Author Identification

Adeela Huma
Department of Computer Science, Virginia Tech
ahuma@vt.edu

***Abstract*—Author identification is a problem of significant importance not only from academic or historic point of view as in cases of disputed authorship of some literary works but in more current and sinister affairs of forensic nature as well. To solve the problem various methodologies have been employed or invented, belonging either to statistic-dedicated computations or machine learning algorithms. This paper investigates the application of Support Vector Machines (SVM) to distinguish the authors of sonnets. The paper also aims to extract features that are computationally efficient.**

***Index Terms*— Machine learning, author identification, stylometry, author attribution, support vector machines (SVM), word cloud, term frequency–inverse document frequency**

I. Introduction

THE authors have distinct linguistic style of writing. Stylometry is the application of the study of linguistic style, usually to written language, but it has successfully been applied to music and to fine-art paintings as well. Stylometry denotes quantitative analysis of some written text that yields information about the style it is composed with and through that about the author of this text. Thus as the main stylometric tasks belonging to written text there are considered author characterization, similarity detection, and author identification.

Author characterization brings conclusions about the author, such as gender, education, social background etc. Similarity detection involves comparing texts of several authors in order to find, if they exist, some properties in common. Author identification (or attribution) means attributing an unknown text to a writer basing on some feature characteristic or measure. It can be used when several people claim to have written some text or when no one is able or willing to identify the real author of this text.

Stylometry is most often used for detection of plagiarism, finding authors of anonymously published texts, for disputed authorship of literature or in criminal investigations within forensic linguistic domain.

Author identification is the task of identifying the author of a given text. It can be considered as a typical classification problem, where a set of documents with known authors is used for training and the aim is to automatically determine the corresponding author of an anonymous text. In contrast to other classification tasks, it is not clear which features of a text should be used to classify an author. Consequently, the main concern of computer-based author identification is to define an appropriate characterization of documents that captures the writing style of authors.

Two critical issues of the stylometric analysis are: selection of descriptors that characterize texts and authors, and analytical techniques applied to the task.

The typical textual analysis procedure (invariant of particular methodology employed) starts with training during which there are used texts of known authors for whom there are computed characteristics of selected features, then follows the stage of verification when for unattributed texts there are obtained the same descriptors to be compared with previously calculated results. Then from the available set of possible authors there is chosen the one that matches most closely.

Features selected in stylometric methods must constitute the writer invariant (called also authorial or author's invariant), a property of a text which is invariant of its author, that is it is similar in all texts of this author and different in texts of different authors. It is generally agreed that writer invariants exist yet establishing what properties of a text should be used is the question that stands open.

In this paper supervised learning using support vector machine is presented to identify authorship. Texts studied for classification are literary works of writers, William Shakespeare [4], Wallace Irwin [5], Elizabeth Barrett Browning [6], William M. MacKeracher [7]. The paper proposes to use lexical features of text that show promising results when used as writer invariants because they are used rather subconsciously and reflect the individual writing style which is difficult to be copied.

II. Background

Author identification has a long history that includes some famous disputed authorship cases and also has forensic applications. There are various author identification techniques used such as counting features such as word length, the use of sentence lengths to judge authorship, syllables per word, a function to relate the frequency of used words and the text length, the style marker of the text, the syntactic and lexical style markers, the proportion of the different word count to the total word count could be a fair measurement.

Manuscript received December 14, 2015. This paper is final report for Adeela Huma is MS student at Virginia Tech. Email: ahuma@vt.edu.

Another related class of techniques that have been applied is machine learning algorithms categorization and other stylistic discrimination tasks. Often, studies have relied on intuitive evaluation of results, based on visual inspection of scatter plots and cluster analysis trees, though recent work has begun to apply somewhat more rigorous tests of statistical significance and cross validation accuracy. Other stylometric features that have been applied include various measures of vocabulary richness and lexical repetition. Most such measures, however, are strongly dependent on the length of the text being studied, and so are difficult to apply reliably. Many other types of features have been applied, including word class frequencies, syntactic analysis, word collocations, grammatical errors, and word, sentence, clause, and paragraph lengths. Many studies combine features of different types using multivariate analysis techniques.

Usually analytical techniques applied to stylometric tasks employ either statistic or machine learning approaches. Statistical computations are used in Markovian Models (MM), Principal Component and Linear Discriminant Analysis (PCA and LDA), cluster analysis, Cumulative Sum (CUSUM or QSUM), while machine learning involves application of Artificial Neural Networks (ANN), Genetic Algorithms (GA), Support Vector Machines (SVM), Rough Set Theory (RST), decision trees, Naïve Baye's Classifier and other methods.

III. Objective

The primary aim of author distinction is to remove uncertainty about the author of some text, which can be used in literary tasks of textual analysis for works edited, translated, with disputed authorship or anonymous, but also with forensic aspect in view to detect plagiarism, forgery of the whole document or its constituent parts, verify ransom notes, etc.

Author identification analysts claim that each writer possesses some unique characteristic, called the authorial or writer invariant, that keeps constant for all texts written by this author and perceivably different for texts of other authors. To find writer invariants there are used style markers that are based on textual properties belonging to either of four categories: lexical, syntactic, structural, and content-specific.

Lexical descriptors provide statistics of total number of words or characters, average number of words per sentence, characters per sentence or characters per word, frequency of usage for individual letters or distribution of word length.

Syntactic features reflect the structure of sentences, which can be simple or complex, or conditional, built with punctuation marks. Structural attributes express the organization of text into paragraphs, headings, signatures, embedded drawings or pictures, and also special font types or its formatting that go with layout.

Content-specific properties recognize some keywords: words of special meaning or significant importance for the given context.

Unfortunately, the convenience of using contemporary word editors and processors works against preserving individual author styles due to its available options of "copy and paste". It makes imitation of somebody else's style much easier and that is why modern author identification techniques aim at exploiting the computational powers of computers to analyze patterns within subconsciously used common parts of speech, as opposed to historical approaches that emphasized some rare standing out elements of a text which could be noticed by virtually anybody and thus likely to be faked.

IV. Dataset and Tools Used

The dataset used for training and testing are literary works (sonnets) of below authors, courtesy of Project Gutenberg (www.gutenberg.org)[1].

- William Shakespeare [4]
- Wallace Irwin [5]
- Elizabeth Barrett Browning [6]
- William M. MacKeracher [7]

A **sonnet** is a poetic form, which originated in Italy. A sonnet usually consists of fourteen lines.

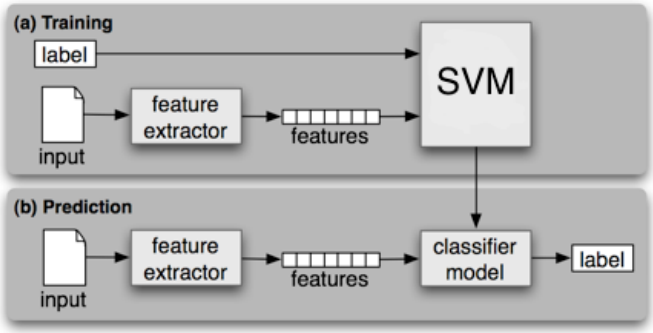
R –language is used for text classification. The source code for the experiments done can be found at below link:

https://www.dropbox.com/sh/h01az8g7ein45qx/AACX1efT3d6DjnmU9aP9G7W8a?dl=0

V. Architecture

Technically text classification is the task of choosing the correct class label for a given input. Following is the framework used for text classification based on supervised learning approach using SVM (Support Vector Machines).

Input is documents/corpus that's a sonnet in our study. Features are extracted using "tf-idf" approach explained below.



VI. Text Analysis / Feature Extraction

The text used for classification is sonnets of William

Shakespeare [4] and other authors [5][6][7].

Text processing is done using **tm** (text mining) package in R- language. This process includes:

- deleting preamble text in each of the documents
- converting all text to lower case
- stop word removal
- punctuation removal
- removing whitespaces

Since text in sonnets contains of a lot of words that do not contribute in identifying authors style such as stop words etc., it's preferable to delete them to avoid curse of dimensionality. After pre-processing the remaining text is converted to Document Term Matrix (DTM).

A **document-term matrix or term-document matrix** is a mathematical matrix that describes the frequency of terms that occur in a collection of documents. In a document-term matrix, rows correspond to documents in the collection and columns correspond to terms. There are various schemes for determining the value that each entry in the matrix should take e.g. term-frequency, tf-idf etc.

For instance if one has the following two (short) documents:
D1 = "I like computers"
D2 = "I hate computers",

Then the document-term matrix would be:

	I	like	hate	computers
D1	1	1	0	1
D2	1	0	1	1

which shows which documents contain which terms and how many times they appear.

To computer the terms in Document Term matrix, **tf-idf** [8][9][10] approach is used. tf–idf, short for t**erm frequency–inverse document frequency**, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus/documents. The tf-idf value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general.

Typically, the tf-idf weight is composed by two terms:

- the first computes the normalized Term Frequency (TF), aka. the number of times a word appears in a document, divided by the total number of words in that document;
- the second term is the Inverse Document Frequency (IDF), computed as the logarithm of the number of the documents in the corpus divided by the number of documents where the specific term appears.

Feature Selection

In the dataset used we have 193 documents, each representing a sonnet. 124 belong to Shakespeare and remaining to a mixture of authors. The only source of data looked was project Gutenberg [1] as it provides license to use free books.

W. Shakespeare Sonnets [4]	Other Author Sonnets [5][6][7].	Total Sonnets
124	69	193

In text classification domain each word is a feature: whether it is present in the document or not or how many times it appears.

After pre-processing, the resulting Document Term Matrix has 4606 terms/words. This means 4606 features, which in turns means very high dimensional data. Most of the matrix has sparse terms. The sparse terms is not a good choice for feature, it merely adds noise. After removing 70% of the sparse terms from the matrix, we get following 4 terms:

"love" "thee" "thou" "thy"

The Terms obtained after removing sparse terms can be used as features to classify Shakespeare's sonnets from the other authors.

Feature Selection Verification

To verify if feature terms obtained from tf-idf approach are good enough for classification or not, some observation is made on documents using:

- **Word-frequency plot** - displays the frequency of words in documents.
- **Word cloud** - Word-cloud or tag cloud is a visual representation of text data. Tags are usually single words and the importance of each tag is shown with font size or color. This format is useful for quickly perceiving the most prominent terms.
- **K-mean clustering of terms** - to cluster most used terms in documents.

Below is the frequency plot for Shakespeare sonnets and it can be seen that words "thy, thou, thee, love" has higher frequency

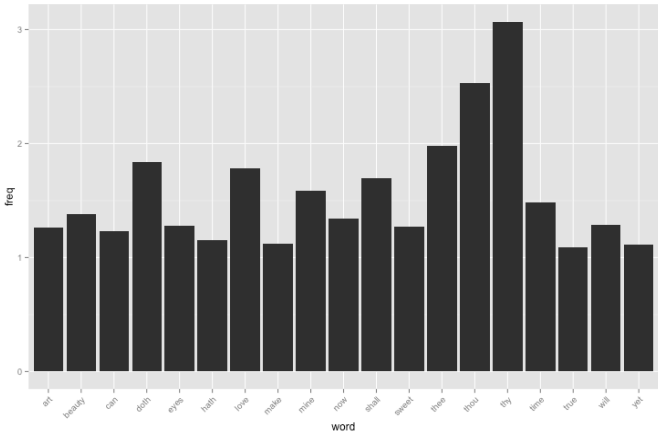


Figure 1: Word frequency plot of Shakespeare sonnets

Below is the word-cloud for Shakespeare sonnets and it can be seen that words “thy, thou” are most frequent terms.

Figure 2: Word cloud of Shakespeare sonnets

Below is the cluster plot for terms used in Shakespeare sonnets and it can be seen that words “thy, thou” are clustered together.

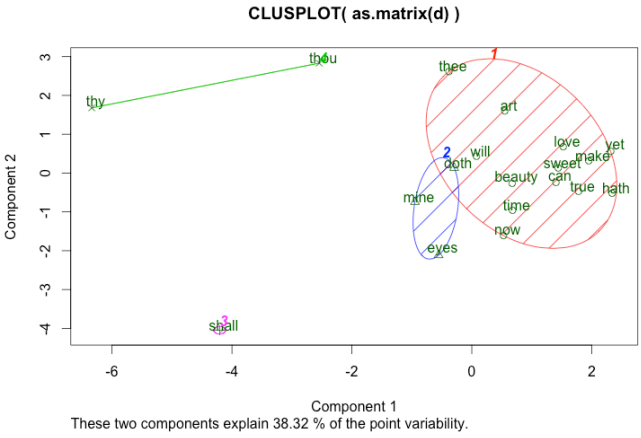


Figure 3: K-mean clustering plot of Shakespeare sonnets

Based on these observations, "love" "thee" "thou" "thy" seems to be correct feature terms for Shakespeare sonnets.

Lets also analyze the sonnets from other authors.

From below frequency plot, it can be seen that similar terms seems to be frequent in other author’s sonnet but their frequency is not as high as in Shakespeare’s sonnets.

Similarly as obvious from word cloud, word “thy” is less prominent in word cloud of other authors.

Cluster plot also show a lot of distance between most frequent word “love” and other words “thou”, “thy” etc.

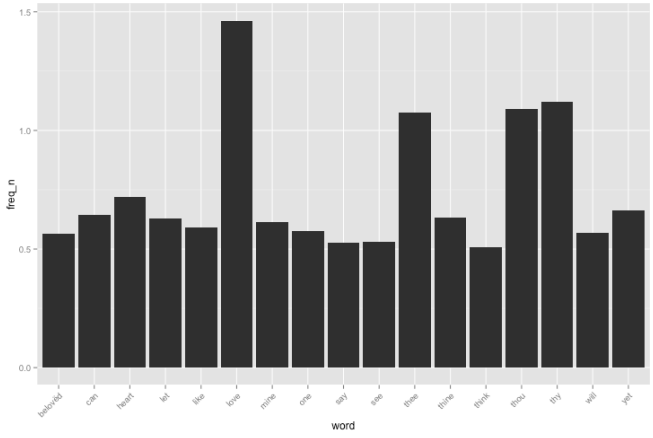


Figure 4: Word frequency plot for other Authors sonnets

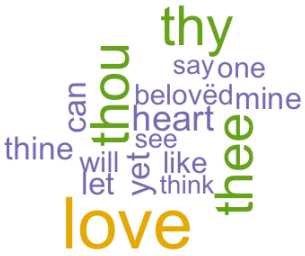


Figure 5: Word cloud for other Author sonnets

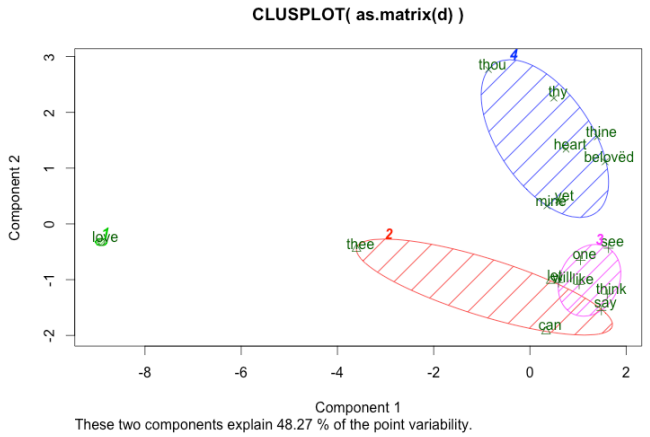


Figure 6: K-mean clustering plot for other Author sonnets

VII. Classification Methodology & Results

After feature extraction Support Vector machines (SVM) is used for binary classification of sonnets using linear kernel function. Linear kernel is suggested for text classification. For binary classification, we have two class labels:

- no-shakespere(-1)
- shakespere (1)

Before data processing each sonnet is assigned a class label. Using **tm** module of R-language we get Document term matrix (DTM - a list of 193 entries) and then DTM is converted to vectors.

For classification svm is trained on 70% data and remaining 30% is used for testing.

Training is done on 70% of randomly chosen vectors using "radial" kernel with cost =1. Following are the result of prediction:

Confusion Matrix:

	no-shakespere	shakespere
no-shakespere	2	6
shakespere	15	35

SVM model's accuracy using radial kernel is: **63.7931%**

Same step is repeated with "linear" kernel and following is the results:

Confusion Matrix:

	no-shakespere	shakespere
no-shakespere	0	0
shakespere	17	41

SVM model's accuracy using linear kernel is: **70.68966%**

A confusion matrix is a table where each cell [i,j] indicates how often label j was predicted when the correct label was i.

Cross Validation Result

Also performed 10 fold cross validation on training data and each time accuracy is improved. Below is summary of **cross-validation from R-console**:

```
svm(formula = ytrain ~ ., data = train[-5], type = "C-classification", kernel = "linear", cost = 1,
    cross = 10)

Parameters:
  SVM-Type:  C-classification
  SVM-Kernel:  linear
       cost:  1
      gamma:  0.25
  Number of Support Vectors:  105
      ( 54 51 )
   Number of Classes:  2
   Levels:
    no-shakespere shakespere

   10-fold cross-validation on training data:

   Total Accuracy: 61.48148
   Single Accuracies:
      76.92308  57.14286  69.23077  50  69.23077  64.28571
   46.15385 64.28571 46.15385 71.42857
```

Result using all terms in documents

As a final test step, used all terms in DTM for classification and results were 100%.

	no-shakespere	shakespere
no-shakespere	18	0
shakespere	0	40

SVM model's accuracy using radial kernel is: **100%**

SVM Prediction Result on Test Data (4 terms)

Following is the graph of SVM's output on test data. X-axis represents documents (sonnets); Y-axis represents decision value (1 to -1).

"+" symbol shows the support vectors and circles shows the class labeled data. It can be seen that data is not classified 100%, there are some errors.

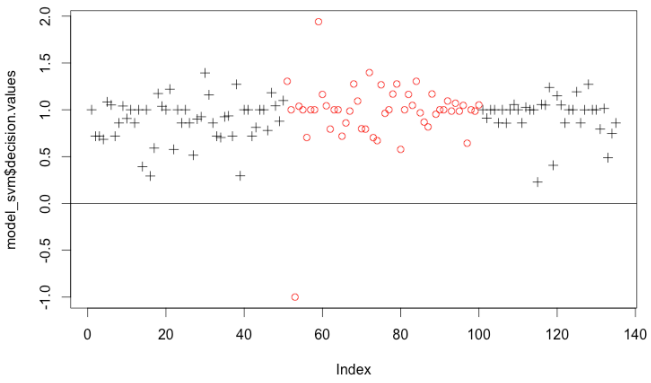


Figure 7: SVM Prediction Result (4 terms)

SVM Prediction Result on Test Data (4606 terms)

Following is the graph of SVM's output on test data. X-axis represents documents (sonnets); Y-axis represents decision value (1 to -1).
"+" symbol shows the support vectors and circles shows the class labeled data. It can be seen that data is classified 100%.

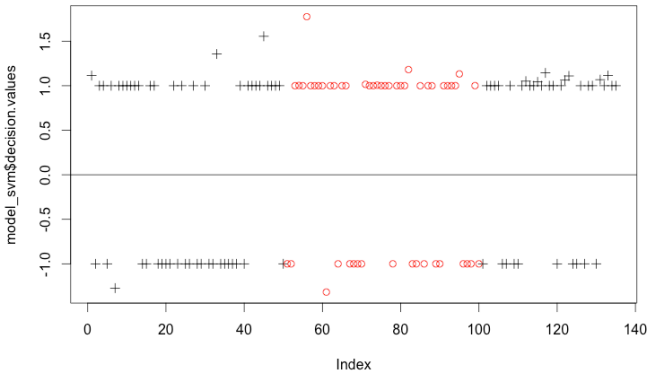


Figure 8: SVM Prediction Result (4606 terms)

VIII. Conclusion

The classifier shows almost 70% accuracy using four features extracted using "tf-idf" approach. If I use all the terms (4606) for text classification and the model took some time to train and predict results but it was 100% accurate. It can be inferred that more the feature terms used in classifier the better results it shows. This study also shows that linear kernels are best to use for text classification as they gave slightly better results than radial kernels.

References

[1] Project Gutenberg (www.gutenberg.org)
[2] Sagae, K. and Lavie, A. A Classifier-Based Parser with Linear Run-Time Complexity. Language Technologies Institute, Carnegie Mellon Unviersity. 2005. http://www.cs.cmu.edu/~alavie/papers/IWPT05-sagae.pdf
[3] Luyckx, K. and Daelemans, W. (2005). Shallow Text Analysis and Machine Learning for Authorship Attribution. In Proceedings of the Fifteenth Meeting of Computational Linguistics in the Netherlands. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.118.5550&rep=rep1 &type=pdf
[4] William Shakespeare Sonnets (http://www.gutenberg.org/ebooks/1041)
[5] The Love Sonnets of a Car Conductor by Wallace Irwin (http://www.gutenberg.org/ebooks/5332)
[6] Sonnets from the Portuguese by Elizabeth Barrett Browning (http://www.gutenberg.org/ebooks/2002)
[7] Sonnets and Other Verse by William M. MacKeracher (http://www.gutenberg.org/ebooks/37365)
[8] K. Sparck Jones. "A statistical interpretation of term specificity and its application in retrieval". Journal of Documentation, 28 (1). 1972.
[9] G. Salton and C. Buckley. "Term-weighting approaches in automatic text retrieval". Information Processing & Management, 24 (5). 1988.
[10] H. Wu and R. Luk and K. Wong and K. Kwok. "Interpreting TF-IDF term weights as making relevance decisions". ACM Transactions on Information Systems, 26 (3). 2008.

Adeela Huma received her B.Sc. Computer Engineering from UET, Lahore, Pakistan. Her research interest includes data mining.