 0 | 1 | 2 | 1 | 0 | 5 | 0 | 0 |
| 0 | 5 | 4 | 1 | 4 | 0 | 0 | 1 | 0 | 0 | 5 | 0 | 0 |
| 0 | 16 | 13 | 1 | 13 | 0 | 2 | 3 | 2 | 1 | 4 | 0 | 0 |
| 0 | 16 | 13 | 1 | 13 | 0 | 2 | 3 | 2 | 1 | 4 | 0 | 0 |
| 0 | 10 | 9 | 1 | 9 | 0 | 1 | 2 | 1 | 1 | 4 | 0 | 0 |

Figure 5.4: Feature Set

#### 5.1.1.3 Model Generation

In third stage classification model is created. Feature set is the input to these stage and output is a binary classifier. Two classification algorithms are used.
1) Naive Bayes
2) Support Vector Machine

The feature set is divided into two for testing and training. $\frac{1}{3rd}$ portion is taken for testing and the rest is taken for training.

#### 5.1.1.4 Gender Identification

From the selected features the system is trained using both algorithms and tested. By these way, two algorithms are compared, and identify which algorithm is best for gender identification.

# CHAPTER 6

# Results and Discussion

The classifier is trained using the different combination of features and the resultant system is compared with other systems. In these approach comparison of different approach for gender identification is performed.

1) Based on n-gram feature

2) Based On Manually extracted 5-set of features

Identify which method is better and from the 5 set of features which feature is more important to predict the gender is also identified.

| Model | Algorithm | Accuracy | Precision | Recall | F-Measure |
|---|---|---|---|---|---|
| N-Gram Model | SVM | 71.07 | 0.98 | 0.71 | 0.82 |
| | NB | 64.52 | 0.69 | 0.77 | 0.73 |
| Feature Extracted Model | SVM | 77.75 | 0.96 | 0.77 | 0.85 |
| | NB | 68.11 | 0.72 | 0.79 | 0.75 |

Table 6.1: Comparison of 2-Approaches

Table 6.1 shows the comparison of two approach based on two different algorithms, in terms of Precision (P), Recall (R) and F-score (F) and Accuracy (A). From the table it is clear that manual extraction of feature model performs better than n-gram model. And feature extracted model is performed better using support vector machine algorithm. In both methods SVM is better. So from which can conclude that SVM classification algorithm performs good for gender identification.

So from the Table 6.1 the best approach is identified, then each feature set is trained seperately to identify which feature is more imported in gender identification.

| Feature Set Included | No.of Features | Accuracy |
|---|---|---|
| Character | 6 | 71.74 |
| Word | 4 | 69.69 |
| Syntactic | 10 | 76.48 |
| Structure | 4 | 68.88 |
| Functional | 35 | 71.75 |

Table 6.2: Comparison- Only one Feature set

First trained using one feature set. Training using syntactic feature set performs well than other feature set in terms of accuracy. Table 6.2 shows the accuracy of training using one feature set.

| Combination of Feature Set | No.of Features | Accuracy |
|---|---|---|
| Character, Word | 10 | 71.69 |
| Character, Syntactic | 16 | 76.59 |
| Character, Structure | 10 | 72.23 |
| Character, Functional | 41 | 72.67 |
| Word, Syntactic | 14 | 75.82 |
| Word, Structure | 8 | 70.22 |
| Word, Functional | 39 | 72.02 |
| Syntactic, Structure | 14 | 77.02 |
| Syntactic, Functional | 45 | 76.66 |
| Structure, Functional | 39 | 72.74 |

Table 6.3: Comparison- Two feature set

Next trained using two feature set. From the combination of features syntactic and structural combination performs better. Table 6.3 shows the accuracy of training using two feature set.

| Combination of Feature Set | No.of Features | Accuracy |
|---|---|---|
| Character, Word, Syntactic | 20 | 76.31 |
| Character, Word, Structure | 14 | 72.31 |
| Character, Word, Functional | 45 | 72.73 |
| Character, Syntactic, Structure | 20 | 76.92 |
| Character, Syntactic, Functioanl | 51 | 76.86 |
| Character, Structure, Functioanl | 45 | 73.05 |
| Word, Syntactic, Structure | 18 | 76.24 |
| Word, Syntactic, Functioanl | 49 | 76.91 |
| Word, Structure, Functioanl | 43 | 73.07 |
| Syntactic, Structure, Functioanl | 49 | 77.36 |

Table 6.4: Comparison- Three feature set

| Combination of Feature Set | No.of Features | Accuracy | Not Included |
|---|---|---|---|
| Character, Word, Syntactic, Structural | 24 | 76.74 | Functional |
| Character, Word, Syntactic, Functioanl | 55 | 76.81 | Structural |
| Character, Word, Structural, Functioanl | 49 | 73.23 | Syntactic |
| Character, Syntactic, Structure, Functioanl | 55 | 77.64 | Word |
| Word, Syntactic, Structure, Functioanl | 53 | 77.64 | Character |

Table 6.5: Comparison- Four feature set

After that using 3 features and 4 features training was performed. In both cases the feature set with syntactic feature performs more than any other combination of features.

## 6.1 Screenshots

Some of the screenshots of the output of the proposed system are shown here.


Figure 6.1: GUI of Proposed System


Figure 6.2: Entering text message to predict gender


Figure 6.3: Predicting Gender

# CHAPTER 7

# Conclusion and Future Work

## 7.1 Conclusion

The proposed system detects gender from the short text messages. Other previous approaches for authorship classification used either N-gram modeling or various classification techniques. Majority of previous experiments were also conducted on larger text documents. In these approach comparison study was performed based on two methods : First one is based on n-gram feature and the second method is based on feature extraction technique. Both the technique was analysed with two classification algorithms: Naive bayes and Support vector machine. The SVM outperforms the Naive Bayes model with an accuracy of 77.75%. A study on the important features for identifying the gender shows that for a combination of 2 features , syntactic and structural features gives the most accuracy of 77.02% . For a 3 feature set combination , syntactic , structural and functional features gives the most accuracy of 77.36%. For a 4 feature combination , word , syntactic ,structural and functional features give the most accuracy with 77.64%. It is evident from the study that syntactic feature is most essential in identifying the gender.

## 7.2 Future Scope

The proposed system can be tested on variety of datasets to understand how the system works on different input. Further-more techniques like can be applied such as Deep-Learning to improve the performance of the system. There are applications like digital forensics to which the proposed system can extend.

# Bibliography


[**1**] Shannon Silessi, Cihan Varol, Murat Karabatak. Identifying Gender From SMS Text Messages. *15th IEEE International Conference on Machine Learning and Applications 2016.*

[**2**] A. Orebaugh and J. Allnutt. Data Mining Instant Messaging Communications to Perform Author Identification for Cybercrime Investigations. *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. Volume 31, pp. 99-110, 2009.*

[**3**] N. Cheng et al. Author gender identification from text. *Digital Investigation Volume 8, Issue 1, pp. 78-88, Jule 2011.*

[**4**] M. Rafi. (2008). SMS Text Analysis: Language, Gender and Current Practice.s *Online Journal of TESOL France. http://www.tesol- france.org/Documents/ Colloque07/SMSTextAnalysisLanguageGenderandCurrent Practice.pdf.*

[**5**] S. Argamon et al. Automatically profiling the author of an anonymous text. *Communications of the ACM - Inspiring Women in Computing Volume 52, Issue 2, pp. 119-123, February 2009.*

[**6**] M. Chen et al. Short text classification improved by learning multi- granularity topics. *Twenty-Second international joint conference on Artificial Intelligence (IJCAI'11), Toby Walsh (Ed.), Vol. Volume Three. AAAI Press 1776-1781.*

[**7**] J. Soler and L. Wanner. How to Use Less Features and Reach Better Performance in Author Gender Identification. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC), Reykjavik, Iceland, 2014, pp. 1315-1319.*

[**8**] R. Ragel et al. Authorship Detection of SMS Messages Using Unigrams. *8th*

*IEEE International Conference on Industrial and Information Systems (ICIIS), Kandy, Sri Lanka, 2013, pp.387-392.*

[**9**] Cheng N, Cheng X, Chandramouli R, Subbalakshmi K.P. Gender identification from e-mails. *In IEEE Symposium on computational intelligence and data mining proceedings, 2009, pp. 154e158.*

[**10**] Mulac A. The gender-linked language effect: do language differences really make a difference?; 1998.

[**11**] Z. Miller et al. Gender Prediction on Twitter Using Stream Algorithms with N-Gram Character Features. *International Journal of Intelligence Science Volume 2, pp. 143-148, 2012.*

[**12**] A. Sun. 2012. Short text classification using very few words. *35th International ACM SIGIR Conference on Research and development in information retrieval (SIGIR '12). New York, NY, USA, pp. 1145-1146.*