# An Exploratory study on Authorship Verification and Learning Assessment in Online Education Systems

Priyanka Singhal

Computer Science

Indraprastha Institute of Information Technology, Delhi (IIIT-D), India

A Thesis Report submitted in partial fulfilment for the degree of

*MTech Computer Science*

29 July 2015

1. Prof. Pushpendra Singh (Thesis Adviser)

2. Prof. Vikram Goyal (Internal Examiner)

3. Dr. Kuldeep Yadav (External Examiner)

Day of the defense: 29 July 2015

Signature from Post-Graduate Committee (PGC) Chair:

# Abstract

Many in class and distance courses offered these days involve the submission of assignments online. These assignments are meant to be assessed to give students timely feedback. However, ensuring that the student who is enrolled in the course is the same who is submitting the assignment work has been a challenge. Since the submitted assignments are in a formal structure and contain some features that can be extracted and accessed automatically, there can be measures like stylometry to study the real author of the submitted work. Once, author verification is done, it is required to assess the verified documents from different aspects.

In this research, two main objectives are being covered in order to see how much actual learning is happening. Firstly, whether the student is actually participating in the course activities as that reflects how comfortable student is with the course. Secondly, how much the student can actually learn in such environments like that of distance learning. With the approach of matching n grams and global weight of terms used in the submissions, we comment on the understanding of a course by the student. Such assessments can help instructor to predict over all learning for the course and what alterations can be made to have effective outcome. A correlation on learning scores and instructor grades is done to show whether such learning measuring techniques can be used to quantify students' learning as a whole.

I dedicate my MTech Thesis to my father Vijay Kumar Singhal who has always encouraged me in all phases of life and is my greatest source of inspiration.

## Acknowledgements

I would take this wonderful oppurtunity to express my deepest gratitude to my advisor Prof. Pushpendra Singh for his continuous guidance, support, constant motivation and patience throughout my thesis. Without his guidance and support this thesis would not have been possible. I feel really blessed to have him as my thesis advisor.

Besides my advisor, I would like to thank Siddhartha Asthana for his timely advice and encouragement, but also for the hard question which incented me to widen my research from various perspectives. I would also like to thank my fellow mate Vibhor Nijhavan for his insightful comments, suggestions and constant support during the course of my thesis. I would like to thank God for all his blessings.

Finally, I would like to thank my parents and sister for their constant support, encouragement, love and trust in me.

# Declaration

This is to certify that the MTech Thesis Report titled **An Exploratory study on Authorship Verification and Learning Assessment in Online Education Systems** submitted by **Priyanka Singhal** for the partial fulfillment of the requirements for the degree of *MTech* in *Computer Science* is a record of the bonafide work carried out by her under my guidance and supervision at Indraprastha Institute of Information Technology, Delhi. This work has not been submitted anywhere else for the reward of any other degree.

**Professor Pushpendra Singh**
**Indraprastha Institute of Information Technology, New Delhi**

# Contents

# List of Figures

# List of Tables

# 1

# Introduction

With the advancement of courses being tutored from face-to-face to online, various institutions have shown interest to adapt the concept of e-learning. This way they can avoid the expense of having workforce and the associated support services for real time supervision. One such application is MOOC(Massive Open Online Course). It is an initiative taken by various universities to bring online education for large number of learners. MOOCs have been able to avail innovation, experimentation and the use of technology for the huge number of participants involved [2]. The report [2] reflects how the number of universities offering MOOCs has doubled over small period of time. It has covered almost 400 universities with offering nearly 2400 courses.

In 2014, Coursera came out as the major contributor ,collaborated with 33 other universities (including Princeton, Penn, Stanford, and the University of Michigan).There are total of 2 million students enrolled for courses offered by Courseera with 70,000 new students signing up per week. MiriadaX being the non U.S. MOOC covered 1 million registered users. The top three providers of MOOC has been Udacity, Coursera and edX. The top 3 subjects in the same year were: Humanities, Computer Science Programming, and Business Management. 80% percent of the current courses are currently being offered in English and 13 different languages contributed the rest of the percentage. As predicted in report [2, 3], MOOCs came out big in the year of 2014 and is still expanding.

MOOCs earlier had programs which were similar in their conduct as college classroom programs, with specific deadlines for graded activities. But in 2012, Udacity offered courses which were more flexible and students could complete them with their own pace[2].

There are certain research questions that were answered regarding MOOC in this report[2]: (a)How ultimate learning is happening for a student?(b)How MOOCs are more or less engaging than in class higher education, or exhibit different patterns overall? (c)Does diversity in learners regarding age, educational attainment, employment type, occupational status and country of residence affect learning.

The above questions in collaboration contribute to ultimately one question of how effectively the deliverance is happening from this side of instructor to the ultimate student at the receiving end. The concept of immediate feedback for a course can be a way to measure how successfully they understand and apply the major concepts. The well-designed online tests in between and at end of course can facilitate instructor to investigate how is students' knowledge of concepts. One other way is to give timely assignments, where there can be set of questions forcing student to enquire answers for some essential concepts.

One major point of consideration in this set up is the point of assessing the work submitted by the students. This is done to give students a timely feedback as part of a quality learning experience. Also, instructor can monitor the over all performance, which can help him to redesign the course for better results in future. One major question is ”are you assessing the right person?”.

There can be multiple forms of academic dishonesty followed by students in the examinations [4]. When we talk about in class examinations, various measures are there to effectively curtail any means of cheating. In a controlled environment, like that of in-classroom examinations, validation of candidate is done by staff member by checking his identity credentials. Other way to validate a student is by issuing a student login as an authorization code. But, the courses offered via an asynchronous, online medium can bring additional challenges in detecting instances of academic dishonesty. Specific concern in this set up even after validating student includes verifying that a student who registered and logged in for MOOC is the same person who completes the work.

For such applications, honor codes, on-site testing, and other practices such as strict surveillance under cameras, this is an area where there are no easy solutions for these difficult questions [5].

### 1.0.1 Research Question

. Aim for this research has been that whether a benchmark can be formed where different measures can be taken in consideration to comment on over all learning of student. A particular challenge was to have corpus as that of MOOC's wherein assignments are submitted online and teacher does not have any means to know the baseline where he can compare different works by the student. As discussed in paper [4], one solution to this is getting a discussion forum assignment in order to know a students style. As every assignment cannot be proctored, this teacher can have a knowledge base for testing further all assignments. In this research, in order to have similar real time scenario, summaries were collected from students for a particular research paper. Similar exercise was repeated throughout the semester and summaries were submitted online through online platform. So for each candidate, they can submit their summaries only after logging through the login details assigned to them. This ensures the validation. For authorship traditionally works has been voluminous and long texts. But in our case these are summaries which are written in an particular format. Students are asked to write answers for set of questions which broadly cover the important aspects of summary. This motivation has been built upon the vision taken by Jeorge-Botana et al., in which they have covered how latent semantic analysis can be used to assess the summaries submitted by students[6].

Going by this vision,the probable solution to have check on any means of cheating is by having a tool which can comment on their writing style in order to have authorship for their submissions. Stylometry is such field where a textual feature or multiple features are used for quantifying writing style. Similarly we can also comment on how their approach is towards writing the summary [1, 7].

This research will broadly cover these questions:

1. How effective is approach of authorship attribution to quantify how many students actually wrote the assignments given to them? How authorship can help us to conquer the question of ultimate student verification?

2. Can learning be judged on the basis of the fact that how many are real attributions for a student?

3. What are the right feature set and classifier when we have similar texts;here summaries derived from single source; here original research paper? How can similar texts affect recognising real author?

4. If students write summaries from original paper, how much actual learning is happening? Is it only copy paste happening in real scenario? How we can comment on confidence regarding writing a summary for a student?

# 2

# Related Work

L. Y. ke et al. [8] has discussed an overall approach on how summaries are being written and what major key points should be stressed in order to streamline a text into a summary. Three main points are deletion, selection and abstraction. In addition, there is a special mention on how summary writers approach for the part they have selected to mention. It can be paraphrasing or nominalising the propositions taken from the original paper or source; replicating linkages mainly mentioned in the original source; mentioning new derived linkages that do not exist at all. In order to estimate how students prioritize their time and effort, one measure that follows is assessment [9]. However, as mentioned in [4], it is very difficult to certify the authorship for the work submitted by students especially in courses offered in environments where student instructor interaction is not face-to-face but through distance or online courses. Also, the paper has briefly mentioned various forms of academic dishonesty. Addressing some of the problems of cheating in such environments, T. Lancaster [10] has explored the context behind each problem and has further suggested technical implementations of intelligent context-aware systems. But, the solution for student verification is still a problem with not much of practical solutions. In order to combat such situations, one possible solution could be to mark the author's personal style.

Fully automated approaches based on concept of authorship attribution can be useful in analysis of many online applications like email, blogs, online forum messages etc. In the survey paper [1], various approaches to quantify the writing style is presented. Two methods; individual and cumulative profile for training text of author has been discussed, each with its own application area. Author has briefly discussed on the important question of how to discriminate between authorship, genre and topic. Many features represent style, other represent context and some represent both. Variety of feature types and categorization methods that have been proposed are covered by [7]. Authorship verification problem and existing techniques to approach it has been discussed in [11, 12]. For all these techniques, size of text is an important aspect. K.Luycks et al. has rigorously researched two main parameters in supervised machine learning that can majorly be a deciding factor for performance of computational attribution. These factors are (1) the number of candidate authors (i.e. the number of classes to be learned), and (2) the amount of training data available per candidate author (i.e. the size of the training data) [13]. On similar lines, R. Ramezani [14] has concluded how different classifiers work with different set of authors.

Further, adding to prospect of assessment of summaries, G. J. Botana et al. [6] has figured out a latent semantic analysis-based automated summary assessment. The correlation between essential information derived from semantic space and human scores gave us an idea on how automated systems using this technique are at power with human assessment. This has highly motivated us to work on how such tools and methodologies can relate to over all learning. In this study, author's confidence in writing are based on the proportion of matches between trigrams in the reading text and the students summary with assumption that student with much clear idea will write more in his words than just copy-paste text. Use of trigram is based on the mention of previous researches as mentioned in [15]. Another measure used is this research is global weight in order to comment how technically student approach for writing i.e. how informative words are. This measure has been used earlier by with respect to..

# 3

# Corpus

## 3.1 Data Collection

Data was collected over a period of four months, based on the assignments allotted to students through a course delivered in college premises. Students were given research papers as reading text with average length of 500 words. and were asked to submit summary regarding the paper on the online platform. Total number of students who enrolled for course were 20. Out of which 2 left the course in between and so the data was considered regarding those 18 students. They were given assignments on regularly basis. The research papers were technical papers belonging to field of Mobile Computing, forming a homogeneous corpus in result. So in totality, the genre of entire corpus was specific for all papers. Students write summaries in English, English being their second language. All of these assignments were graded, so assumption is they will attempt it with vision of achieving more. Students include males and females both and belong to same age group.

Collection of data set retrieval has been divided into two phases. One collection is done in strict vigilance in order to have the ground truth. This makes up the 50 percent of the existing corpus and is used to build the training platform. Rest of the 50 percent includes the testing documents used for attribution in order to comment on the real authorship that happens.

## 3.2 Summary Representation

Summaries collected have specific restricted format, with length minimum of 1 page up to maximum of 2. There were a certain set of questions which were required by students to answer broadly covering all aspects of summary. Point worth noticing is that these question are asked in such a way to stress the learning. Redaction of questions was not done as they contained the style as each student approached differently while mentioning the same set of questions. Length of answers varies for all students, required they stick to format of having maximum of two pages. Size of summaries submitted by students online have an average of 550 words. They are given flexible time wherein they can submit within a gap of 3 days. We believe this flexibility in time can help student to read research paper of such length in detail and they should be able to deliver largely.

Total summaries collected were 494 in number, but after removing for those two students they were 474 in number. During the entire they were given 31 research papers in total as assignment, out of which for only 5 research papers all 18 students wrote summary. Otherwise, at maximum a student has written 28 summaries and 23 in minimum. Spelling mistakes are not done frequently and so has not been our evaluation criteria.

# 4

# Research Framework and Solution Approach

Currently, the students assignments are collected online and by many prevailing automated and semi-automated techniques submissions are accessed. However, student verification has always been a challenge in such platforms where student instructor interaction is minimum. On the same note how to comment on actual learning when the submitted work is not of the registered student (figure 4.1).

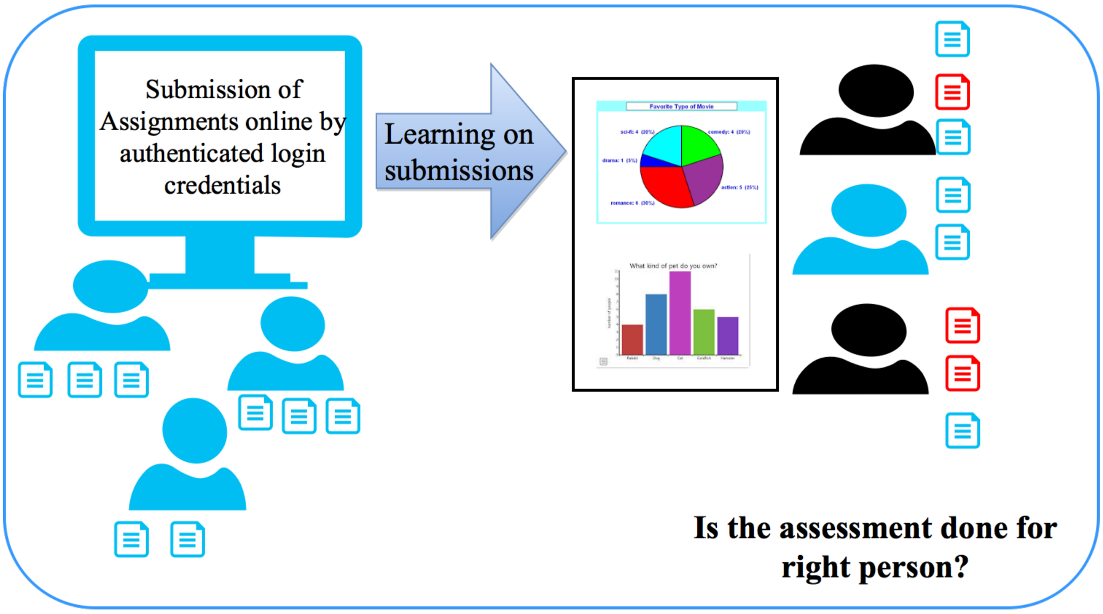

**Figure 4.1:** Current approach for assessing learning.

With the current understanding, the main prospect of this research was to see if verification can be handled carefully and later learning measures were applied on the verified document only. This can suffice our assumption that if a student is confident enough with the course, student shall write the summary by their own. With the two measures discussed later, the main objective was to see even if student is writing summary by own what is the approach they follow and how they go for summarising strategies (figure 4.2).

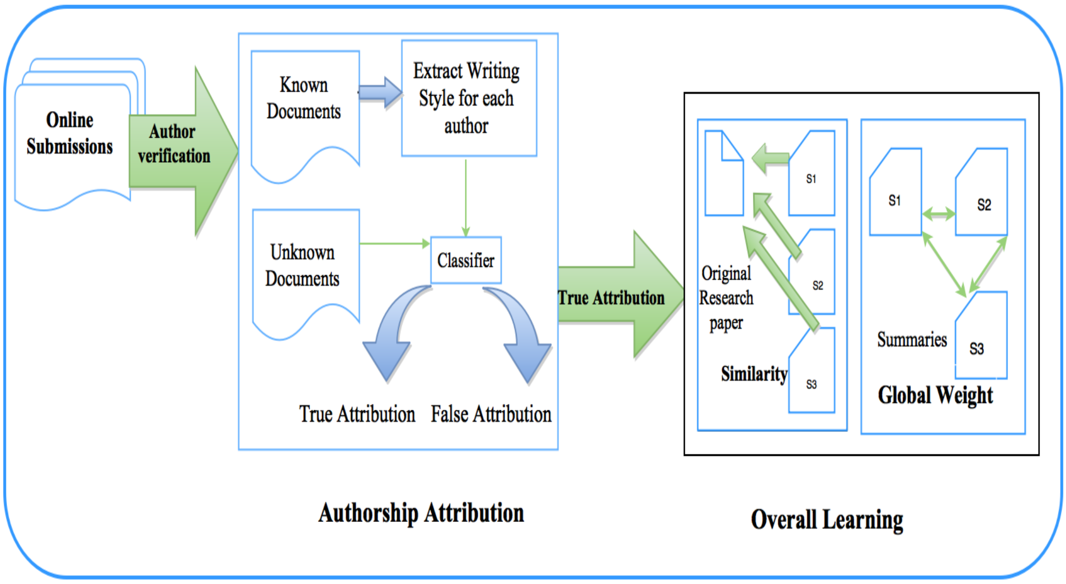

**Figure 4.2:** After attribution approach

## 4.1 Authorship Attribution

Authorship attribution is a field of stylometry,which learns the style from the samples of a predefined candidate set of authors and aims for deciding the real author. Every author has his own particular writing style. Various factors like word frequency,the frequency of characters,vocabulary richness, and the length of sentences decide the sentence construction. Authorship attribution differs from text classification as former deals with writing style and later is more content specific. Here features and classifiers are text dependant. The feature set is not prior deterministic as that of text categorization.

The basic assumption of authorship attribution is that: every author possesses some particular characteristics in his writing style, which are distinct and cannot be manipulated as this is adapted by the author unconsciously[**?** ]. Believing this, it is possible to have an automated process which can learn such patterns and can conclude authorship of the supposed candidate on the unknown texts. One of the paper [16] has talked about how writeprints can be generated and therefore can be used to identify one person from another. Popular applications of authorship are plagiarism detection, resolving disputes over authorship, as well as attributing content that is malicious.

This is a particular case of one-class problem; also called as authorship verification. Here there are closed set of candidate authors whose training data is available as ground truth and one amongst them is responsible for the test document. It is different from open class problem where the potential candidate set is open and in response for an anonymous text there can be an answer of ”none of the above”. Authorship verification is a special case of this open class where there is only one candidate whom the authorship is checked for. Either he has written it someone else has, where that someone else can be anyone in the whole world[11].

### 4.1.1 AA instance based and profile based

There are mainly two categories in which Stamatatos[1] has classified all the methods. One is profile-based where concatenation of all training texts corresponding to an author is done. This generates a profile for author and hence features are extracted from this generated profile. Further, generated features are given to attribution model to train it, and then when new disputed text comes it decides on the basis of training.
Other is instance-based, where each text is considered as an instance and further features are extracted from each instance to fed to model. Instance-based is most of the time used in contemporary authorship attribution applications. It is believed to retain more information than profile-based method. The methodology of the instance-based approach can be seen in figure

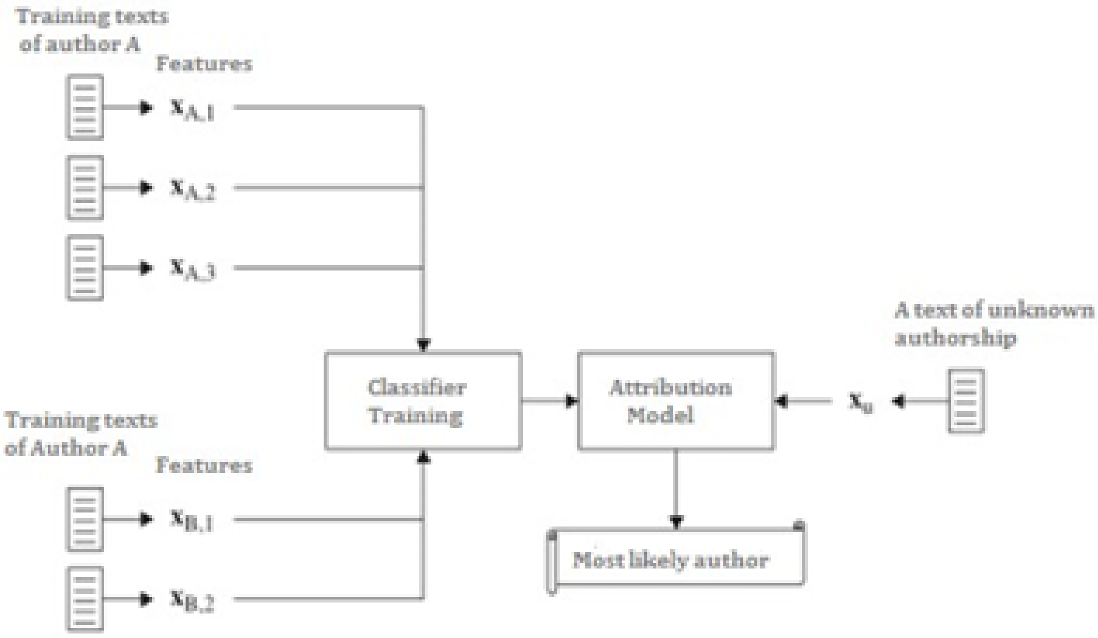

**Figure 4.3:** Instance-Based approach ([1])

### 4.1.2 Selection of feature

Finding appropriate features has been a challenging task and various studies has been performed to find most efficient feature set for a specific corpus. The so called " style markers" are used to quantify the writing style of an author. Some of the main stylometric features are: lexical features, character features, syntactic features, semantic features, application-based features, and idiosyncratic features.

**Lexical features**

Text written usually contains tokens namely word, number, punctuation etc. The usage of such token represent style and their usage depends from author to author. These features are with such advantage that they can be applied to any language and corpus. Lexical features represent style at large and does not contribute much to dimensionality when it comes to style based categorization as compared to classic text classification. Function words are topic-independent words that are not controlled and are used unconsciously by the authors [1]. This way function words are purely stylistic based. There have been many lists ranging from 150 to

675 words in researches done till now. However, in this research a list of function words has been used [17].

A simple and very successful method to define a lexical feature set for authorship attribution is to extract the most frequent words in the available corpus (comprising all the texts of the candidate authors). Then, a decision has to be made about defining the value of 'x' in most frequent 'x' words that will be used as features. In order to have contextual content, word n-grams are generated out of text. Also, with spelling mistakes and other such formatting errors one can have idiosyncratic features for an author.

**Character Features**

From a text various features can contribute to this category, like counts of alphabetic characters in a word, digits, uppercase and lowercase characters, punctuation marks etc. Character 'n' gram captures style and contextual information and captures any grammatical errors for an text. For instance, if we have used a word 'becuase' instead of 'because', then it will not be affected in case we are using character n-gram with value of n=2. In this category we use calculating most frequent as we do with words feature. Also, as word feature this is also tool independent.

An important aspect of character n-gram is to find appropriate value of 'n' where value of 'n' can range from 2,3,4 and so on. A large n would lead to increase in dimensionality in data, however it would better capture lexical and contextual information. On the other hand, a small n will represent sub-word information, but would not be appropriate for finding contextual information. Various researchers have tried with variable lengths of 'n' as well. These value of n is also dependent on languages of corpus, as in English n=4 has been found performing better than others.[1]

**Syntactic Features**

Syntactic feature is considered more reliable than other features as it has been observed that author often follows similar syntactic style unconsciously. Function words has been really successful in order to give syntactic information.
Moreover, function words as mentioned earlier are used by author unconsciously and so its hard to deceive by altering it.

### Idiosyncratic Features

These features refer to spelling mistakes or formatting errors that are done by an author. This can help to catch author as if the frequency of error is quite high then its more likely that author makes same mistake usually. There are many lists available with correct spellings and so it is not difficult to access whether word is written correctly or not.

### Application Based Features

Besides aforementioned features, other features which are particularly specific to application can also be used to better understand the style. Many structural measures revealed the possibility to define like mentioning greetings and farewells in particular way, signature types, use of long paragraphs etc. in the application domains such as e-mail messages and online-forum messages [1]. Moreover, these measures are important more in short texts as other features cannot be extracted from them efficiently. For areas where topic for all topics is same, to capture the style of an author, content-specific keywords can be used. In more clear way it can be understood that if texts belong to certain genre, one can use some words related to topic more often.

Table 3.1 gives just a brief introduction to various feature categories

**Table 4.1:** Type of stylometric features.

| Feature Category Name | Feature Type | Represents content or Style |
|---|---|---|
| Lexical | Sentence length,<br>Word length,<br>frequency of word,<br>word n grams,<br>function words etc. | Style except word n gram can capture content-specific information. |
| Character | Character types<br>(letters, digits, punctuation),<br>Character n-grams<br>( variable and fixed),<br>character count per sentence etc. | Style majorly with hints of contextual information if value of n in n-gram is large. |
| Syntactic | Parts of Speech(POS),<br>POS n-gram, | Style |
| Semantic | Synonyms,<br>semantic dependencies etc. | Style |
| Idiosyncratic | Errors: Collections of common spelling mistakes and grammatical mistakes. | Style |
| Application Based | Structural,<br>Content Specific words,<br>Language Specific | Style |

### 4.1.3 Dimensionality Reduction

There is a problem with classification algorithms, they tend to over-fit the training data when dimensionality of problem increases. However, existing machine learning algorithms like SVM can deal with the increased feature set size [1].

There are many reduction measures like mutual information, chi square, frequency thresholding, information gain, and term strength. Amongst these most productive are mutual information and information gain.

Information gain represents the entropy reduction given a certain feature, which means measuring the number of bits of information gained about the category prediction by knowing the presence or absence of a term in a document[18, 19]

$$IG(t_k, c_i) = \sum_{c \in \{c_i, \overline{c_i}\}} \sum_{t \in \left\{t_k, \overline{t_k}\right\}} P(t,c).log \frac{P(t,c)}{P(t).P(c)}$$

### 4.1.4 Evaluation Metric

We used Precision, Recall, and F1 for evaluation authorship attribution results. These measure can be used to find the rank of different settings. These measure are standard information retrieval measures

**Precision** Precision can be defined for a author A as fraction of correct attribution to A made by a system.

$$P_A = \frac{correct(A)}{attributions(A)}$$

**Recall** Recall can be defined for a author A as fraction of documents that are correctly attributed to A by total documents by A.

$$R_A = \frac{correct(A)}{documents - by(A)}$$

**F Measure** F1 measure combines precision and recall to give a balanced score, by taking harmonic means of both.

$$F_1 = \frac{2P_AR_A}{P_A + R_A}$$

**Macro average** Macro average is used to get a cumulative score of results for all authors.

$$macro - avg_M(\{A_i\}) = \frac{1}{n}\sum_i M_{A_i}$$

where n is total no of results and MAi is metric result for author Ai. Macro averaging gives same accuracy to all author irrespective of no of documents written by them.

**Micro Average** Micro average is also used to get a cumulative score of results for all authors.

$$micro - avg_M(\{A_i\}) = \frac{1}{k}\sum_i k_iM_{A_i}$$

where k is total documents by Author Ai. Micro averaging gives more weight to accuracy for authors with more test documents.

## 4.2 Learning overall

### 4.2.1 Cosine Similarity

To find similarity between two documents, there is correlation between term vectors in which document is represented. In order to find similarity, we calculate cosine theta ( theta being the angle between two documents). Given two documents A and B, cosine similarity is given as:

$$SIM_C(\overrightarrow{t_a}, \overrightarrow{t_b}) = \frac{\overrightarrow{t_a}.\overrightarrow{t_b}}{|\overrightarrow{t_a}|^2 \times |\overrightarrow{t_b}|^2}$$

where ta and tb are m-dimensional vectors over the term set T = t1,...,tm. [20]. Cosine similarity is always positive ranging between [0,1].

### 4.2.2 Jaccard Distance

Jaccard coefficient is fraction of shared terms sum to the summing term weights of both documents alone but are not shared.

$$SIM_J(\overrightarrow{t_a}, \overrightarrow{t_b}) = \frac{\overrightarrow{t_a}.\overrightarrow{t_b}}{|\overrightarrow{t_a}|^2 + |\overrightarrow{t_b}|^2 - \overrightarrow{t_a}.\overrightarrow{t_b}}$$

jaccard coefficient ranges from 0 to 1. When two documents are same than it is 1 and when they are disjoint it is 0.

### 4.2.3 Global weight

Global Weight gives more weight to important words or words those are more informative.

$$GW = 1 + \sum_{j=1} \frac{p_{ij} log p_{ij}}{log(ndocs)}$$

where pij is defined as

$$p_{ij} = \frac{tf_{ij}}{gf_i}$$

where tfij is term frequency of term i in document j and gfi is term frequency of term i in whole corpus.

$$g_{total} = \frac{\sum_{i=1}^{w} g_i}{w}$$

where w is total words in summary. gtotal tells that how much informative or unique a summary is.

### 4.2.4 Spearman rank correlation

For a sample of size n, the n raw scores Xi, Yi are converted to ranks xi, yi, and is computed from:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where di = xi - yi, is the difference between ranks.

# 5

# Implementation Details and Evaluation Results

Following the methodology described in chapter 4, experiments were conducted for authorship verification on 50 percent ground truth data set obtained in order to examine the performance of four classifiers. Best setting produced in above set was used to determine authorship on left over 50 percent testing documents. Further investigation was done to see learning happening by weighting it in terms of Jaccard similarity, Cosine similarity and global weight scores for each author for each of their summary.

## 5.1 Authorship Verification

### 5.1.1 Preprocessing

The data collected was not very noisy as assignments were graded, so this required small preprocessing of text. For authorship verification three main preprocessing were done: punctuation separator, unify case, normalise white-space. Punctuation separator consider each punctuation as a separate character and adds space before and after it. Normalise white-space is used to reduce extra spaces with single space. Unify case helps to convert all uppercase characters to lowercase.

Next, all the texts corresponding to same author were collected in order to train the classifier for all the 18 authors.

### 5.1.2 Feature Selection

With varying techniques already been worked out using individual and iterative feature sets, we tested each feature separately over the corpus. This was done in order to test which features were most useful in identifying authors on the data set that has been taken. These include character n-gram with n having values from 2 to 5, sentence length, Stanford POS, Stanford POS n-gram with n ranging from 2 to 5, Words, word N gram ranging from 2 to 5, and function words. For all the four classifiers, these 16 features were fed to determine style and hence best setting of classifier and corresponding feature set was taken for final attribution on testing samples. The function words used in our study are listed in website of sequence publishing[17].

### 5.1.3 Feature Reduction

Here, for both training and testing corpus, following was done: (a)For the start,a large set of unique features(words) are fetched for each feature in feature set. In case the feature set is huge, final features were obtained by first fetching 15,000 most frequent and then top 2,000 with highest Info Gain value. This was similar to the approach they followed in[19]. (b) Otherwise, just 3,000 features with highest info gain were considered where the unique features were less than 15,000 but more than 10,000. (c) In cases where they were less than 1,000, they were considered as it is.

### 5.1.4 K-fold approach for finding best setting on this data set

All experiments for training purpose are done using 3-fold cross-validation, with each fold of equal length. This allows us to get a reliable setting and an overall idea of how well the classifier will work on the new test set. Cross validation is done when data is limited and when we wan to swap the roles of test and train document both. The ground truth data is divided into 3 parts for each author. So one setting includes 18 times one of the subsets (1 out of 3) for each author as test case and other 18 times two subset( rest 2 out of 3) for each author as training. This is done so that each document vector comes as a test case at least once.

An over all view for the approach is given in figure 5.1 which shows that how authorship verification is done with the best setting obtained from training data over different classifying algorithms as Burrows Delta, SVM, Naive Bayes and Decision Tree.

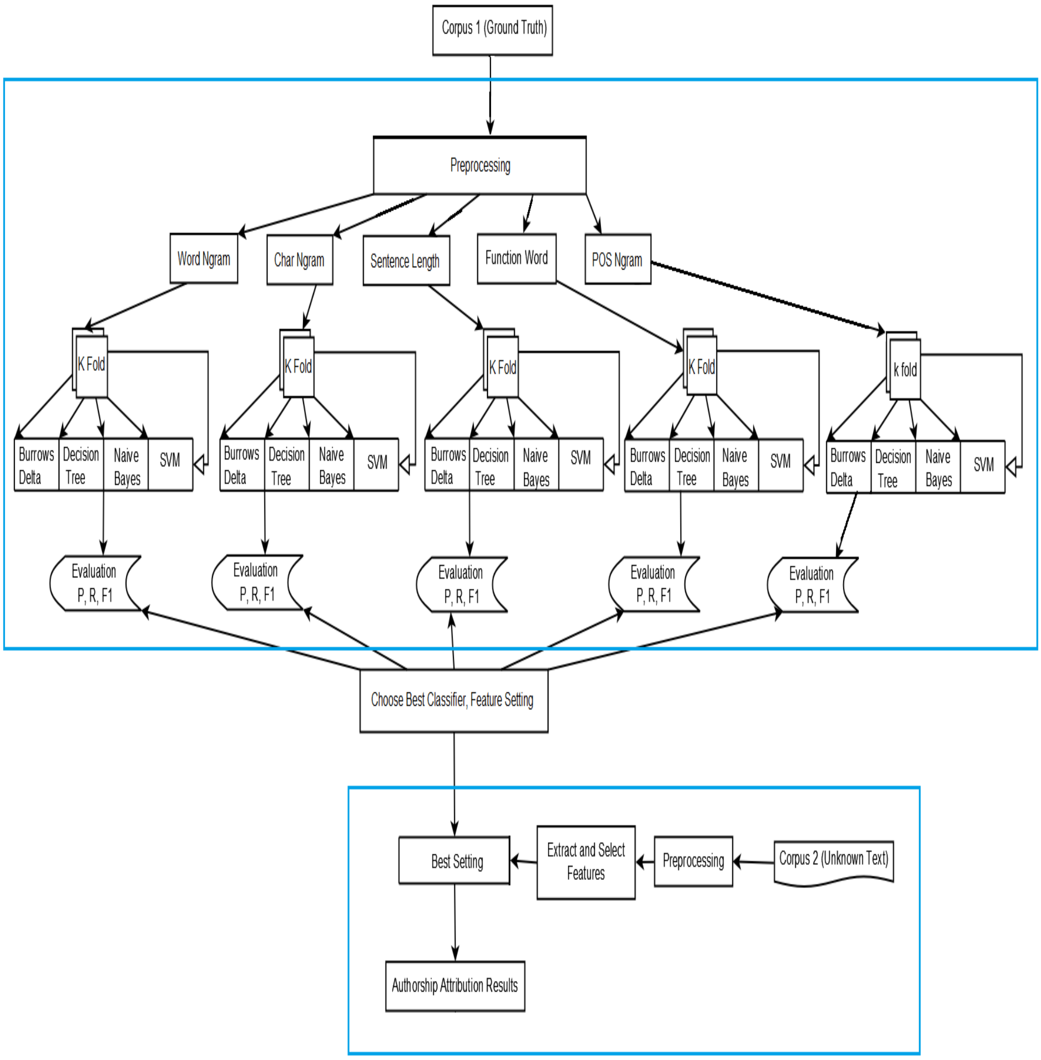

**Figure 5.1:** Overall approach on authorship attribution

### 5.1.5 Experiment on ground truth for best setting

In order to evaluate the performance of features for each classifier, 3 sets of 16 experiments were conducted i.e. a total of 48 experiments. Here, we call one feature plus specific classifier as one setting, for example: character n gram 2 on classifier Burrows Delta is one setting. So for a particular classifier there will be 16 settings and for each setting 3 fold experiments were run. After classifying test data in that fold, precision, recall and F1 measure are calculated for cumulatively all authors by considering macro-averaging and micro-averaging methods. So in total there are six scores corresponding 1 setting in 1 fold i.e. micro precision, micro recall, micro F1, macro precision, macro recall and macro F1. After all 3 folds, average of values of precision, Recall, F1 for both micro and micro is done for that particular setting. Now ranks were calculated for each classifier amongst all the features. Individual ranking was done within each micro precision, micro recall, micro F1, macro precision, macro recall and macro F1 and then combined ranking was generated for each setting related to each classifier. This was done with the vision provided by PAN in the paper [21].

All the results are represented in form of tables with values of precision, recall and f1 measure for features and corresponding classifiers. The obtained values reflect how is the performance of that classifier in the setting as to how correctly texts are attributed. The first column in the table tells about the 16 features that have been mentioned earlier.

Table 5.1 and figure 5.2 show the results for the aforementioned settings where the outcome are in decreasing order of overall ranks with least rank as the best performer and last placed as the weakest. For Burrows Delta, the best setting came out with 0.770 macro precision, 0.733 macro recall, 0.705 macro f1, 0.797 micro precision, 0.740 micro recall and 0.725 micro f1 measure. There is no much difference in micro and macro scores as classes are almost balance. Ranks obtained for macro is 3, for micro is 10 and overall it is 13.
On contrary, table 5.2 for SVM had best result for word n gram 3 with highest score that of micro precision with value 0.889 and ranking got improved with 3 of macro, 3 of micro and 6 of overall. These observations can also be interpreted through graphical representation as given in figure 5.5.
Data provided in table 5.3 and graphical representation in figure 5.3 reflect how well decision tree achieve results with highest score of 0.971 micro precision on setting of word n gram 4. In this the overall ranking improved to 8, which is an

interesting outcome.

Though Naive Bayes perform really well with some corpus, but here in our case it was more or less similar to Burrows Delta. However, the best setting of Naive Bayes came out with word n gram with micro precision value of 0.847 as is apparent from the data given in table 5.4 and presented in figure5.4. The ranks for this classifier were comparable to that of SVM.

The top 3 best settings for all classifiers were considered and re-ranking was done with their respective values of precision, recall and f1 measure. Table 5.5 shows the respective ranks of all classifiers with the respective setting chosen. Decision tree with all the three settings outperform others, best being the word n-gram 4 with overall rank of 8. The error rate was minimum for decision tree classifier with setting of word n gram 4.

**Table 5.1:** Values of Precision, Recall, F1 and final ranking on all features with Burrows Delta as classifier.

| | | Macro Averaging | | | Micro Averaging | | A | B | A+B |
|---|---|---|---|---|---|---|---|---|---|
| Feature Set | F Measure | Precision | Recall | F Measure | Precision | Recall | Macro (F+P+R) rank | Micro (F+P+R) rank | Overall Rank |
| Character N gram 4 | 0.705 | 0.770 | 0.733 | 0.725 | 0.797 | 0.740 | 3 | 10 | 13 |
| Character N gram 5 | 0.646 | 0.728 | 0.668 | 0.672 | 0.764 | 0.684 | 6 | 19 | 25 |
| Character N gram 3 | 0.483 | 0.625 | 0.497 | 0.623 | 0.825 | 0.619 | 12 | 19 | 31 |
| Word N gram 5 | 0.288 | 0.346 | 0.342 | 0.659 | 0.813 | 0.750 | 21 | 10 | 31 |
| Character N gram 2 | 0.466 | 0.591 | 0.497 | 0.600 | 0.774 | 0.622 | 13 | 24 | 37 |
| Word N gram 4 | 0.245 | 0.307 | 0.287 | 0.643 | 0.831 | 0.714 | 25 | 12 | 37 |
| Stanford POS | 0.532 | 0.545 | 0.598 | 0.552 | 0.559 | 0.617 | 11 | 33 | 44 |
| Stanford POS Ngram 2 | 0.373 | 0.388 | 0.449 | 0.501 | 0.628 | 0.720 | 18 | 29 | 47 |
| Word N gram 3 | 0.155 | 0.192 | 0.200 | 0.637 | 0.779 | 0.788 | 34 | 14 | 48 |
| Stanford POS Ngram 4 | 0.191 | 0.283 | 0.214 | 0.505 | 0.764 | 0.590 | 29 | 33 | 62 |
| Stanford POS Ngram 5 | 0.074 | 0.095 | 0.119 | 0.527 | 0.663 | 0.814 | 40 | 22 | 62 |
| Word N gram 2 | 0.146 | 0.243 | 0.182 | 0.495 | 0.792 | 0.601 | 35 | 30 | 65 |
| function words | 0.215 | 0.269 | 0.294 | 0.430 | 0.538 | 0.595 | 27 | 42 | 69 |
| Stanford POS Ngram 3 | 0.069 | 0.114 | 0.104 | 0.441 | 0.695 | 0.631 | 42 | 32 | 74 |
| Words | 0.004 | 0.002 | 0.055 | 0.086 | 0.045 | 1 | 48 | 33 | 81 |
| Sentence Length | 0.071 | 0.068 | 0.080 | 0.219 | 0.248 | 0.355 | 44 | 46 | 90 |

**Table 5.2:** Values of Precision, Recall, F1 and final ranking on all features with SVM as classifier.

| | | Macro Averaging | | | Micro Averaging | | A | B | A+B |
|---|---|---|---|---|---|---|---|---|---|
| Feature Set | F Measure | Precision | Recall | F Measure | Precision | Recall | Macro (F+P+R) Rank | Micro (F+P+R) Rank | Overall Rank |
| Word N gram 3 | 0.841 | 0.876 | 0.851 | 0.862 | 0.889 | 0.849 | 3 | 3 | 6 |
| Word N gram 4 | 0.825 | 0.869 | 0.832 | 0.853 | 0.879 | 0.825 | 6 | 6 | 12 |
| Stanford POS Ngram 5 | 0.793 | 0.828 | 0.808 | 0.831 | 0.849 | 0.821 | 10 | 10 | 20 |
| Word N gram 5 | 0.777 | 0.835 | 0.783 | 0.820 | 0.859 | 0.792 | 13 | 12 | 25 |
| Stanford POS Ngram 2 | 0.784 | 0.812 | 0.800 | 0.811 | 0.829 | 0.808 | 13 | 14 | 27 |
| Words | 0.763 | 0.799 | 0.779 | 0.777 | 0.805 | 0.778 | 19 | 19 | 38 |
| Character N gram 4 | 0.737 | 0.804 | 0.740 | 0.773 | 0.815 | 0.748 | 20 | 20 | 40 |
| Character N gram 5 | 0.718 | 0.774 | 0.737 | 0.741 | 0.779 | 0.724 | 24 | 25 | 49 |
| Word N gram 2 | 0.694 | 0.761 | 0.681 | 0.738 | 0.780 | 0.706 | 30 | 28 | 58 |
| Character N gram 2 | 0.700 | 0.755 | 0.716 | 0.718 | 0.757 | 0.707 | 28 | 31 | 59 |
| Stanford POS Ngram 3 | 0.676 | 0.725 | 0.700 | 0.725 | 0.754 | 0.723 | 32 | 30 | 62 |
| Stanford POS Ngram 4 | 0.592 | 0.673 | 0.589 | 0.643 | 0.687 | 0.608 | 39 | 38 | 77 |
| Stanford POS | 0.620 | 0.658 | 0.640 | 0.634 | 0.665 | 0.634 | 38 | 40 | 78 |
| Character N gram 3 | 0.590 | 0.659 | 0.603 | 0.636 | 0.682 | 0.614 | 40 | 39 | 79 |
| function words | 0.565 | 0.605 | 0.577 | 0.597 | 0.627 | 0.582 | 45 | 45 | 90 |
| Sentence Length | 0.179 | 0.199 | 0.190 | 0.344 | 0.365 | 0.356 | 48 | 48 | 96 |

**Table 5.3:** Values of Precision, Recall, F1 and final ranking on all features with Decision Tree as classifier.

| | | Macro Averaging | | | Micro Averaging | | A | B | A+B |
|---|---|---|---|---|---|---|---|---|---|
| Feature Set | F Measure | Precision | Recall | F Measure | Precision | Recall | Macro (F+P+R) Rank | Micro (F+P+R) Rank | Overall Rank |
| Word N gram 4 | 0.950 | 0.960 | 0.962 | 0.962 | 0.971 | 0.967 | 4 | 4 | 8 |
| Word N gram 2 | 0.941 | 0.949 | 0.952 | 0.951 | 0.959 | 0.954 | 7 | 7 | 14 |
| Word N gram 5 | 0.936 | 0.960 | 0.945 | 0.949 | 0.972 | 0.950 | 7 | 7 | 14 |
| Word N gram 3 | 0.930 | 0.947 | 0.942 | 0.942 | 0.959 | 0.946 | 12 | 12 | 24 |
| Words | 0.914 | 0.927 | 0.929 | 0.925 | 0.941 | 0.929 | 15 | 15 | 30 |
| Character N gram 3 | 0.835 | 0.862 | 0.838 | 0.841 | 0.869 | 0.844 | 18 | 18 | 36 |
| Stanford POS Ngram 5 | 0.797 | 0.820 | 0.802 | 0.817 | 0.836 | 0.824 | 23 | 22 | 45 |
| Stanford POS Ngram 4 | 0.794 | 0.824 | 0.798 | 0.813 | 0.839 | 0.819 | 24 | 23 | 47 |
| Character N gram 5 | 0.791 | 0.817 | 0.807 | 0.804 | 0.832 | 0.812 | 25 | 27 | 52 |
| Character N gram 4 | 0.722 | 0.764 | 0.744 | 0.726 | 0.777 | 0.737 | 31 | 34 | 65 |
| Stanford POS Ngram 3 | 0.709 | 0.755 | 0.721 | 0.738 | 0.784 | 0.741 | 34 | 31 | 65 |
| Character N gram 2 | 0.684 | 0.787 | 0.685 | 0.705 | 0.810 | 0.709 | 34 | 34 | 68 |
| Stanford POS Ngram 2 | 0.588 | 0.645 | 0.609 | 0.630 | 0.693 | 0.638 | 39 | 39 | 78 |
| function words | 0.464 | 0.503 | 0.478 | 0.535 | 0.583 | 0.546 | 42 | 42 | 84 |
| Stanford POS | 0.293 | 0.355 | 0.296 | 0.370 | 0.455 | 0.372 | 45 | 45 | 90 |
| Sentence Length | 0.100 | 0.098 | 0.106 | 0.284 | 0.283 | 0.296 | 48 | 48 | 96 |

**Table 5.4:** Values of Precision, Recall, F1 and final ranking on all features with Naive Bayes Classifier as classifier.

| | | Macro Averaging | | | Micro Averaging | | A | B | A+B |
|---|---|---|---|---|---|---|---|---|---|
| Feature Set | F Measure | Precision | Recall | F Measure | Precision | Recall | Macro (F+P+R) Rank | Micro (F+P+R) Rank | Overall Rank |
| Word N gram 5 | 0.783 | 0.839 | 0.792 | 0.791 | 0.847 | 0.791 | 3 | 3 | 6 |
| Word N gram 4 | 0.731 | 0.790 | 0.740 | 0.738 | 0.800 | 0.737 | 6 | 6 | 12 |
| Word N gram 3 | 0.700 | 0.777 | 0.713 | 0.700 | 0.783 | 0.703 | 9 | 9 | 18 |
| Character N gram 5 | 0.461 | 0.510 | 0.509 | 0.574 | 0.643 | 0.628 | 18 | 16 | 34 |
| Character N gram 4 | 0.439 | 0.549 | 0.459 | 0.513 | 0.639 | 0.539 | 20 | 20 | 40 |
| Stanford POS | 0.457 | 0.522 | 0.467 | 0.512 | 0.581 | 0.524 | 18 | 26 | 44 |
| Words | 0.443 | 0.557 | 0.455 | 0.505 | 0.638 | 0.518 | 19 | 25 | 44 |
| Character N gram 2 | 0.430 | 0.521 | 0.444 | 0.519 | 0.594 | 0.499 | 25 | 25 | 50 |
| Word N gram 2 | 0.388 | 0.578 | 0.356 | 0.500 | 0.751 | 0.457 | 25 | 25 | 50 |
| Stanford POS Ngram 2 | 0.424 | 0.585 | 0.433 | 0.458 | 0.636 | 0.467 | 22 | 31 | 53 |
| Stanford POS Ngram 4 | 0.236 | 0.348 | 0.245 | 0.446 | 0.677 | 0.470 | 39 | 26 | 65 |
| Character N gram 3 | 0.309 | 0.401 | 0.334 | 0.406 | 0.531 | 0.438 | 33 | 39 | 72 |
| Stanford POS Ngram 5 | 0.192 | 0.321 | 0.201 | 0.404 | 0.737 | 0.462 | 43 | 29 | 72 |
| Stanford POS Ngram 3 | 0.204 | 0.363 | 0.198 | 0.363 | 0.650 | 0.352 | 41 | 37 | 78 |
| function words | 0.279 | 0.318 | 0.302 | 0.384 | 0.433 | 0.418 | 39 | 43 | 82 |
| Sentence Length | 0.131 | 0.170 | 0.132 | 0.303 | 0.372 | 0.314 | 48 | 48 | 96 |

**Table 5.5:** Comparing top three results of all four classifiers and corresponding ranks.

| | | Macro Averaging | | | Micro Averaging | | A | B | A+B |
|---|---|---|---|---|---|---|---|---|---|
| Feature Set | F Measure | Precision | Recall | F Measure | Precision | Recall | Macro (F+P+R) Rank | Micro (F+P+R) Rank | Overall Rank |
| Word N gram 4 Decision Tree | 0.950 | 0.960 | 0.962 | 0.962 | 0.971 | 0.967 | 4 | 4 | 8 |
| Word N gram 2 Decision Tree | 0.941 | 0.949 | 0.952 | 0.951 | 0.959 | 0.954 | 7 | 7 | 14 |
| Word N gram 5 Decision Tree | 0.936 | 0.960 | 0.945 | 0.949 | 0.972 | 0.950 | 7 | 7 | 14 |
| Word N gram 3 SVM | 0.841 | 0.876 | 0.851 | 0.862 | 0.889 | 0.849 | 12 | 12 | 24 |
| Word N gram 4 SVM | 0.825 | 0.869 | 0.832 | 0.853 | 0.879 | 0.825 | 15 | 15 | 30 |
| Stanford POS Ngram 5 SVM | 0.793 | 0.828 | 0.808 | 0.831 | 0.849 | 0.821 | 19 | 18 | 37 |
| Word N gram 5 Nave bayes | 0.783 | 0.839 | 0.792 | 0.791 | 0.847 | 0.791 | 20 | 21 | 41 |
| Word N gram 4 Nave bayes | 0.731 | 0.790 | 0.740 | 0.738 | 0.800 | 0.737 | 24 | 26 | 50 |
| Character N gram 4 Burrows | 0.705 | 0.771 | 0.733 | 0.726 | 0.797 | 0.740 | 28 | 27 | 55 |
| Word N gram 3 Nave bayes | 0.700 | 0.777 | 0.713 | 0.700 | 0.783 | 0.703 | 29 | 31 | 60 |
| Character N gram 5 Burrows | 0.646 | 0.728 | 0.668 | 0.673 | 0.764 | 0.684 | 33 | 34 | 67 |
| Character N gram 3 Burrows | 0.483 | 0.625 | 0.497 | 0.623 | 0.825 | 0.619 | 36 | 32 | 68 |

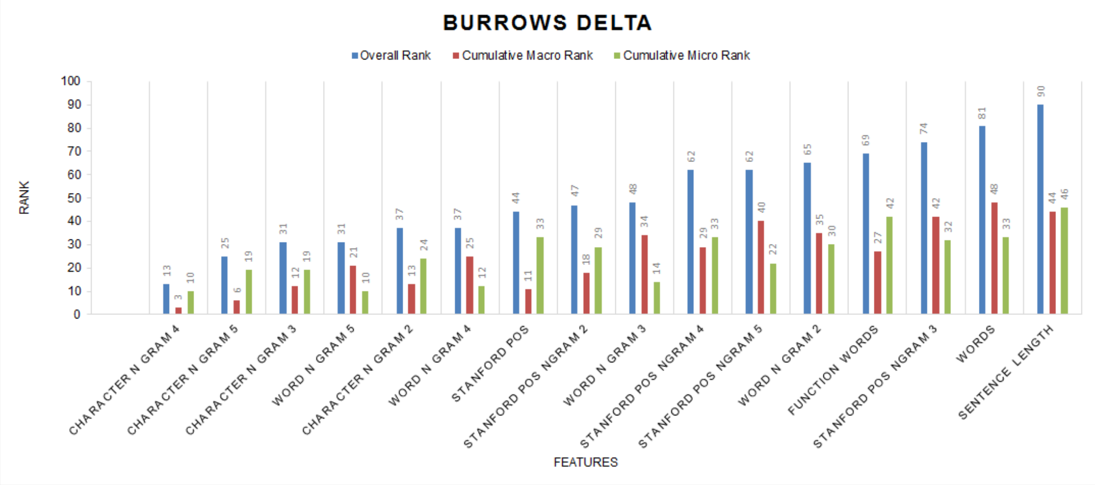

**Figure 5.2:** Burrows Delta with ranking amongst features.

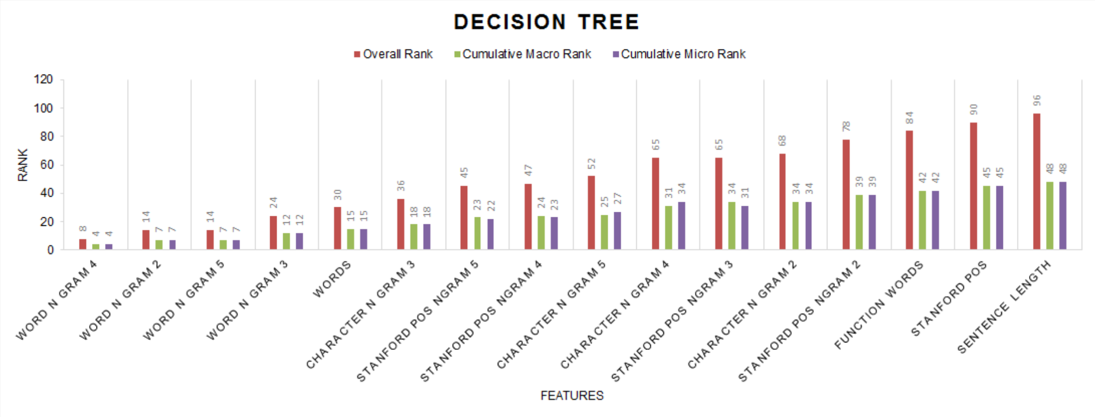

**Figure 5.3:** Decision tree with ranking amongst features

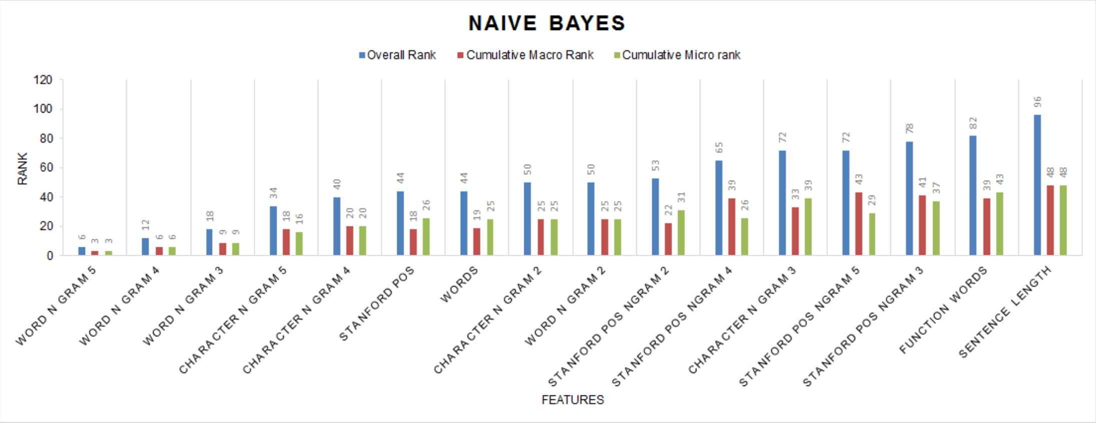

**Figure 5.4:** Naive Bayes with ranking amongst features

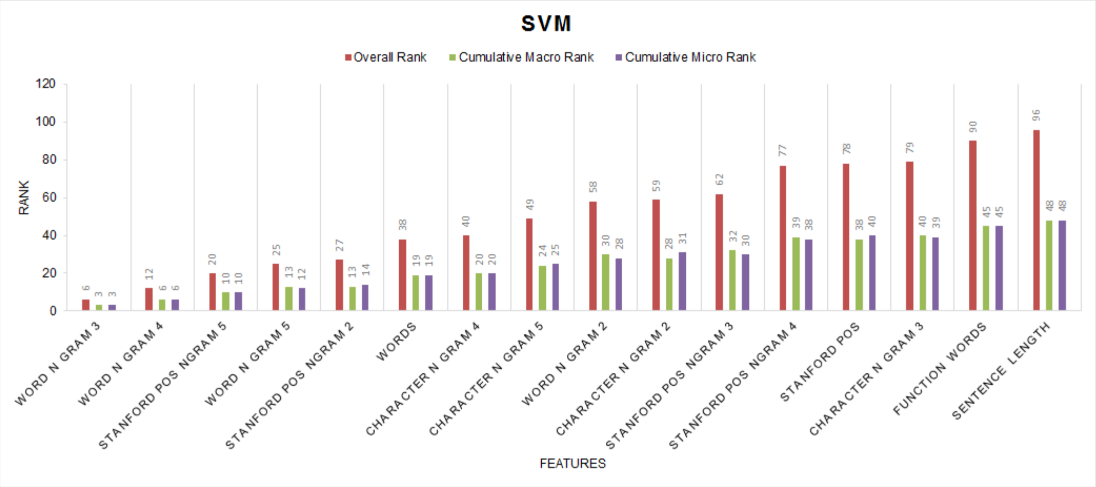

**Figure 5.5:** SVM with ranking amongst features

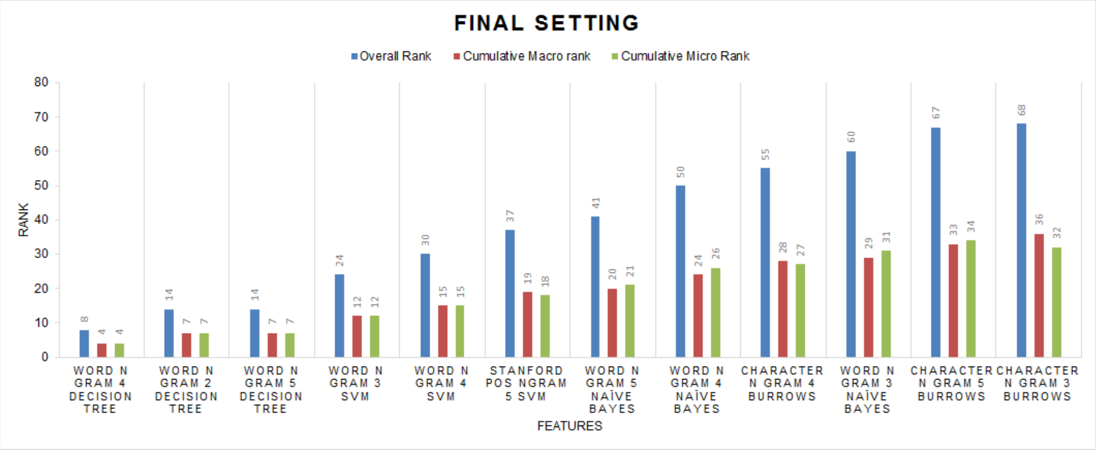

**Figure 5.6:** Over all with ranking amongst features

## 5.2 Attributions for testing data

From the above setting, Decision tree classifier was chosen with word n gram 4 feature to do attributions on the testing data. Out of 18 authors only three were there who had false attributions (Table 5.6). It For author 5 and 10 there were 2 documents each which were suspicious and for author 15 it was only one such document.

**Table 5.6:** Final attributions done for all 18 authors with decision tree classifier having feature setting of word n gram 4.

| Name of Author | Original number of Test Documents for each author | Number of documents attributed to same author | Number of documents not written by the author |
|---|---|---|---|
| Author 1 | 13 | 13 | 0 |
| Author 2 | 13 | 13 | 0 |
| Author 3 | 13 | 13 | 0 |
| Author 4 | 13 | 13 | 0 |
| Author 5 | 13 | 11 | 2 |
| Author 6 | 12 | 12 | 0 |
| Author 7 | 12 | 12 | 0 |
| Author 8 | 12 | 12 | 0 |
| Author 9 | 13 | 13 | 0 |
| Author 10 | 13 | 11 | 2 |
| Author 11 | 13 | 13 | 0 |
| Author 12 | 13 | 13 | 0 |
| Author 13 | 13 | 13 | 0 |
| Author 14 | 13 | 13 | 0 |
| Author 15 | 14 | 13 | 1 |
| Author 16 | 13 | 13 | 0 |
| Author 17 | 12 | 12 | 0 |
| Author 18 | 13 | 13 | 0 |
| Total | 231 | 226 | 5 |

## 5.3 Learning

After verifying all authors, only 5 documents were there which got removed as they were not written by the same author who submitted it. The outcome of such less documents that were not self written by student could be because the corpus was of such an institute where strict measures are taken against any kind of plagiarism. This satisfies the result that largely documents are written by the student themselves who are submitting it as well. In next step, we are checking whether learning is actually happening even if students are writing summary by themselves. For this, distance(Cosine and Jaccard) between submitted summary by an author and corresponding original paper is calculated. Similarly after calculating scores for every summary submitted, we can have scores for an particular author for all his summaries in respect to all original research papers(figure 5.7).

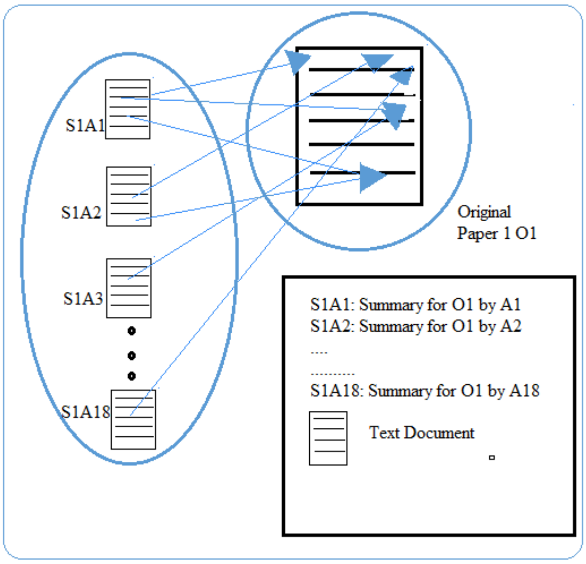
S1A1
S1A2
S1A3
S1A18
Original
Paper 1 O1
S1A1: Summary for O1 by A1
S1A2: Summary for O1 by A2
....
..........
S1A18: Summary for O1 by A18
Text Document

**Figure 5.7:** Similarity between submitted summaries by all students and corresponding original research paper.

Another measure that was considered was to calculate global weight for each summary of each author. This was calculated by calculating global weight for a summary against all similar submitted summaries corresponding to one research paper. This way again, profile for an author can be made where scores are generated for each of his summary. However, scores generated here are with respect to other similar summaries submitted by other authors regarding the same original research paper. Fig 5.8 helps to understand the criteria which is followed.

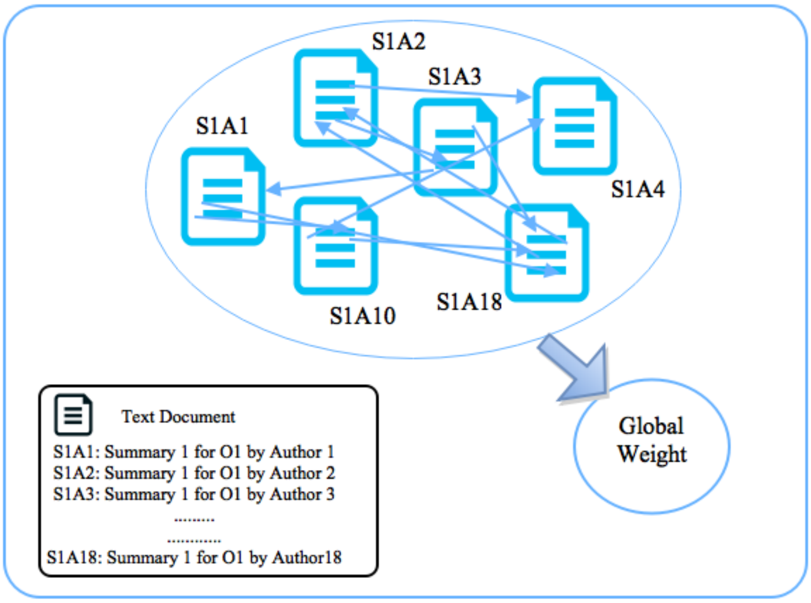
S1A2
S1A3
S1A1
S1A4
S1A10
S1A18
Global Weight
Text Document
S1A1: Summary 1 for O1 by Author 1
S1A2: Summary 1 for O1 by Author 2
S1A3: Summary 1 for O1 by Author 3
.........
............
S1A18: Summary 1 for O1 by Author18

**Figure 5.8:** Global weight between all similar summaries submitted by students corresponding to original research paper.

### 5.3.1 Using Similarity scores

By going on the assumption that every author writes summary after reading the paper, there must be some percentage by which author gets influenced or might collect something from paper which does not attributes to his own style.

In totality there is a situation where O1,O2,O3.........On: are the set of original papers, A1,A2,A3.....Am: are set of authors writing summaries for these papers, and S1,S2,S3.....St: are the corresponding summaries written by these authors for the original set of papers. Note every author has not written every summary.

We can say suppose we have summaries in a way as:

A1: S1,S2,S3,S6... for O1,O2,O3,O6.....

A2: S1,S3,S4,S5... for O1,O3,O4,O5.....

A3: S1,S2,S3,S4... for O1,O2,O3,O4.....

A4: S1,S2,S4,S7... for O1,O2,O4,O7..... and so on.

Scores were calculated for every author by matching trigram[6] between the two texts, below say for author A1 scores are:

Score 1: for S1 and O1

Score 2: for S2 and O2
Score 3: for S3 and O3
.....
...........
Score t: for St and Ot.
Similarly for A2{Score 1,Score 2.....Score t} ,A3{..}... and so on.

For a single author, AUC(area under curve ) was calculated for all his scores generated for all summaries submitted. Figure 5.9 shows an overall view on how similarity scores are generated and overall learning is measured.

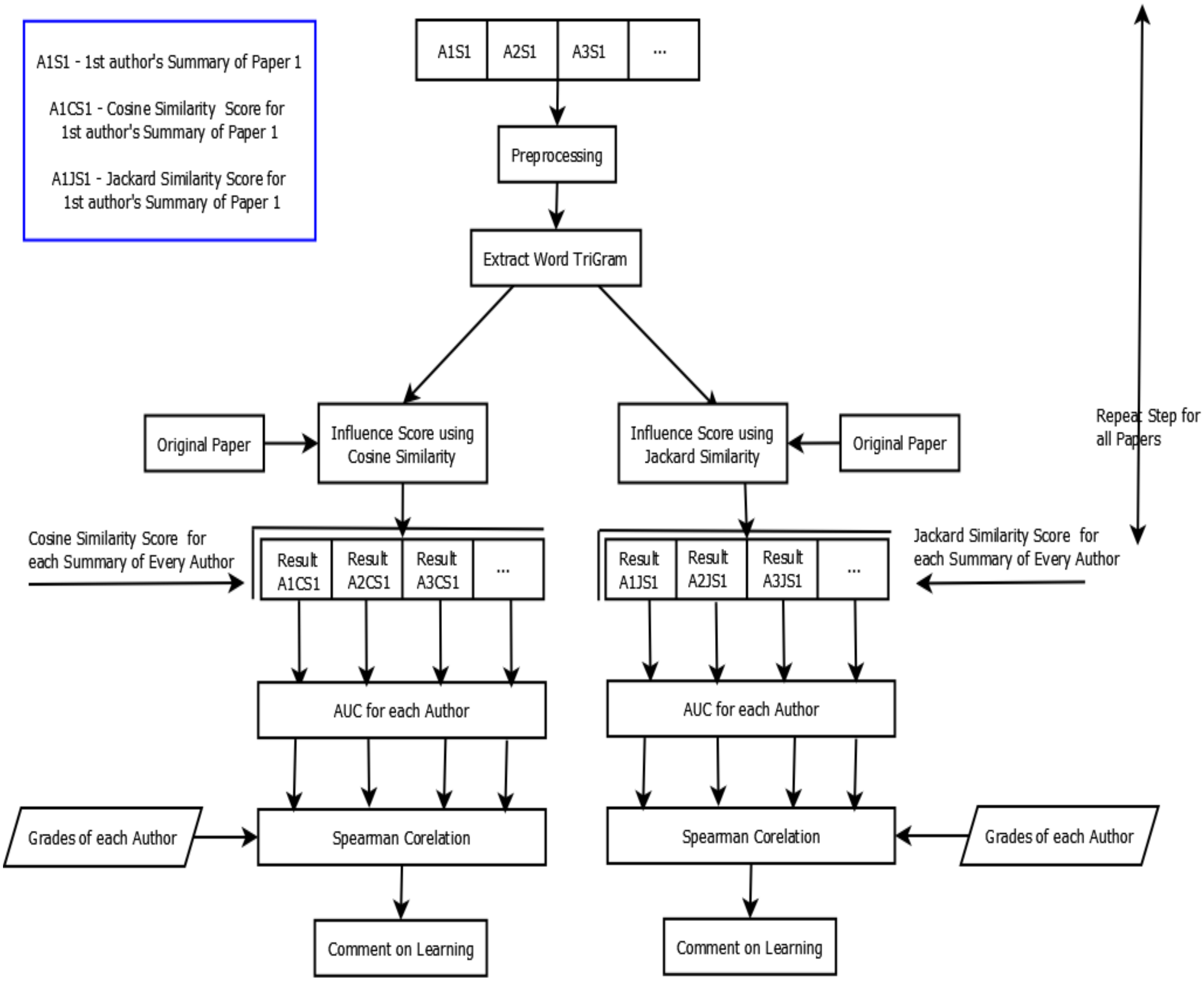
A1S1 - 1st author's Summary of Paper 1
A1CS1 - Cosine Similarity Score for 1st author's Summary of Paper 1
A1JS1 - Jackard Similarity Score for 1st author's Summary of Paper 1
A1S1
A2S1
A3S1
...
Preprocessing
Extract Word TriGram
Original Paper
Influence Score using Cosine Similarity
Influence Score using Jackard Similarity
Original Paper
Repeat Step for all Papers
Cosine Similarity Score for each Summary of Every Author
Result A1CS1
Result A2CS1
Result A3CS1
...
Result A1JS1
Result A2JS1
Result A3JS1
...
Jackard Similarity Score for each Summary of Every Author
AUC for each Author
AUC for each Author
Grades of each Author
Spearman Corelation
Spearman Corelation
Grades of each Author
Comment on Learning
Comment on Learning

**Figure 5.9:** Similarity scores using Jaccard and Cosine and further relating their AUC to human graded assignments.

This way, AUC scores for all authors are correlated with human graded scores using Spearman's rank correlation.

Table 5.7 gives us the Jaccard distance corresponding to each author and his summary with the original paper as mentioned above, where paper1, paper2, paper3.....paper31 corresponds to all original papers. Places left blank signifies that author has not submitted summary for that paper. Cosine similarity scores are shown in table 5.8 calculated in similar fashion as above.

**Table 5.7:** Jaccard Distance Similarity

| | | | | | | | | Name | of | Authors | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Paper Name | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 | A9 | A10 | A11 | A12 | A13 | A14 | A15 | A16 | A17 | A18 |
| Paper1 | | 0.005 | 0.005 | 0.013 | 0.009 | 0.002 | 0.004 | 0.011 | 0.006 | 0.054 | 0.007 | 0.003 | 0.001 | | 0.050 | 0.009 | 0.012 | 0.015 |
| Paper2 | 0.018 | 0.008 | 0.010 | 0.009 | 0.010 | 0.003 | | 0.018 | 0.007 | | 0.011 | 0.011 | 0.004 | 0.018 | 0.014 | 0.006 | | 0.008 |
| Paper3 | 0.010 | 0.008 | | 0.003 | 0.005 | 0.001 | 0.004 | 0.006 | 0.003 | 0.003 | 0.008 | 0.020 | 0.002 | | 0.005 | 0.005 | 0.007 | 0.009 |
| Paper4 | 0.015 | 0.007 | 0.004 | 0.013 | | 0.003 | | 0.008 | 0.007 | 0.007 | 0.008 | 0.006 | 0.004 | | | | 0.008 | 0.005 |
| Paper5 | | | 0.003 | 0.003 | 0.004 | 0.007 | 0.002 | 0.010 | 0.002 | | 0.010 | 0.010 | 0.003 | 0.005 | 0.015 | 0.004 | | 0.006 |
| Paper6 | | 0.005 | | | 0.004 | | 0.004 | 0.008 | 0.004 | 0.006 | 0.006 | 0.012 | 0.005 | 0.004 | 0.010 | 0.005 | 0.008 | |
| Paper7 | 0.042 | 0.027 | | | 0.012 | | 0.003 | 0.030 | 0.007 | | | 0.007 | 0.003 | 0.007 | | 0.011 | 0.010 | 0.008 |
| Paper8 | 0.011 | 0.007 | 0.005 | 0.005 | 0.003 | 0.001 | 0.005 | 0.004 | | 0.005 | 0.008 | 0.006 | 0.003 | 0.007 | 0.004 | | 0.004 | 0.004 |
| Paper9 | 0.007 | 0.008 | 0.007 | 0.003 | 0.002 | 0.001 | 0.005 | | 0.006 | 0.004 | 0.007 | | 0.002 | 0.001 | 0.009 | 0.003 | 0.007 | 0.008 |
| Paper10 | 0.022 | 0.017 | 0.009 | 0.011 | | 0.012 | 0.003 | | 0.010 | 0.016 | 0.016 | 0.018 | | 0.013 | 0.016 | 0.008 | | 0.010 |
| Paper11 | 0.021 | 0.012 | 0.008 | 0.012 | 0.008 | 0.005 | 0.005 | 0.011 | 0.005 | 0.004 | 0.005 | 0.010 | 0.007 | 0.009 | 0.018 | 0.005 | 0.011 | 0.010 |
| Paper12 | | 0.020 | | | 0.010 | | 0.007 | 0.026 | 0.008 | 0.013 | 0.018 | 0.014 | 0.010 | 0.010 | 0.031 | 0.012 | 0.016 | 0.009 |
| Paper13 | 0.009 | 0.009 | 0.006 | 0.004 | 0.005 | 0.003 | | 0.005 | 0.006 | 0.012 | 0.004 | 0.007 | 0.005 | 0.007 | | 0.005 | 0.007 | 0.006 |
| Paper14 | 0.017 | 0.003 | 0.003 | 0.009 | | | 0.004 | | 0.008 | 0.011 | 0.004 | 0.013 | 0.004 | 0.011 | 0.024 | 0.003 | 0.010 | 0.014 |
| Paper15 | 0.036 | 0.028 | 0.024 | 0.025 | 0.015 | 0.014 | 0.038 | 0.019 | 0.021 | 0.041 | 0.032 | 0.024 | 0.013 | 0.020 | 0.057 | | 0.016 | 0.035 |
| Paper16 | 0.023 | 0.025 | 0.019 | 0.012 | 0.012 | 0.010 | 0.015 | | 0.011 | 0.016 | 0.014 | | 0.007 | | 0.037 | 0.006 | 0.016 | 0.022 |
| Paper17 | 0.012 | 0.007 | 0.004 | 0.006 | 0.006 | 0.002 | 0.004 | | 0.006 | 0.005 | 0.003 | 0.010 | | 0.004 | 0.005 | 0.005 | | 0.007 |
| Paper18 | | 0.020 | 0.009 | 0.009 | 0.006 | | 0.005 | | 0.008 | 0.011 | 0.004 | 0.011 | 0.004 | 0.005 | 0.005 | 0.005 | 0.008 | |
| Paper19 | 0.018 | 0.007 | 0.016 | 0.009 | 0.009 | 0.003 | 0.010 | 0.010 | 0.014 | 0.115 | 0.008 | 0.010 | 0.004 | 0.015 | 0.014 | 0.007 | 0.008 | 0.008 |
| Paper20 | 0.026 | 0.008 | 0.007 | 0.003 | 0.005 | 0.002 | 0.002 | 0.009 | 0.003 | 0.009 | 0.006 | 0.008 | 0.005 | 0.004 | 0.011 | 0.002 | 0.006 | 0.005 |
| Paper21 | 0.013 | 0.008 | 0.005 | | | | 0.005 | 0.007 | 0.003 | 0.008 | 0.005 | 0.012 | 0.003 | 0.007 | 0.009 | | 0.008 | |
| Paper22 | 0.013 | 0.016 | 0.015 | 0.011 | 0.009 | 0.010 | 0.011 | 0.012 | 0.013 | 0.010 | 0.012 | | | 0.011 | 0.028 | 0.008 | 0.012 | 0.007 |
| Paper23 | 0.011 | 0.006 | 0.009 | 0.012 | | 0.004 | 0.009 | 0.010 | 0.005 | | 0.008 | 0.013 | 0.003 | 0.009 | 0.015 | 0.006 | 0.012 | 0.010 |
| Paper24 | 0.006 | 0.006 | 0.002 | 0.012 | 0.005 | 0.003 | 0.010 | 0.006 | | 0.034 | 0.005 | 0.006 | 0.003 | | 0.020 | 0.004 | 0.007 | 0.009 |
| Paper25 | 0.004 | | 0.001 | 0.003 | | | 0.001 | 0.002 | 0.001 | 0.001 | | 0.005 | 0.001 | | 0.008 | 0.001 | 0.003 | 0.004 |
| Paper26 | 0.021 | 0.012 | 0.005 | 0.011 | 0.005 | 0.005 | | | 0.011 | 0.004 | | 0.011 | 0.003 | 0.005 | 0.023 | | 0.007 | 0.014 |
| Paper27 | 0.011 | 0.007 | 0.009 | 0.008 | 0.007 | 0.004 | | | 0.012 | 0.037 | 0.007 | 0.008 | 0.006 | 0.006 | 0.010 | 0.002 | 0.007 | 0.010 |
| Paper28 | 0.010 | 0.002 | | 0.002 | 0.003 | 0.002 | 0.009 | 0.008 | | | 0.002 | 0.008 | 0.002 | 0.009 | 0.011 | 0.005 | | 0.012 |
| Paper29 | 0.013 | | 0.010 | 0.004 | 0.006 | 0.003 | 0.003 | 0.019 | 0.002 | | 0.006 | 0.018 | 0.002 | 0.007 | 0.011 | 0.005 | | 0.012 |
| Paper30 | 0.010 | | 0.003 | 0.004 | | | | 0.007 | 0.003 | 0.007 | | 0.002 | 0.002 | 0.006 | 0.049 | 0.004 | 0.005 | 0.003 |
| Paper31 | 0.023 | 0.013 | 0.011 | 0.009 | 0.007 | 0.005 | 0.007 | 0.001 | 0.006 | 0.004 | 0.007 | 0.028 | 0.002 | 0.008 | 0.013 | 0.007 | 0.004 | 0.011 |

**Table 5.8:** Cosine Similarity

| | Name of Authors | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Paper Name | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 | A9 | A10 | A11 | A12 | A13 | A14 | A15 | A16 | A17 | A18 |
| Paper1 | | 0.025 | 0.022 | 0.058 | 0.038 | 0.010 | 0.020 | 0.045 | 0.026 | 0.229 | 0.029 | 0.015 | 0.006 | | 0.165 | 0.044 | 0.049 | 0.052 |
| Paper2 | 0.067 | 0.036 | 0.040 | 0.037 | 0.042 | 0.018 | | 0.079 | 0.030 | | 0.045 | 0.046 | 0.019 | 0.068 | 0.054 | 0.025 | | 0.029 |
| Paper3 | 0.041 | 0.043 | | 0.017 | 0.025 | 0.006 | 0.021 | 0.031 | 0.015 | 0.012 | 0.041 | 0.097 | 0.011 | | 0.019 | 0.026 | 0.033 | 0.038 |
| Paper4 | 0.037 | 0.022 | 0.010 | 0.035 | | 0.009 | | 0.024 | 0.020 | 0.021 | 0.022 | 0.018 | 0.011 | | | | 0.023 | 0.014 |
| Paper5 | | | 0.014 | 0.013 | 0.018 | 0.040 | 0.011 | 0.051 | 0.011 | | 0.051 | 0.047 | 0.013 | 0.018 | 0.063 | 0.018 | | 0.026 |
| Paper6 | | 0.024 | | | 0.019 | | 0.018 | 0.035 | 0.025 | 0.029 | 0.027 | 0.059 | 0.023 | 0.014 | 0.041 | 0.022 | 0.034 | |
| Paper7 | 0.145 | 0.104 | | | 0.056 | | 0.013 | 0.130 | 0.030 | | | 0.036 | 0.012 | 0.027 | | 0.046 | 0.041 | 0.034 |
| Paper8 | 0.045 | 0.039 | 0.025 | 0.022 | 0.016 | 0.008 | 0.023 | 0.019 | | 0.026 | 0.038 | 0.032 | 0.015 | 0.033 | 0.019 | | 0.024 | 0.018 |
| Paper9 | 0.030 | 0.042 | 0.033 | 0.015 | 0.012 | 0.008 | 0.025 | | 0.028 | 0.021 | 0.036 | | 0.007 | 0.006 | 0.031 | 0.013 | 0.031 | 0.034 |
| Paper10 | 0.077 | 0.073 | 0.037 | 0.047 | | 0.058 | 0.013 | | 0.046 | 0.079 | 0.069 | 0.074 | | 0.048 | 0.058 | 0.032 | | 0.039 |
| Paper11 | 0.072 | 0.050 | 0.028 | 0.045 | 0.028 | 0.023 | 0.019 | 0.042 | 0.018 | 0.021 | 0.021 | 0.037 | 0.026 | 0.032 | 0.055 | 0.023 | 0.043 | 0.029 |
| Paper12 | | 0.084 | | | 0.039 | | 0.030 | 0.106 | 0.049 | 0.055 | 0.079 | 0.059 | 0.037 | 0.036 | 0.108 | 0.045 | 0.058 | 0.032 |
| Paper13 | 0.036 | 0.045 | 0.030 | 0.020 | 0.028 | 0.016 | | 0.027 | 0.041 | 0.058 | 0.018 | 0.034 | 0.027 | 0.033 | | 0.022 | 0.029 | 0.025 |
| Paper14 | 0.060 | 0.020 | 0.014 | 0.037 | | | 0.019 | | 0.047 | 0.058 | 0.019 | 0.064 | 0.021 | 0.045 | 0.094 | 0.015 | 0.043 | 0.060 |
| Paper15 | 0.096 | 0.100 | 0.063 | 0.065 | 0.045 | 0.054 | 0.133 | 0.063 | 0.054 | 0.119 | 0.102 | 0.079 | 0.041 | 0.051 | 0.126 | | 0.044 | 0.078 |
| Paper16 | 0.089 | 0.122 | 0.081 | 0.053 | 0.058 | 0.068 | 0.073 | | 0.048 | 0.078 | 0.067 | | 0.030 | | 0.132 | 0.033 | 0.064 | 0.089 |
| Paper17 | 0.042 | 0.035 | 0.019 | 0.031 | 0.030 | 0.017 | 0.019 | | 0.028 | 0.024 | 0.013 | 0.050 | | 0.016 | 0.023 | 0.021 | | 0.027 |
| Paper18 | | 0.084 | 0.033 | 0.030 | 0.023 | | 0.021 | | 0.028 | 0.043 | 0.016 | 0.039 | 0.014 | 0.015 | 0.013 | 0.018 | 0.026 | |
| Paper19 | 0.057 | 0.031 | 0.052 | 0.030 | 0.036 | 0.015 | 0.035 | 0.036 | 0.059 | 0.296 | 0.024 | 0.034 | 0.017 | 0.053 | 0.053 | 0.029 | 0.031 | 0.026 |
| Paper20 | 0.074 | 0.034 | 0.021 | 0.010 | 0.018 | 0.007 | 0.007 | 0.028 | 0.013 | 0.028 | 0.020 | 0.029 | 0.016 | 0.013 | 0.033 | 0.009 | 0.022 | 0.018 |
| Paper21 | 0.054 | 0.044 | 0.024 | | | | 0.024 | 0.032 | 0.017 | 0.048 | 0.027 | 0.062 | 0.016 | 0.030 | 0.040 | | 0.037 | |
| Paper22 | 0.045 | 0.075 | 0.063 | 0.047 | 0.040 | 0.052 | 0.046 | 0.060 | 0.064 | 0.044 | 0.057 | | | 0.047 | 0.103 | 0.034 | 0.045 | 0.027 |
| Paper23 | 0.034 | 0.025 | 0.030 | 0.044 | | 0.016 | 0.035 | 0.037 | 0.017 | | 0.025 | 0.055 | 0.013 | 0.030 | 0.042 | 0.019 | 0.040 | 0.031 |
| Paper24 | 0.014 | 0.013 | 0.005 | 0.030 | 0.011 | 0.008 | 0.024 | 0.014 | | 0.070 | 0.013 | 0.014 | 0.008 | | 0.040 | 0.010 | 0.015 | 0.018 |
| Paper25 | 0.008 | | 0.003 | 0.007 | | | 0.003 | 0.006 | 0.003 | 0.003 | | 0.013 | 0.002 | | 0.017 | 0.004 | 0.008 | 0.010 |
| Paper26 | 0.075 | 0.064 | 0.024 | 0.042 | 0.023 | 0.026 | | | 0.047 | 0.023 | | 0.047 | 0.015 | 0.019 | 0.074 | | 0.029 | 0.050 |
| Paper27 | 0.042 | 0.036 | 0.040 | 0.034 | 0.032 | 0.021 | | | 0.060 | 0.147 | 0.031 | 0.036 | 0.030 | 0.028 | 0.041 | 0.011 | 0.029 | 0.037 |
| Paper28 | 0.036 | 0.013 | | 0.009 | 0.013 | 0.011 | 0.042 | 0.039 | | | 0.009 | 0.034 | 0.010 | 0.036 | 0.047 | 0.026 | | 0.052 |
| Paper29 | 0.047 | | 0.043 | 0.019 | 0.033 | 0.017 | 0.016 | 0.100 | 0.011 | | 0.030 | 0.088 | 0.008 | 0.030 | 0.044 | 0.026 | | 0.050 |
| Paper30 | 0.034 | | 0.015 | 0.014 | | | | 0.027 | 0.015 | 0.029 | | 0.008 | 0.009 | 0.026 | 0.150 | 0.019 | 0.021 | 0.012 |
| Paper31 | 0.060 | 0.044 | 0.036 | 0.030 | 0.024 | 0.023 | 0.021 | 0.003 | 0.022 | 0.019 | 0.023 | 0.092 | 0.006 | 0.024 | 0.038 | 0.023 | 0.016 | 0.033 |

Values of AUC for each author under Jaccard and Cosine are shown in table 5.10 in combination with AUC for global weight covered in next subsection.

### 5.3.2 Using Global Weight

As discussed earlier, global weight is a cumulative score of every unique term considered in a summary for an author. Score is calculated by assigning some weight to each term used by that author in proportion to the fact that how frequently other authors have used the same term. For frequent term gives less information and so author with more global weight has less number of such terms. This can give a measure to say how comfortable is author with the subject as he generalises the paper and then write, rather than a person use trivial examples and is less technical [22].

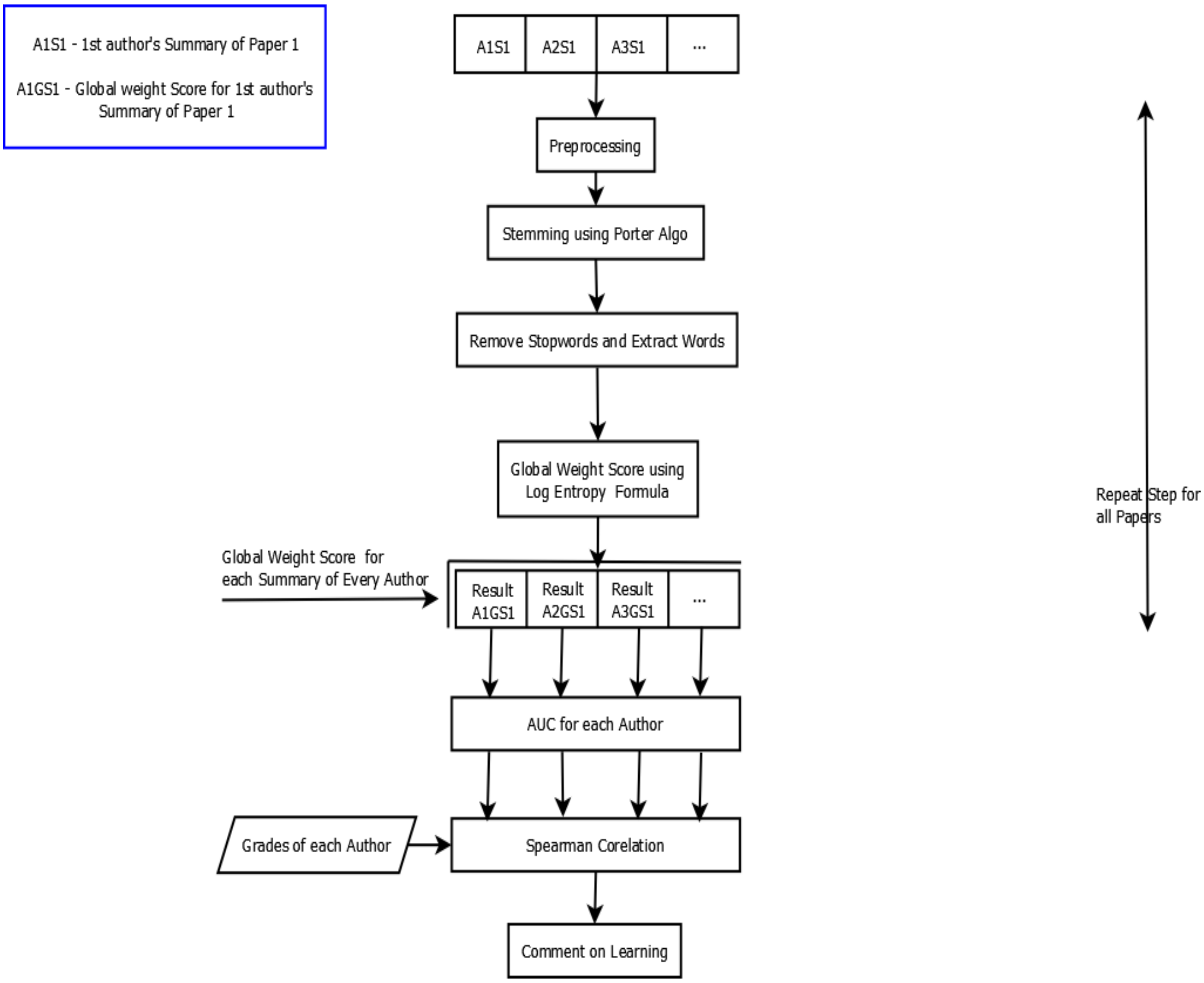
A1S1 - 1st author's Summary of Paper 1
A1GS1 - Global weight Score for 1st author's Summary of Paper 1
A1S1
A2S1
A3S1
...
Preprocessing
Stemming using Porter Algo
Remove Stopwords and Extract Words
Global Weight Score using Log Entropy Formula
Global Weight Score for each Summary of Every Author
Result A1GS1
Result A2GS1
Result A3GS1
...
AUC for each Author
Grades of each Author
Spearman Corelation
Comment on Learning
Repeat Step for all Papers

**Figure 5.10:** Similarity scores using Jaccard and Cosine and further relating their AUC to human graded assignments.

Figure above shows how global weight is calculated for every author and so as in case of similarity scores, author profile can be made containing all scores corresponding to an author for all his summaries. Further AUC is calculated, which can give one cumulative score for each author and the AUC's for all authors are correlated with human grades using Spearman's rank correlation.

Below is table 5.9 for global weight scores for every summary for each author, where paper1, paper2, paper3.....paper31 corresponds to all original papers. Places left blank signifies that author has not submitted summary for that paper.

**Table 5.9:** Global weight

| | | | | | | | | Name | of | Author | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Paper Name | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 | A9 | A10 | A11 | A12 | A13 | A14 | A15 | A16 | A17 | A18 |
| Paper1 | | 0.373 | 0.440 | 0.429 | 0.419 | 0.387 | 0.392 | 0.433 | 0.364 | 0.414 | 0.439 | 0.447 | 0.477 | | 0.519 | 0.411 | 0.432 | 0.504 |
| Paper2 | 0.464 | 0.330 | 0.441 | 0.450 | 0.388 | 0.516 | | 0.438 | 0.456 | | 0.453 | 0.373 | 0.446 | 0.514 | 0.555 | 0.400 | | 0.462 |
| Paper3 | 0.447 | 0.361 | | 0.390 | 0.373 | 0.333 | 0.400 | 0.433 | 0.432 | 0.436 | 0.367 | 0.404 | 0.508 | | 0.597 | 0.387 | 0.390 | 0.508 |
| Paper4 | 0.512 | 0.329 | 0.415 | 0.424 | | 0.452 | | 0.405 | 0.360 | 0.363 | 0.476 | 0.385 | 0.518 | | | | 0.375 | 0.521 |
| Paper5 | | | 0.377 | 0.463 | 0.359 | 0.345 | 0.413 | 0.447 | 0.388 | | 0.335 | 0.367 | 0.515 | 0.528 | 0.527 | 0.455 | | 0.473 |
| Paper6 | | 0.406 | | | 0.433 | | 0.444 | 0.487 | 0.338 | 0.380 | 0.425 | 0.385 | 0.568 | 0.596 | 0.443 | 0.496 | 0.417 | |
| Paper7 | 0.429 | 0.409 | | | 0.337 | | 0.459 | 0.472 | 0.391 | | | 0.295 | 0.488 | 0.532 | | 0.405 | 0.422 | 0.426 |
| Paper8 | 0.507 | 0.314 | 0.426 | 0.384 | 0.356 | 0.359 | 0.416 | 0.441 | | 0.380 | 0.374 | 0.322 | 0.521 | 0.496 | 0.567 | | 0.279 | 0.489 |
| Paper9 | 0.446 | 0.317 | 0.395 | 0.392 | 0.346 | 0.416 | 0.407 | | 0.382 | 0.367 | 0.357 | | 0.492 | 0.587 | 0.576 | 0.392 | 0.414 | 0.476 |
| Paper10 | 0.445 | 0.324 | 0.419 | 0.469 | | 0.369 | 0.464 | | 0.387 | 0.278 | 0.351 | 0.394 | | 0.539 | 0.516 | 0.435 | | 0.474 |
| Paper11 | 0.410 | 0.279 | 0.416 | 0.445 | 0.396 | 0.348 | 0.366 | 0.423 | 0.393 | 0.351 | 0.381 | 0.385 | 0.455 | 0.475 | 0.516 | 0.307 | 0.365 | 0.478 |
| Paper12 | | 0.301 | | | 0.425 | | 0.374 | 0.480 | 0.269 | 0.299 | 0.323 | 0.275 | 0.512 | 0.516 | 0.511 | 0.411 | 0.423 | 0.497 |
| Paper13 | 0.452 | 0.354 | 0.417 | 0.439 | 0.388 | 0.406 | | 0.415 | 0.276 | 0.394 | 0.409 | 0.341 | 0.513 | 0.492 | | 0.451 | 0.405 | 0.467 |
| Paper14 | 0.477 | 0.330 | 0.429 | 0.438 | | | 0.411 | | 0.324 | 0.307 | 0.455 | 0.354 | 0.475 | 0.532 | 0.507 | 0.396 | 0.411 | 0.484 |
| Paper15 | 0.436 | 0.284 | 0.427 | 0.430 | 0.351 | 0.446 | 0.333 | 0.398 | 0.436 | 0.365 | 0.343 | 0.304 | 0.396 | 0.559 | 0.545 | | 0.471 | 0.542 |
| Paper16 | 0.423 | 0.290 | 0.432 | 0.411 | 0.323 | 0.284 | 0.403 | | 0.416 | 0.313 | 0.377 | | 0.508 | | 0.554 | 0.367 | 0.408 | 0.483 |
| Paper17 | 0.493 | 0.381 | 0.455 | 0.446 | 0.361 | 0.302 | 0.383 | | 0.366 | 0.345 | 0.391 | 0.376 | | 0.578 | 0.498 | 0.440 | | 0.534 |
| Paper18 | | 0.365 | 0.429 | 0.451 | 0.383 | | 0.393 | | 0.401 | 0.327 | 0.385 | 0.410 | 0.476 | 0.543 | 0.604 | 0.415 | 0.403 | |
| Paper19 | 0.478 | 0.367 | 0.409 | 0.458 | 0.360 | 0.278 | 0.396 | 0.480 | 0.360 | 0.519 | 0.482 | 0.410 | 0.446 | 0.479 | 0.538 | 0.419 | 0.490 | 0.445 |
| Paper20 | 0.477 | 0.320 | 0.438 | 0.455 | 0.369 | 0.334 | 0.375 | 0.478 | 0.313 | 0.418 | 0.464 | 0.345 | 0.490 | 0.488 | 0.514 | 0.384 | 0.370 | 0.444 |
| Paper21 | 0.481 | 0.293 | 0.423 | | | | 0.476 | 0.417 | 0.353 | 0.295 | 0.325 | 0.312 | 0.499 | 0.520 | 0.383 | | 0.315 | |
| Paper22 | 0.473 | 0.330 | 0.414 | 0.427 | 0.376 | 0.386 | 0.432 | 0.423 | 0.373 | 0.328 | 0.349 | | | 0.506 | 0.473 | 0.479 | 0.458 | 0.460 |
| Paper23 | 0.465 | 0.369 | 0.451 | 0.401 | | 0.420 | 0.444 | 0.533 | 0.367 | | 0.456 | 0.331 | 0.414 | 0.479 | 0.536 | 0.423 | 0.501 | 0.451 |
| Paper24 | 0.438 | 0.363 | 0.446 | 0.389 | 0.409 | 0.282 | 0.380 | 0.451 | | 0.482 | 0.345 | 0.452 | 0.375 | | 0.555 | 0.337 | 0.470 | 0.486 |
| Paper25 | 0.462 | | 0.417 | 0.409 | | | 0.356 | 0.377 | 0.394 | 0.366 | | 0.413 | 0.481 | | 0.613 | 0.369 | 0.370 | 0.454 |
| Paper26 | 0.545 | 0.289 | 0.442 | 0.506 | 0.401 | 0.364 | | | 0.485 | 0.314 | | 0.401 | 0.437 | 0.570 | 0.623 | | 0.378 | 0.480 |
| Paper27 | 0.461 | 0.279 | 0.429 | 0.411 | 0.422 | 0.366 | | | 0.325 | 0.480 | 0.389 | 0.399 | 0.430 | 0.475 | 0.488 | 0.382 | 0.476 | 0.486 |
| Paper28 | 0.481 | 0.288 | | 0.446 | 0.377 | 0.398 | 0.444 | 0.457 | | | 0.456 | 0.377 | 0.475 | 0.487 | 0.466 | 0.413 | | 0.474 |
| Paper29 | 0.462 | | 0.438 | 0.449 | 0.334 | 0.395 | 0.379 | 0.463 | 0.390 | | 0.411 | 0.419 | 0.519 | 0.527 | 0.553 | 0.417 | | 0.449 |
| Paper30 | 0.484 | | 0.413 | 0.384 | | | | 0.423 | 0.451 | 0.383 | | 0.426 | 0.510 | 0.516 | 0.557 | 0.446 | 0.337 | 0.519 |
| Paper31 | 0.498 | 0.341 | 0.388 | 0.398 | 0.358 | 0.341 | 0.423 | 0.596 | 0.350 | 0.298 | 0.374 | 0.372 | 0.491 | 0.514 | 0.512 | 0.381 | 0.359 | 0.431 |

ROC curves for Jaccard, Cosine and Global weight scores were generated following the measures described in [11, 12]. Following are the curves for 5 highest and 5 lowest scorers.

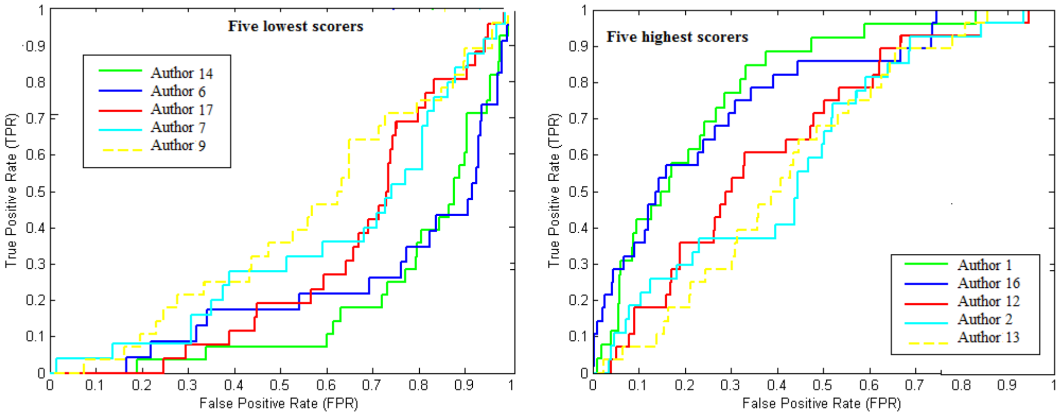

**Figure 5.11:** ROC curve for highest and lowest Jaccard similarity scorers.

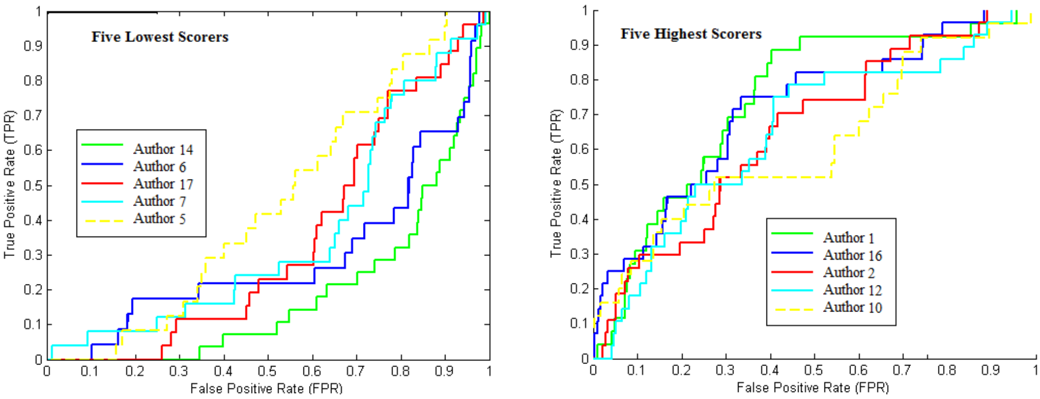

**Figure 5.12:** ROC curve for highest and lowest Cosine similarity scorers.

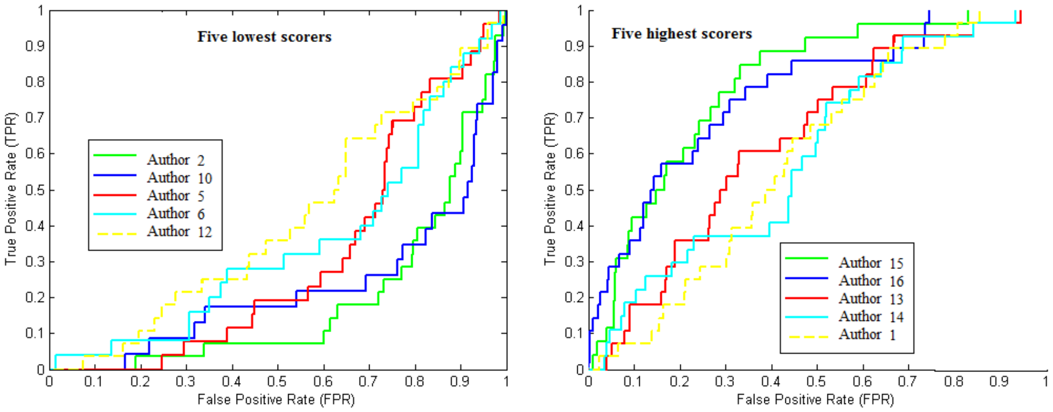

**Figure 5.13:** ROC curve for highest and lowest Global weight scorers.

### 5.3.3 Comparing Jaccard Cosine and Global scores with Human grades

**Table 5.10:** Values of AUC scores against human grades

| Name of author | AUC for each author for Jaccard Distance of all submitted summaries. | AUC for each author for global weight of all submitted summaries. | Grades given by instructor | Equivalent Numeric grade |
|---|---|---|---|---|
| Author 1 | 0.792064595 | 0.736673034 | A+ | 11 |
| Author 2 | 0.601390984 | 0.105161723 | A- | 9 |
| Author 3 | 0.45858656 | 0.521705157 | A | 10 |
| Author 4 | 0.494134406 | 0.545165075 | A | 10 |
| Author 5 | 0.43071161 | 0.263951311 | A | 10 |
| Author 6 | 0.223825307 | 0.269448236 | B | 8 |
| Author 7 | 0.35027027 | 0.419009009 | A | 10 |
| Author 8 | 0.584031975 | 0.638623513 | A | 10 |
| Author 9 | 0.414237123 | 0.284823453 | A | 10 |
| Author 10 | 0.582252252 | 0.258198198 | B | 8 |
| Author 11 | 0.483157366 | 0.372381431 | A | 10 |
| Author 12 | 0.642695173 | 0.273485585 | A- | 9 |
| Author 13 | 0.588921283 | 0.787252996 | A | 10 |
| Author 14 | 0.187965662 | 0.779964367 | A | 10 |
| Author 15 | 0.499279279 | 0.91981982 | A- | 9 |
| Author 16 | 0.767330742 | 0.915127956 | A | 10 |
| Author 17 | 0.309776003 | 0.436186838 | A | 10 |
| Author 18 | 0.552612613 | 0.424864865 | A+ | 11 |

# 6

# Discussion

In this work on the aspect of driving best setting for the collected corpus, decision tree classifier with word n gram 4 feature gave interesting results with micro precision of 0.971. This can be due to the fact that summaries has certain set of questions, but the style of introducing question for each author is different. Word n gram 4 can catch such content specific style for each author.For all other classifiers except Burrows Delta, word n gram with variable value of n showed highest results. This can well be understood as word n grams have been usually confirmed to cover stylistic and content-based text.

Experiments on training corpus showed the following remarkable points:

(a) for only burrows delta character came at top and for rest word n gram performed well.

(b) Unlikely in many cases, function word did not performed well. This might be because of the constrictive nature of summary.

(c) Although, the text were really small(words per text), the performance of classifier was very satisfactory.

Experiment was conducted on the test summaries, classifier implemented was Decision tree with word n gram 4 as the feature to do final attributions. As students are largely writing their own, this can reflect their confidence in the course and interest. Also, when they are writing it by themselves this can be assumed that learning is actually happening at some point.

As in our test cases, no similar summaries were there for training, so we believe that our results are satisfactory as the content is not the deciding criteria and style is being considered as well.

With regards to learning measures, Jaccard had maximum AUC score for author 1 with value 0.792064595 and Cosine had maximum score with value 0.91981982 for author 16. In terms of learning, Spearman rank correlation gave positive result with highest 'R' value of 0.39049 for global weight and actual human grades amongst all three measures i.e. Jaccard, Cosine and Global weight . We cannot comment that two measures that we used were by far the best measures that can be compared with human grades, but they were able to conclude to learning aspect to some point.

# 7

# Future Work and Limitations

In our research, need to collect ground truth is necessary. For case where ground truth is not available, unsupervised machine learning techniques can help. Above situation with no ground truth comes with a classic problem that what if student has not written any of the summaries. In current scenario, strict formatted summaries regarding to one topic are considered. In future, can there be an automated system which can comment for a particular author regardless of the topic.

# References

[1] E. Stamatatos. **A Survey of Modern Authorship Attribution Methods**. *Journal of the American Society for Information Science and Technology*, **60**(3):538–556, 2009. vi, 3, 5, 11, 12, 13, 14, 15

[2] H. Davis J. Wintrup, K. Wakefield. **Engaged learning in MOOCs: a study using the UK Engagement Survey**. pages 1–72, 2015. 1, 2

[3] W. D. Milheim. **Massive Open Online Courses (MOOCs): Current Applications and Future Potential**. *Educational Technology*, **53**(3):39–42, 2013. 1

[4] T. Ellis W. Hafner. **Work in progressAuthenticating authorship of student work: Beyond plagiarism detection**. *Proc. 35th Frontiers in Education Conf., Indianapolis, IN*, pages F1H – 25–6, Oct, 2005. 2, 3, 5

[5] K. Mangan. **MOOC Mania. The Chronicle of Higher Education [Special Report: Online Learning]**. page B4B5, 2012. 2

[6] I. G-Veiga J. I. M-Cordero G.J-Botana, J. M. Luzon. **Automated LSA Assessment of Summaries in Distance Education: Some Variables to Be Considered**. *Journal of Educational Computing Research*, **52**:341–364, 2015. 3, 6, 32

[7] S. Argamon M. Koppel, J. Schler. **Computational Methods in Authorship Attribution**. *Journal of the American Society for Information Science and Technology*, **60**:9–26, 2009. 3, 6

[8] M. Hoey L. H. ke. **Strategies of writing summaries for hard news texts: A text analysis approach**. *SAGE*, **16**(1):89–105, 2014. 5

[9] D. Challis. **Committing to quality learning through adaptive online assessment**. *Assessment Evaluation in Higher Education*, **30**(5):519 – 527, 2005. 5

[10] T. Lancaster. **The Application of Intelligent Context-Aware Systems to the Detection of Online Student Cheating**. *Seventh International Conference on Complex, Intelligent, and Software Intensive Systems*, 2013. 5

[11] E. Stamatatos P. Juola. **Overview of the International Authorship Identification Competition at PAN-2013**. *Proceedings of fifth international workshop on uncovering plagiarism, authorship, and social software misuse*, 2013. 6, 11, 37

[12] W. Daelemans E. Stamatatos B. Stein M. A. S-Perez A. B-Cedeo P. Juola, M. Potthast. **Overview of the Author Identification Task at PAN 2014**. *Proceedings of Sixth international workshop on uncovering plagiarism, authorship, and social software misuse*, 2014. 6, 37

[13] W. Daelemans K. Luyckx. **The Effect of Author Set Size and Data Size in Authorship Attribution**. *Literary and Linguistic Computing*, **26**:35–55, 2011. 6

[14] M. Kahani R. Ramezan, N. Sheydaei. **Evaluating the Effects of Textual Features on Authorship Attribution Accuracy**. *3rd International Conference on Computer and Knowledge Engineering.* 6

[15] P. Rosso A. Barron-Cedeno. **On automatic plagiarism detection based on ngrams comparison**. *Advances in Information Retrieval.* 6

[16] B. C.M. Fung M. Debbabi F. Iqbal, H. Binsalleeh. **A unified data mining solution for authorship analysis in anonymous textual communications**. *Information Sciences*, **231**:98–112. 11

[17] **Retrieved from http://www.sequencepublishing.com/academic.html**. 13, 20

[18] O. Pedersen Y. Yang. **A comparative study on feature selection in text categorization**. *ICML*, 1997. 15

[19] E. Stamatatos J. Houvardas. **N-Gram Feature Selection for Authorship Identification**. *Artificial Intelligence: Methodology, Systems, and Applications*, **4183**:77–86, 2006. 15, 20

[20] A. Huang. **Similarity measures for text document clustering**. *Proceedings of the Sixth New Zealand Computer Science Research Student Conference*, **179**:49–56, 2008. 17

[21] P. Juola S. Argamon. **Overview of the International Authorship Identification Competition at PAN-2011**. *Proceedings of international workshop on uncovering plagiarism, authorship, and social software misuse*, 2011. 23

[22] S. T. Dumais T. K. Landauer. **A solution to Platos problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge**. *PSYCHOLOGICAL REVIEW.* 36