Journal of Scientometric Res. 2019;8(2s):s7-s38
http://www.jscires.org

Analyzing the Common Wisdom of Binarization Doctrine in Internationality Classification of Journals: A Machine Learning Approach

Gambhire Swati Sampatrao[1], Sudeepa Roy Dey[1*], Abhishek Bansal[2], Sriparna Saha[2]

[1]Department of computer science and engg, PESIT-BSC, Bangalore-560100 (affiliated to Visvesvarya Technical University, Belagavi), Karnataka, INDIA.
[2]Department of Computer Science, IIT Patna, Bihar, INDIA.

ABSTRACT

Evaluating and identifying "Internationality" of peer reviewed journals is a hotly debated topic. The problem broadly focuses on whether a journal is international or not, indicating a strong tilt toward binary classification doctrine. The manuscript investigates the doctrine, for the first time. The authors have validated their study further by using minimum error rate classifier, investigated theoretical lower and upper bounds of classification error in the context of internationality. The novel approach has rich ramifications in Scientometrics. Further, we propose a new principle of classification that results in greater accuracy fortifying the assertion.

Keywords: Journal Internationality, Binary classification, Bayesian Error rate function, web scraping, supervised learning, Non-Local Influence Quotient (NLIQ), Source-Normalized Impact per Paper (SNIP), Other Citation Quotient (OCQ), Unified Granular Neural Network (UGNN).

Correspondence
Sudeepa Roy Dey
Department of computer science and engg, PESIT-BSC, Bangalore-560100 affiliated to Visvesvarya Technical University, Belagavi, Karnataka, INDIA.
Email: sudeepar@gmail.com

Received: 09-01-2019
Revised: 11-06-2019
Accepted: 16-09-2019
DOI: 10.5530/jscires.8.2.22

INTRODUCTION

Defining and measuring internationality as a function of influence diffusion of scientific journals is an open problem. Until last year, there was no metric to rank journals based on the extent or scale of internationality.[1] Measuring internationality is qualitative, vague, open to interpretation and is limited by vested interests. With the tremendous increase in the number of journals in various fields and the unflinching desire of academicians across the globe to publish in "international" journals, it has become an absolute necessity to evaluate, rank and categorize journals based on internationality. In recent times, various authors and research scholars have been exploring means to find suitable and reputed "international" journals for publication of their research work. The drive behind this is to own appreciation or award for the quality work that they do. Also, institutional assessment and evaluation depends heavily on peer-reviewed publications, academia or research labs alike. Thus, evaluating internationality is an open problem owing to the fact that such journals are vast in number; a plethora of such entities claim to be international but citations, influence and other indicators are a bare minimum.

Data collected from IEEE Xplore in the year 2009 showed an exponential increase of 25% in international journal publications (www.journalmetrics.com), when compared with previous years. A study conducted by Buchandiran[2] reveals an enormous increase in publication of journals between the years 2004 and 2009, whereby in the year 2009, 6,132 Indian institutions have contributed 23,745 papers out of which 15,880 were from academic institutions. This clearly shows that academic institutions contribute to the majority of such published work. Elsevier's Scopus covered 15,376 publications till 2014 and Thomson Reuters Web of Science covered 8,262 publications in the same field. The raw numbers are encouraging but pose serious challenges as most of the journals claim to be "international". To aggravate the landscape further, predatory publishers who unethically and unprofessionally exploit the open access publishing model for financial rewards, have crowded the scientific publishing space. This is a period of publication explosion with nearly 2.5 million new scientific papers being published each year. The increasing pressure on authors and research scholars to publish in international journals or perish have lead to this deluge. The anxiety of the scholars are aptly put to use by various predatory publishers

hiding behind the title International for attracting submissions by giving impressions of Internationality through origin of editors, ISSN or editorial members rather than the quality. Unfortunately, the authors of the manuscript could not find unambiguous guidelines or scholarly references to the "internationality classification of journals". The popular perception encourages us to believe in the doctrine that "A journal is either International or NOT!" advocating binary discrimination philosophy. This paper dispels the fallacy of national and international journals by using supervised Machine Learning techniques like binary classification and also investigates the premises of classification by posing the following questions

1. Is journal classification by internationality a binary problem or much more complex and granular?
2. Is journal labeling based on internationality a post-facto analysis?
3. How reasonable is an "a-priori" perceptive understanding of journal internationality index based on non quantifiable information?

The proposed methodology will test the hypothesis of binarization based on a Bayesian approach. In machine learning, classification is the problem of identifying belongingness of a new observation to a set of categories, on the basis of a training set of data containing observations (or instances) whose category membership is known a-priori. If instances are given with known labels, then the learning is called supervised learning and the data set is known as training set. When the instances are classified into two categories based on whether an instance has some qualitative property (features), it is known as binary or binomial classification. Binary classification generally falls into the domain of supervised learning since the training dataset is labeled. In binary classification, only 2 classes are involved i.e. instances can be assigned to one of the two classes. If we pose the problem of classifying journals as national or international based on internationality index, this binary classification logic would be fallacious. There must be proper discrimination between international and national journals (rather a graded and much finer classification paradigm among International journals), guided by the principles of statistical Machine Learning aided and abetted by the features. These features or indicators are well studied in.[3-8]

Issues in existing binarization doctrine: The existing classification methods of journals as national and "international" is merely based on some thumb rules such as: Impact factor, nationality of journal and H-index. Such doctrine is not reliable and many a times lead to blurry conclusions and easy manipulations. The authors listed a few pointers in this direction that raise alarm and investigates the existing doctrine being practiced.

- Cases where Jeaffrey Beall classified Frontier journals as predatory since they are open access!! Is the problem, journal classification, so naive? The authors respectfully disagree.
- The habitability problem, before major quantification initiatives were taken up, was not posed as a binary problem. It could easily have been formulated as exoplanets being habitable or not.[9] But it was not, posed instead as a three class problem and later modified to be formulated as multi (more than three) class problem.
- Any one feature or two are not adequate to solve such a problem, rather the problem is non linear and dependent upon the complex dynamics of multiple features and class annotations
- The solution best case scenario within such a binarization doctrine may throw up some insight into the reliability of such a doctrine

We adopt a machine learning driven approach to tackle the problems stated above. The field of Machine Learning integrates many distinct approaches such as probability theory, logic, combinatorial optimization, randomized search, statistics, reinforcement learning and control theory well explained in.[10-12] The anatomy of a typical machine learning problem relies on data as input and the learning algorithm to produce a model as output. Predominantly, there are two kinds of Machine Learning algorithms: Supervised and Unsuper-vised. Supervised learning is the machine learning task of inferring a function from labeled training data. The training data consist of a set of training examples. On the contrary, unsupervised learning allows the algorithm divide the input dataset into subsets and clusters them depending on its own similarity computation. There is no label associated with the input.

Binary Classification would generally fall into the domain of Supervised Learning since the training dataset is labeled. As the name suggests, it is simply a special case in which there are only two classes. Some typical examples include credit Card Fraudulent Transaction detection (Fraud or no fraud), medical diagnosis and Spam detection (Spam and regular email). There exists various paradigms that are used for learning binary classifiers. These include Decision Trees, Neural Networks, Bayesian Classification and Support Vector Machine among other methods. Multi-class classification deals with the scenario where each training point belongs to one of the N different classes. The goal is to construct a function which, given a new data point, will correctly predict the class the new point belongs to. Some typical examples include handwriting recognition, sentiment analysis and document classification.

PROBLEM DEFINITION

Journal internationality is broadly classified in to two categories: journals are *perceived* to be International or National. This nomenclature is not solidly founded upon quantitative analysis since measures of internationality and the required metrics for such measure are either not clear or debatable due to lack of statistical evidence. The absence of clear quantifiers regarding the binary doctrine of internationality is a good enough reason to investigate the doctrine of "binarization". The paper is inspired by an earlier work where inter-nationality of journals was quantified and modeled for the first time.[13] However, in contrast to the cited scholarly work, the current manuscript dwells on the binary classification problem from a machine learning perspective, without considering the explicit inter-nationality score of any journal, as proposed by Ginde. *et al*.[13] The research problem is settled in the following shape:

- Journal internationality is a riveting topic as thousands, if not millions, of journals claim to be international. Scientific inquiry, let alone discussion, is not possible without a rigorous, quantitative analysis of the doctrine of binarization.
- The significance of internationality quantification seems prudent as the other option of considering journals published from Mexico or Canada while geophysically located in the North Americas is simply not Science!
- The idea of classifying scholarly journals into two classes may be an oversimplification, since internationality score is granulated, as demonstrated in.[13]
- For the same reason stated above, the binary classification doctrine seem unfair as some journals may lose out marginally.
- For all of the above reasons, the binary classification *"theory"* must be put to stringent scrutiny by adopting a machine classification approach.
- Some ground rules founded on sophisticated mathematical and modeling principles are in order to avoid frequent misclassification of journals based on internationality. To serve that purpose well, the binarization hypothesis, if it is a major impediment to that process, need to be studied in detail.
- Some baseline for future debate and scientific critique regarding journal internationality needs to be drawn.
- Finally, some groundwork must be laid to explore an alternative, multi-class discrimination paradigm for *journal internationality*. This is, however, beyond the scope of the current manuscript and marked as future work.

Remark: One might argue the literature on existing classification approaches is basically missing. This is where the manuscript intends to score. There is no comparative studies available on internationality classification problem. There is a hand-waving theory that journals are International or National. The literature on existing classification approaches is basically missing because there is no literature. No classification approach has ever been tried on the journal interna-tionality problem. There is no machine learning based classification approach to compare with. Our approach borders on a completely new postulate and that statement is quite categorically established through the remainder of the text.

OUR CONTRIBUTION

We aim to accomplish and show the following using Bayesian minimum error rate classifier and perform theoretical estimation of error bound.

Technical contribution

- Construct decision rules by converting a priori class probability ω_i into a measurement conditioned probability, $P(\omega_i/x)$. This is known as the posterior probability.
- Formulate a measure of expected classification risk, known as classification error.
- Choose decision rule that minimizes the risk i.e the risk of misclassifying a national journal as international journal and vice versa.
- Establish bounds of risk and show that there exists an one-dimensional risk bound for the multidimensional density function under the assumption of Gaussian distribution.

Scientometric contribution

- Under the risk-minimal scheme of classification of journals, identify journals which are claimed as national or inter-national wrongly and establish the hypothesis that journal classification based on internationality should not be posed as a binary problem.
- Establish the fact that Internationality of a journal is not orthogonal to quality, impact and influence. The grading parameters are different. The process of evaluation inter-nationality intersects with journal influence on academic community. Therefore, instead of classifying journals as National or International, we should classify internationality in to different grades, similar to what Scimagojr does to the quality grading (Q1- Q4) of journals.[14] The interna-tionality gradation and ranking is not present in theory or practice, even by the leaders such as Clarivate Analytics and Scimago Labs.

It is imperative to rather admit that the classification problem is more granular. Therefore, more than two classes based on features and data seem fair. We set up the problem as a two class discrimination and demonstrate the fallacy of such scheme. *This approach is a probable fundamental change in the way internationality of journals is studied. Beyond reasonable doubt, this is our greatest contribution in terms of originality.*

The remainder of the paper is organized as follows. We present a few standard definitions and metrics, commonly used and widely known. Next, the flaws and limitations are discussed followed by remedial and novel metric definitions. Finally, the methodology adopted and fallacy of classification doctrine are justified. The flow of arguments between sections is interlaced with theoretical bound estimation.

BASIC CONCEPTS, DEFINITIONS, REMEDIAL METRICS

4.1 Basic concepts:

- **Linearly separable:** The two sets are linearly separable if there exist at least one line in the plane which divides data points. In statistics and machine learning, classifying certain types of data is a problem for which good algorithms exist that are based on this concept.
- **ShapiroWilk test:** The ShapiroWilk test is a test of normality in frequentist statistics. It was published in 1965 by Samuel Sanford Shapiro and Martin Wilk. The ShapiroWilk test utilizes the null hypothesis principle to check whether a sample x_1, ..., x_n from a normally distributed population. The following test statistic is used.

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

where x_i: *is the i-th order statistic, i.e., the i-th smallest number in the sample;*

$\bar{x} = (x_1 + \ldots\ldots + x_n)/n$ *is the sample mean. The constants* a_i *are given by*

$$(a_1, \ldots., a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{1/2}}$$

where $m = (m_1, \ldots., m_n)^T$ and $m_1, \ldots., m_n$ are the expected values of the order statistics of independent and identically distributed random variables sampled from the standard normal distribution, and V is the covariance matrix of those order statistics.

- **Minimum Error Rate Classifier:** In classification problems, each state of nature is associated with a different one of the classes, and the action a_i is usually interpreted as the decision that the true state of nature is w_i. If action a_i is taken and the true state of nature is w_i then the decision is correct if $i = j$ and in error if $i \neq j$. If errors are to be avoided, it is natural to seek a decision rule that minimizes the probability of error i.e. the error rate.
- **NLIQ:** Papers published in one journal cite papers from the same journal much more often than those from different journals, regardless of the journals SNIP value. This, too, leads to a cycle wherein an individual journals prestige is increased by virtue of increased citations from within. It should be noted that journals of higher SNIP value have a lower NLIQ value compared to journals of lower SNIP value meaning citations are mostly restricted to the same journal they originate from. This in no way implies that there is a correlation between the two;[15] merely revealing that journals most people would consider to be highly ranked (i.e by having higher SNIP values) exhibit only a low level of non-local influence. Evidently, information about the internationality of these journals is incomplete – whether the authors are from the same institution or the same country or merely citing their previous works or those of colleagues due to reciprocity, as previously mentioned – is not known. Echoing the definition from Ginde *et al*,[1] NLIQ is the number of citations made by articles published in a journal X to articles published in different journals divided by the total number of citations made by all papers in that journal X. Clearly, higher the number of external citations made by articles in a journal, higher the NLIQ of that journal.
- **OCQ (Other Citation Quotient):** reflects a journals integrity owing to the fact that no legitimate journal will promote authors and allow them to indiscreetly cite their own work.
- **H Index:** The h index expresses the journals number of articles (h) that have received at least h citations
- **ICQ (International collaboration quotient)** accounts for the articles that have been produced by researchers from several countries. In order to compute this parameter, we first extracted the country information of the journal and then the author affiliations for each one of the published articles in that journal. Every authors country is matched with the country of publishing journal. Ratio is calculated on the basis of weights assigned to different combination of authors affiliation and origin of the publishing journal

These four parameters, NLIQ, OCQ, ICR and H-index constitute the feature vector, X in the classifier design.

METHODOLOGY

Let us begin by accepting the binarization doctrine and analyze a best case solution scenario suitable to validate the doctrine; the outcome of such solution scheme (classification scheme) may support the doctrine or may produce counter examples! In case we encounter counter examples, given that there may not exist a better scheme to turn around the counter examples thus produced, it is reasonable to doubt the doctrine. Discussing the fallacy of the doctrine would be the next step (evidently no clear demarcation between National and International journals). Thus by invalidating the doctrine of binary classification, we justify the problem as multi-class and construct a baseline for future research i.e propose the possibility of constructing a multi-class classifier by using deep learning or gradient boosted methods. These methods, hopefully shall capture the inherent complexity of such classification problem.

We propose a framework for binary classification of journals and further prove that, due to granularity of internationality indicator values, journals should be classified into three classes of internationality namely, Low, Moderate and High. This would be a multi-class classification problem. Figure(1) shows the steps to follow for binary classification of journals as national or international.

X is a d dimension feature vector i.e $X \in R^d$ belonging to the feature space which is *the space of Scientometric indicators.*

We construct a ***minimum error rate classifier*** which helps us in deciding the internationality or nationality of journal on the basis of values of different features.The features for journal we consider are OCQ (Other Citation Quotient), NLIQ (Non Local Influence Quotient), IC (International Collaboration Ratio) and HINDEX (defined in the previous section).

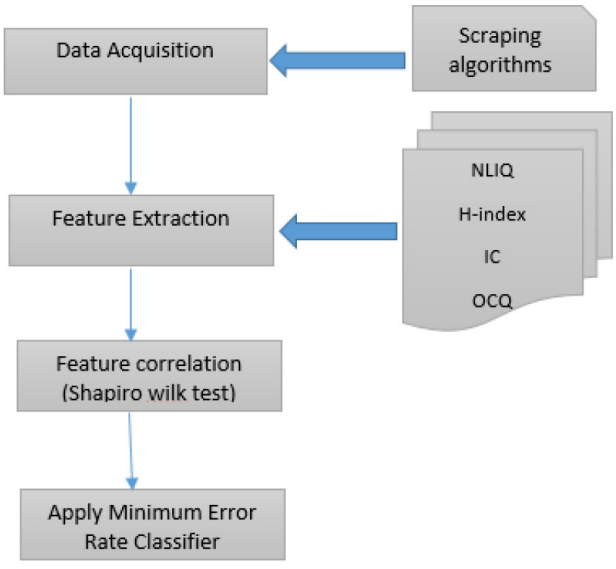

Figure 1: A high level view of methodology.

Proposed Model

Minimum Error Classifier

Let us design an autonomous classification scheme to identify two types of journals, based on internationality, an indicator hard to quantify, until recently by.[15] Supervised classification works with samples from different classes. We consider some samples from each of the classes and make a set of samples known as training set. The class labels of each sample are know a-priori. A measurable quantity is considered as a feature. Let X be a feature vector. X can be one dimensional or multidimensional. Let us consider IC and NLIQ to begin with, to construct 2 dimensional feature space, which may be extended by including the features OCQ and H-index. We shall illustrate the feature extraction later in the algorithm section. The selected features should give good discrimination between two classes. The selected features mentioned above are typical of the type of journals. However some entities in both classes do exist which are atypical. We hope that such cases don't happen too often.

We consider some samples from every class ω_i and measure the value of X for every class and compute the probability density function (pdf) of X for every class ω_i i.e $p(X \mid \omega_i)$ where i varies from 1 to c, the total number of classes. If any unknown sample needs to be identified, we have to measure the feature X for that sample and take the decision in favor of an appropriate class. By probability theory,

$$p(\omega_i, x) = p(\omega_i \mid X)_p(X) \tag{1}$$

or

$$p(\omega_i, X) = p(X \mid \omega_i)P(\omega_i) \tag{2}$$

action		state/class		associated probability
α_1	$\Rightarrow$	ω_1	$\Rightarrow$	$P(\alpha_1 \mid \omega_1)$
α_2	$\Rightarrow$	ω_2	$\Rightarrow$	$P(\alpha_2 \mid \omega_2)$

where $X = [x_1\ x_2 \ldots . x_d]^t$; d is dimension of feature vector; $p(\omega_i \mid X)$ is the probability that class is ω_i given the feature x and $p(X \mid \omega_i)$ is the class conditional density function for class ω_i.

Using equations (1) and (2), we obtain

$$P(\omega_i \mid X)p(X) = p(X \mid \omega_i)P(\omega_i)$$
$$P(\omega_i \mid X) = p(X \mid \omega_i)P(\omega_i)/p(X)$$

where $P(\omega_i \mid X)$ is the posteriori probability; $P(\omega_i)$ is the prior probability; $p(X) = \sum_{j=1}^{c} p(x \mid \omega_i)P(\omega_i)$ and c is the total number of classes. *Under this setting,* ***Bayes decision rule*** *states that we construct a decision in favor of the class having maximum posterior probability.*

If we consider the two class case (binary), then

$$P(\omega_1 \mid X) > P(\omega_2 \mid X) \quad \Rightarrow \text{class } \omega_1$$
$$P(\omega_1 \mid X) < P(\omega_2 \mid X) \quad \Rightarrow \text{class } \omega_2$$

If we assign any new unknown sample to any one class then there is finite probability of its association to other class. This indicates the measure of error in the *Bayes Decision Rule.* It is also common knowledge that error in classification will crop up whenever features overlap. Therefore the classification risk, P(error) indicating the likelihood of an incorrect decision, is introduced. If for any unknown sample X, $P(\omega_2 \mid X) > P(\omega_1 \mid X)$ then the classifier may decide in favor of class ω_2 producing an error of the form $P(\omega_1 \mid X)$. OTOH, for any unknown sample X, $P(\omega_2 \mid X) < P(\omega_1 \mid X)$ then the classifier takes decision in favor of class ω_1 and $P(\omega_2 \mid X)$ is the error. We construct a Bayesian decision rule that ensures the error be minimized. This is known as the ***Minimum Error Classifier.***

$$P(\text{error}) = \int P(\text{error}, X)dx$$
$$P(\text{error}) = \int P(\text{error} \mid X)p(X)dx$$
$$P(\text{error}) = \int \min[P(\omega_1 \mid X), [P(\omega_2 \mid X)]p(X)dX$$

If $P(\omega_1 \mid X) = P(\omega_2 \mid X)$, the classifier fails to arrive at a decision. Please note, *even if the posterior densities are continuous (we assume this to be true), this form ofconditional error makes the integrand discontinuous when the full error is computed. This is the reason, bounds of error(risk) need to be evaluated.*

5.1.2 Minimum Risk Classifier

Some loss is always incurred if classifier takes a wrong decision regarding the classification task. We define this loss by formulating a loss function $\lambda(\alpha_i \mid \omega_j)$. Loss function is more general compared to the probability of error. It quantifies the loss incurred in taking an action α_i when the action (or class-*National or international Journal, in our case)* is ω_j. The action α_i implies assigning unknown samples to one of the classes ω_i among all the *c* classes where *c* is the total number of classes. **We may represent loss function** $\lambda(\alpha_i \mid \omega_j)$ as λ_{ij}. *Expected Loss can be defined as*

$$R(\alpha_i \mid X) = \sum_{j=1}^{c} \lambda_{ij} P(\omega_j \mid X)$$

where ω_j (we don't know the true action(or class)) is the true action (or class), $R(\alpha_i \mid X)$ is the expected loss or risk function known as conditional risk, α_i is the action in favor of ω_i, X is the feature vector of dimension *d*, $P(\omega_j \mid X)$ is the posterior probability and *c* is the total number of classes. λ_{ij} should be equal to zero when *i* = *j* because we take the action α_j (i.e we assign unknown sample to class ω_j) and the correct action or class is ω_j. However, as noted in minimum error rate classifier, there exists some non zero probability of taking action in favor of the other class in a binary setting. Let us consider a binary classification problem i.e *c* = 2 and ω_1 and ω_2 are the two classes.

$$R(\alpha_1 \mid X) = \lambda_{11} P(\omega_1 \mid X) + \lambda_{12} P(\omega_2 \mid X)$$
$$R(\alpha_2 \mid X) = \lambda_{21} P(\omega_1 \mid X) + \lambda_{22} P(\omega_2 \mid X)$$

If $R(\alpha_1 \mid X) > R(\alpha_2 \mid X)$, we assign the unknown sample to the class ω_2 else if $R(\alpha_1 \mid X) < R(\alpha_2 \mid X)$, we assign the unknown sample to the class ω_i.

This type of a decision rule works on the principle that if we take decision in favor of that class unknown samples are assigned to that minimizes the risk in the process. *This type of classifier is known as* ***Minimum Risk Classifier.***

Let us consider the case when $R(\alpha_1 \mid X) > R(\alpha_2 \mid X)$. A modified decision rule is derived as follows:

$$\lambda_{11} P(\omega_1 \mid X) + \lambda_{12} P(\omega_2 \mid X) > \lambda_{21} P(\omega_1 \mid X) + \lambda_{22} P(\omega_2 \mid X)$$
$$(\lambda_{12} - \lambda_{22}) P(\omega_2 \mid X) > (\lambda_{21} - \lambda_{11}) P(\omega_1 \mid X)$$

where $(\lambda_{12} - \lambda_{22}) > 0$ and $(\lambda_{21} - \lambda_{11}) > 0$. This decision rule sounds accurate since the loss incurred in taking the wrong decision is always greater than the loss incurred in taking the right decision. Similarly, a decision rule for the case may be constructed where $R(\alpha_1 \mid X) < R(\alpha_2 \mid X)$.

$$\lambda_{11} P(\omega_1 \mid X) + \lambda_{12} P(\omega_2 \mid X) < \lambda_{21} P(\omega_1 \mid X) + \lambda_{22} P(\omega_2 \mid X)$$
$$(\lambda_{12} - \lambda_{22}) P(\omega_2 \mid X) < (\lambda_{21} - \lambda_{11}) P(\omega_1 \mid X)$$

Minimum Error Rate Classifier

Define the zero-one loss function as:

$$\lambda_{ij} = 1 \quad \text{when} \quad i \neq j$$
$$\lambda_{ij} = 0 \quad \text{when} \quad i = j$$

$1 \leq i \leq c \text{ and } 1 \leq j \leq c$ where *c* is the total number of classes. The expressions for risk are as follows:

$$R(\alpha_i \mid X) = \sum_{j=1}^{c} P(\omega_j \mid X)$$
$$R(\alpha_i \mid X) = 1 - P(\omega_i \mid X)$$

Clearly, $P(\omega_i \mid X)$ has to be maximized in order to minimize $R(\alpha_i \mid X)$. The posterior probability corresponding to the class w_i shall determine the outcome of the classifier in favor of that class.

Bound Under Min Max Risk Criterion:

We now proceed to analyze the bounds of risk incurred during the classification task. Define R_1 as the region in the feature space where the classifier decides ω_1 and R_2 as the region in the feature space where the classifier decides ω_2. We may define risk as an integral over the decision space,

$$R = \int_{R_1} \lambda_{11} P(\omega_1 \mid X) + \lambda_{12} P(\omega_2 \mid X) + \int_{R_2} \lambda_{21} P(\omega_1 \mid X) + \lambda_{22}(\omega_2 \mid X)$$

Equivalently,

$$R = \int_{R_1} \lambda_{11} P(\omega_1) P(X \mid \omega_1) + \lambda_{12} P(\omega_2) p(X \mid \omega_2) + \int_{R_2} \lambda_{21} P(\omega_1) p(X \mid \omega_1) + \lambda_{22} P(\boldsymbol{\omega}_2) P(X \mid \omega_2)$$

The prior/posterior probabilities for the classes in a two class problem are related by

$$P(\omega_2) = 1 - P(\omega_1)$$

&

$$\int_{R_1} p(X \mid \omega_1) dX = 1 - \int_{R_2} p(X \mid \omega_1) dx$$

Since the loss incurred in taking the wrong decision is always greater than the loss incurred in taking the right decision, we have

$$\lambda_{12} > \lambda_{11} \;\&\; \lambda_{21} > \lambda_{22}$$

&

$$R < \int_{R_1} \lambda_{12}(P(\omega_1) p(X \mid \omega_1) + (1 - P(\omega_1)) p(X \mid \omega_2)) dX \int_{R_2} (\lambda_{21} P(\omega_1) p(X \mid \omega_1) + (1 - P(\omega_1)) P(X \mid \omega_2)) dX)$$

Under zero one loss function, the risk is recomputed as

$$< \int_{R_2} (\lambda_{21} P(\omega_1) p(X \mid \omega_1) + (1 - P(\omega_1)) P(X \mid \omega_2)) dX) + \int_{R_2} (P(\omega_1) p(X \mid \omega_1) + (1 - P(\omega_1)) P(X \mid \omega_2)) dX$$

This is the bound on risk under ***Min-Max Criterion.***

$$R < \int_{R_1 + R_2} (P(\omega_1) p(X \mid \omega_1) + (1 - P(\omega_1)) p(X \mid \omega_2)) dX$$

Error probabilities and the classification problem:

Since the problem is posed as binary, let us gain some insight into the source of errors, for the two class problem. This is equivalent to partitioning the decision space into two regions R_1 and R_2 and investigating the affiliation of an observation point, x (sample journal from the test set) to R_2 when the true class is ω_1, or $X \in R_1$ and the true class being ω_2. These are mutually exclusive and exhaustive. Note,

$$P(\text{error}) = \int_{R_2} p(X \mid \omega_1) P(\omega_1) dx + \int_{R_1} p(X \mid \omega_2) P(\omega_2) dX$$

If $P(x \mid \omega_1) P(\omega_1) > P(x \mid \omega_2) P(\omega_2)$, classifying $x \in R_1$ is better since the smaller quantity will contribute to the error integral. However, if the journal classification problem is not posed as a binary problem, there are more ways to be wrong than to be right. We then compute the probability of being correct i.e.

$$P(\text{correct}) = \sum_{i=1}^{c} P(x \in R_i, \omega_i)$$

$$P(\text{correct}) = \sum_{i=1}^{n} \int_{R_1} p(x \mid \omega_i) P(\omega_i) dx$$

The decision rule guarantees the lowest average error rate. In the two-class case, the general error integral may be approximated analytically to provide an upper bound. We present the result for error in the Gaussian case (two category, multi dimensional data for the journal classification problem). The general error integral may be approximated by an analytical upper bound. We shall present a lemma which verifies this claim.

Lemma 1: *P(error),* error of misclassification in the binary classification problem is bounded above by

$$P_{(\omega_1)}^{\beta} P_{(\omega_2)}^{1-\beta} \int p^{\beta}(x \mid \omega_1) p^{1-\beta}(x \mid \omega_2) dx, \quad 0 \le \beta \le 1$$

The integral spans the feature space as there is no need to impose limits of integration corresponding to decision boundaries. Since we assumed class-conditional probabilities to be normal, it may be shown that the integral on the right equals $e^{-k(\beta)}$ implying the error is optimized in 1-D space even though the distribution belongs to a space of higher dimensions, where

$$k(\beta) = \frac{\beta(1-\beta)}{2} (\mu_2 - \mu_1)^T \left[\beta \sum_1 + (1-\beta) \sum_2 \right]^{-1} (\mu_2 - \mu - 1) + \frac{\frac{1}{2} \ln \left| \beta \sum_1 + (1-\beta) \sum_2 \right|}{\left| \sum_1 \right|^{\beta} \left| \sum_2 \right|^{\beta}}$$

Remark: One dimensional error bound is possible in a multi variate problem. This is encouraging.

Proof: We require the following identity:

$$\min[a, b] \le \alpha^{\beta} b^{1-\beta}, a, b \ge 0, 0 \le \beta \le 1$$

Assume,

$$a \ge b$$

Then,

$$\left(\frac{a}{b}\right)^{\beta} \geq 1$$
$$\Rightarrow \left(\frac{a}{b}\right)^{\beta} \geq b$$
$$\Rightarrow \alpha^{\beta} b(1-\beta) \geq b$$

We know,

$$P(\text{error}) = \int_{-\infty}^{\infty} P(\text{error}, x)dx = \int_{-\infty}^{\infty} P(\text{error} \mid x)p(x)dx$$

Applying the inequality to the integral above, we obtain

$$P(\text{error}) \leq P^{\beta}(\omega_1)P^{1-\beta}(\omega_2)\int p^{\beta}(x \mid \omega_1)p^{1-\beta}(x \mid \omega_2)dx$$

An illustrative example: β can't be identically 0 or 1, since that would indicate the prior probabilities w.r.t classes ω_1 and ω_2 respectively to be heavily biased toward one particular class. This is not the case as the problem considers samples almost equally distributed to two classes (please refer to the data set in the appendix). The prior probability in this case is generated from the algorithm and the data set. Table 1 presents different values of upper bound for different β values. Careful manipulations would lead to tighter error bounds by fixing convenient choices of the likelihood. We prove the following lemma to establish the assertion.

(β)	k(β)	*exp(–k(β))*	Upper Bound
0.6	3.22	0.04	0.02
0.7	3.54	0.029	0.012
0.8	0.796	0.451	0.231
0.9	0.347	0.707	0.364

Table 1: The upper bound of error for different values; prior probabilities are calculated from historical data of National and International Journals. $P(\omega_1) = 0.52$ and $P(\omega_2) = 0.48$: Classification problem is posed nicely as $P(\omega_1)$ and $P(\omega_2)$ are close to each other and there is no clear class domination. The upper bound values are empirical evidence of the theoretical guarantee of minimum accuracy of classification.

Lemma 2:: Tighter analytical lower and upper bound on errors: We obtain the following lower and upper bounds on the error:

- Lower bound is computed as

$$e_L(p) = \frac{1}{\beta}\ln\left[\frac{1+e^{-\beta}}{e^{-\beta p}+e^{-\beta(1-p)}}\right],$$

  for any $\beta > 0$ with a suitable choice of the likelihood, p to tighten the bound.

- It can also be shown that, the upper bound is given by

$$e_u(p) = e_L(p) + [1 - 2e_L(0.5)]e_{c_1}(p)$$

were,

$$e_{c_1}(p) \geq \min[p, 1-p]; e_{c_1}(p) = e_{c_1}(1-p)$$
$$e_{c_1}(0) = e_{c_1}(1) = 0; e_{c_1}(0.5) = 0.5$$

Proof of lower bound: Set $p = p(x \mid \omega_1)$ and note that

$$e_L(p) = \frac{1}{\beta}\ln\left[\frac{1+e^{-\beta}}{e^{-\beta p}+e^{-\beta(1-p)}}\right]$$

is symmetric with respect to the interchange p and $(1 - p)$ and thus to $p = 0.5$. This implies, the limit probabilities on $\left[0, \frac{1}{2}\right]$ applied to p will hold on the interval $\left[\frac{1}{2}, 1\right]$ applied to p. It's easy to check that $\min[p, 1 - p] = p$ for $p \in \left[0, \frac{1}{2}\right]$. Therefore,

$$e_L(p) = \frac{1}{\beta}\ln\left[\frac{1+e^{-\beta}}{1+e^{-\beta}}\right] = \frac{1}{\beta}\ln[1]$$
$$\frac{\partial}{\partial L}e_L(p) = \frac{e^{\beta} - e^{2\beta p}}{e^{\beta} + e^{2\beta p}} < 1 = \frac{\partial}{\partial p}\min[p, 1-p]$$

Next,

$$\lim_{\beta \to \infty}\frac{\partial}{\partial p}e_L(p) = \lim_{p \to 1}\frac{e^{\beta} - 2e^{2\beta p}}{e^{\beta} + e^{2\beta p}} = 1, p \in \left[0, \frac{1}{2}\right]$$

Proof of upper bound: Let

$$e_L(p) = p - \theta(p), \theta(p) \geq 0, \text{ with degree at least 1.}$$
$$e_{c_1} = p + \psi(p) \text{ where } \psi(p) \geq 0 \text{ and } \psi(0) = \psi\left(\frac{1}{2}\right) = 0$$

Then,

$$e_u(p) = p - \theta(p) + \left[1 - 2\left(\frac{1}{2} - \theta\left(\frac{1}{2}\right)\right)\right](p + \psi(p))$$
$$\text{i.e. } e_u(p) = p - \theta(p) + \theta\left(\frac{1}{2}\right)(p + \psi(p))$$

$$e_u(p) - p = -\theta + \theta\left(\frac{1}{2}\right)p + \theta\left(\frac{1}{2}\right)\psi(p) > 0$$

$$e_u(p) = e_L(p) + [1 - 2e_L(0.5)]e_{c_1}(p)$$

Hence the proof. Table 2 contains the computed values of the error bounds for a particular choice of the likelihood, consistent with the derived error bounds.

β	p	Lower Bound $e_L(p)$	Upper Bound $e_u(p)$
0.6	0.5	0.074	0.5
0.7	0.5	0.086	0.5
0.8	0.5	0.974	0.5
0.9	0.5	0.1087	0.5

Table 2: The upper and lower bound of error for different β values; an illustrative example of the tighter bounds when the samples are distributed among the two journal classes, International and National.

Probability of Error under Cauchy Conditional Density: Thicker tailed distribution

If there exists minimum overlap between the features, we may assume two identical distributions in one dimension to accommodate the two-class problem (this holds good once feature ranking is accomplished and it is ascertained that certain features may be ignored, a straightforward task). Cauchy density represents data with thicker tails. Fat tails may be a possibility and represent skewness in data which may well be the case in the context of the problem. Therefore, instead of using multiple one dimensional normal density, Cauchy distribution is used. Estimating/approximating normal distributions from a fat-tailed one (such as Cauchy distribution) could prove untenable. This is the reason, we investigate the probability of error bound under the assumption of Cauchy conditional density,

$$p(x \mid \omega_i) = \frac{1}{\pi b} \frac{1}{1 + \left(\frac{|x - a_i|}{b} \right)^2}; i = 1,2, \text{ where } \omega_1, \omega_2 \text{ are the two}$$

classes, national and international, a_i are the peaks of the two distributions, b is the width, and without loss of generality, $a_1 > a_2$. This is mathematically equivalent to Bayesian error rate with the Neyman-Pearson condition. *Sporadic labeling may cause extreme events. The model should explain such events, however small the probability of those events be. The risk/payoff could very well be unwise to ignore! Imagine a journal being assigned different labels by different indexing/database services.*

Artificial Balancing of data: From a data analytic point of view, if most of the samples in the dataset belong to one class, it is known as *data bias* and can lead to over-fitting, i.e., when a classifier becomes overly complex and extremely sensitive to the nuances in the data. Over-fitting is a problem that needs to be dealt with carefully and not be overlooked as an administrative task. If a dataset has number of samples belonging to one class over a thousand times the total number of samples belonging to all the other classes, just reporting the numeric accuracy obtained by directly feeding the data to train a classifier would be an incorrect methodology. To counter the potential problems due to the dominance by a single class, *artificially balanced* datasets are used by considering random samples from the dominating classes and other classes with the total number of samples belonging to one class being equal to the number of samples in the *other* class, as it has the least number of samples. Then this balanced dataset is divided in a suitable ratio where the larger portion was that of the training set. This cycle of balancing the data set artificially, dividing it, training and testing a classifier was performed multiple times, and the mean accuracy of all the trials was considered to be representative of the potential of a classifier. By artificial balancing, the reported accuracies are also more reliable than without balancing. The model and error bound are considered for balanced data as practice i.e the prior probability of both classes being equal. If this is not the case such as the present data set where, $P(\omega_1) = 0.52$ and $P(\omega_2) = 0.48$, the error bound may be affected. We consider the error bound assuming equal prior probabilities under Bayesian error rate with the Neyman-Pearson condition and Cauchy conditional distribution *(Lemma 3)*. The imbalance in data (different prior probabilities) is then accounted for by modeling the prior probability and a new error bound is computed. It is then shown that, as long as the amount of imbalance is not significant to the extent of imparting noteworthy class bias, the error bound is identical to the original error bound in the limiting sense *(Lemma 4)*. If the number of international journals equals the number of national journals i.e. if $P(\omega_1) = P(\omega_2) = \frac{1}{2}$, then we have the following result.

Lemma 3: The Probability(error) $= \frac{1}{2} - \frac{1}{\pi} \arctan\left(\frac{a_2 - a_1}{2b} \right)$.

Proof of Lemma 3: Risk is defined as,

$$R = \int_{R_1} p(x \mid \omega_2) P(\omega_2) dx + \int_{R_2} p(x \mid \omega_1) P(\omega_1) dx$$

$$R = \frac{1}{2} \int_{R_1} p(x \mid \omega_2) dx + \frac{1}{2} \int_{R_2} p(x \mid \omega_1) dx$$

$$Since, \int_{R_1} p(x \mid \omega_1) dx = 1 - \int_{R_2} p(x \mid \omega_1) dx$$

$$R = \frac{1}{2}[1 - \int_{R_1} p(x \mid \omega_1) dx] + \frac{1}{2} \int_{R_1} p(x \mid \omega_2) dx$$

$$= \frac{1}{2}[1 + \int_{R_1} p(x \mid \omega_2) - p(x \mid \omega_1) dx]$$

$$= \frac{1}{2} + \frac{1}{2\pi b}[\int_{R_1} \frac{1}{1 + \frac{(x-a_2)^2}{b}} - \frac{1}{1 + \frac{(x-a_1)^2}{b}} dx]$$

$$R_1 \Rightarrow x = a_1 \text{ or } x = a_2 \Rightarrow \frac{(x - a_1)(x - a_2)}{b^2} = 0$$

$$probability(error) = \frac{1}{2} - \frac{1}{\pi} \arctan|\frac{a_1 - a_2}{2b}|$$

Remark: In the context of the problem, prior probabilities of two classes ω_1 and ω_2 are not exactly same, but very close. $p(\omega_1) = 0.52$, $p(\omega_2) = 0.48$. Let's model the prior probability in the following way:

$p(\omega_1) = \frac{1}{2} + \in$; $p(\omega_2) = \frac{1}{2} - \in$; $\in > 0$; where $\in = 0.02$.

In that case , what would be the probability(error)? The answer is contained in the following lemma.

Lemma 4: Probability of error varies linearly with the magnitude of data imbalance, in the event samples are not equally distributed among classes.

Proof of Lemma 4:

$$R = \int_{R_1} p(x \mid \omega_2)\left(\frac{1}{2} - \in\right)dx + \int_{R_2} p(x \mid \omega_1)\left(\frac{1}{2} + \in\right)dx$$

$$R = \frac{1}{2}\int_{R_1} p(x \mid \omega_2) + \frac{1}{2}\int_{R_2} p(x \mid \omega_1)dx + \in\left[\int_{R_2} p(x \mid \omega_1)dx - \int_{R_1} p(x \mid \omega_2)dx\right]$$

From the previous lemma, where the samples are equally distributed in two classes (number of journals in class national and international is equal).

$$\text{So, } \frac{1}{2}\int_{R_1} p(x \mid \omega_2) + \frac{1}{2}\int_{R_2} p(x \mid \omega_1)dx = \frac{1}{2} - \frac{1}{\pi}\arctan\left|\frac{a_1 - a_2}{2b}\right|$$

Consider,

$$\in\left[\int_{R_2} p(x \mid \omega_1)dx - \int_{R_1} p(x \mid \omega_2)dx\right]$$

$$\int_{R_2} p(x \mid \omega_1)dx = 1 - \int_{R_1} p(x \mid \omega_1)dx$$

$$\text{Therefore } \int_{R_2} p(x \mid \omega_1)dx - \int_{R_1} p(x \mid \omega_2)dx$$

$$= \in 1 - \int_{R_1} p(x \mid \omega_1)dx - \int_{R_1} p(x \mid \omega_2)dx$$

$$= \in\left[1 - \left[\int_{R_1} p(x \mid \omega_1) + p(x \mid \omega_2)dx\right]\right]$$

$$= \in\left[1 - \left[\int_{R_1} \frac{1}{\pi b}\frac{1}{1 + \left(\frac{x_1 - a_1}{b}\right)^2} + \frac{1}{\pi b}\frac{1}{1 + \left(\frac{x - a_2}{b}\right)^2}dx\right]\right]$$

$$= \in\left[1 - \frac{1}{\pi b}\left[\int_{R_1} \frac{1}{1 + \left(\frac{x - a_1}{b}\right)^2} + \frac{1}{1 + \left(\frac{x - a_2}{b}\right)^2}dx\right]\right]$$

To evaluate and simplify the integral, we follow the following steps:

$$b\arctan\left(\frac{x - a_1}{b}\right) + b\arctan\left(\frac{x - a_2}{b}\right)$$

$$= \left[\arctan\left(\frac{x - a_1}{b}\right) + \arctan\left(\frac{x - a_2}{b}\right)\right]$$

$$= b\arctan\left(\frac{\frac{x - a_1}{b} + \frac{x - a_2}{b}}{1 - \frac{(x - a_1)(x - a_2)}{b^2}}\right)$$

$$= b\arctan\left(\frac{2x - \frac{(a_1 + a_2)}{b}}{1 - \frac{(x - a_1)(x - a_2)}{b^2}}\right)$$

$$R_1 \Rightarrow x = a_1 \text{ or } x = a_2$$

Therefore the integral becomes $b\arctan\left(\frac{a_1 - a_2}{b}\right)$ or $b\arctan\left(\frac{a_2 - a_1}{b}\right)$.

Consequently, the expression for risk is obtained.

$$R_{new} = \frac{1}{2} - \frac{1}{\pi}\arctan\left|\frac{a_1 - a_2}{2b}\right| + \in 1 - \frac{1}{\pi}\arctan\left|\frac{a_1 - a_2}{2b}\right|$$

$$= \left(\frac{1}{2} + \in\right) - \frac{1}{\pi}\arctan\left|\frac{a_1 - a_2}{2b}\right|(\in + 1)$$

$$\|R_{new} - R\| = \in\left[1 - \frac{1}{\pi}\arctan\left|\frac{a_1 - a_2}{2b}\right|\right]$$

Remark: *P(error)* is maximized when $\left|\frac{a_2 - a_2}{2b}\right| = 0$ & equals $\frac{1}{2}$. This happens if $a_1 = a_2$ i.e. when both distributions are same. Since *arctan* is strictly bounded by $\frac{\pi}{2}$, $\|R_{new} - R\| \leq \frac{\in}{2}$

Remark 1: As long as *e* is bounded by a small number, the risk will vary insignificantly compared to the risk in the balanced sample. It can be easily shown that $|R_{new} - R| < \frac{\in}{2}$ since *arctan* has a strict upper bound of $\frac{\pi}{2}$. This implies that the difference

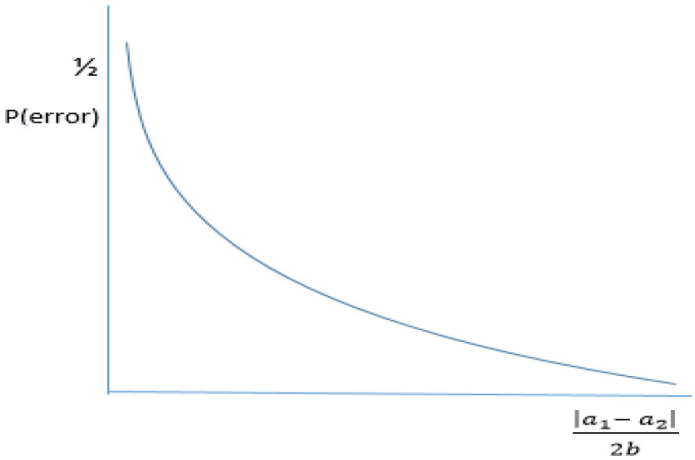

Figure 2: P(error) decreases as a function of $\left|\frac{a_1 - a_2}{2b}\right|$

between the two risks is negligible as long as the amount of class imbalance is insignificantly small as shown in Figure 2.

Remark2: Covariance & Covariance matrix: Covariance is a measure of the joint variability of two random variables. Given two random variables X and Y, their co-variance is, $cov(X,Y) = E[X-E(X)(Y-E(Y)]$ where, $E(X)$ & $E(Y)$ are means of X & Y. It is well known that, cross-covariances are zero for statistically independent variables. Covariance matrix is a matrix whose element in the i, j position is the covariance between the i-th and j-th elements of a random vector. The covariance matrix provides a succinct way to summarize the covariances of all pairs of variables. In our case, the covariance matrix between the four features is the following :

$$\begin{pmatrix} 0.008149 & 0.005732 & 0.01128 & 0.01872 \\ 0.005732 & 0.040904 & 0.01666 & 0.02246 \\ 0.011283 & 0.016665 & 0.07531 & 0.06635 \\ 0.018726 & 0.022460 & 0.06635 & 0.10588 \end{pmatrix}$$

It is observed that the off diagonal entries in the above matrix are close enough to 0. It shows that the features are weakly correlated and hence may be assumed as statistically independent. This allows multivariate normal density function reformulated as product of univariate normal density functions. This ensures tight error bound as the cumulative error is computed as the product of the error (bounded) due to each feature.

Remark 3: Multivariate normal density, when features are statistically independent (covariance matrix is diagonal), can be written as product of independent univariate normal density functions.

Proof: *Multivariate normal density pdf* is written as:

$$p(X) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}((X-\mu)^t \Sigma^{-1}(X-\mu)\right]$$

where, X is a d dimensional feature vector; Σ is a d × d matrix called the covariance matrix, μ is the mean vector of the feature vector X i.e $\mu = E(X)$, and $(X-\mu)^t \Sigma^{-1}(X-\mu)$ is called square of the Mahalanobis distance.

In case of diagonal covariance matrix (cross covariances are either zero or close enough), $\sigma_{ij} = 0$, for $i,j = 1,....,n$

$$p(x) = \frac{1}{(2\pi)^{d/2}\left|\sigma_{ii}^2\right|_{n\times n}^{\frac{1}{2}}} \exp\left(\frac{-1}{2}[x_i-\mu_i]_{n\times n}^T \left[\frac{1}{\sigma_{ii^2}}\right] n\times n [x_i-\mu_i] n\times n\right)$$

$$p(x) = \frac{1}{(2\pi)^{d/2}\sigma_1\sigma_2\ldots\sigma_n} \exp\left(\frac{-1}{2}[x_i-\mu_i]_{n\times n}^T \left[\frac{1}{(\sigma_i)^2_{n\times n}}(x_i-\mu_i)\right] n\times n\right)$$

Simplifying further,

$$p(x) = \frac{1}{\sqrt{2\pi}^{d/2}\sigma_1\sigma_2\cdots\sigma_n} \exp\left(\frac{-1}{2\sigma_1^2}(x_1-\mu_1)^2\right)$$
$$\frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(\frac{-1}{2\sigma_2^2}(x_2-\mu_2)^2\right)\cdots\frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(\frac{-1}{2\sigma_n^2}(x_n-\mu_n)^2\right)$$

Thus, the multivariate density becomes a product of independent univariate density functions.

5.1.6 Discriminant Function

Assume, there exists 2 classes according to the common wisdom of classifying internationality. The evaluation of discriminant function for every class is the next task needs to be accomplished. This is necessary since classifier assigns unknown sample to the class having maximum value for the discriminant function.

$R(\alpha_i \mid X)$ has to be minimized while maximizing the discriminant function in order to design the minimum risk classifier. This implies, $g_i(X) = -R(\alpha_i \mid X)$. It is known that, minimum error rate classifier requires $P(\omega_i \mid X)$ to be maximized while discriminant function needs to be maximized. Therefore, $g_i(X) = P(\omega_i \mid X)$. We also note that, the discriminant function is not unique because if we consider f(g_i(X)) to be monotonically increasing, we may use either g_i(X) or f(g_i(X)). This is advantageous because if g_i(X) is hard to compute, we may resort to a comparatively easy calculation of f(g_i(X)). We use minimum error rate classifier and proceed further by rewriting the discriminant function as

$$g_i(X) = P(\omega_i \mid X)$$
$$g_i(X) = \frac{p(X \mid \omega_i)P(\omega_i)}{p(X)}$$

p(X) appears in the denominator of every discriminant function. This is a scale factor and has no effect. Therefore, we can exclude this so that $g_i(X) = p(X \mid \omega_i)P(\omega_i)$. The product in the expression is difficult to analyze compared to additive terms. The problem is solved by considering a monotonically increasing function such as the log function,

$$g_i(X) = \ln p(X \mid \omega_i) + \ln P(\omega_i)$$

Now consider the two class journal classification problem i.e $c = 2$, ω_1 and ω_2 being the two classes, National and International. We obtain the following decision rule,
$g_1(X) > g_2(X) \Rightarrow$ class ω_1 else, $g_1(X) < g_2(X) \Rightarrow$ class ω_2.

Next, decision boundary between the two classes is defined as $g_1(X) = g_2(X)$ i.e. $g_1(X) - g_2(X) = 0$.

Let $g(X) = g_1(X) - g_2(X)$.

This yields the expression for the discriminant function, $g(X) = P(\omega_1 \mid X) - P(\omega_2 \mid X)$

$$g(X) = \ln \frac{p(X \mid \omega_1)}{p(X \mid \omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)} \quad (3)$$

We assume probability density function as normal whose pdf is:

$$p(X) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left[-\frac{1}{2}((X-\mu)^t \Sigma^{-1}(X-\mu) \right]$$

In order to establish bounds of risk and to show that there exist an one-dimensional risk bound for the multidimensional density function, we assume that the features follow multi-dimensional Gaussian distribution. To check normality of features we perform Shapiro-Wilk test. The Shapiro-Wilk test utilizes the null hypothesis principle to check whether samples $x_1, x_2, ..., x_n$ from a normally distributed population. The null-hypothesis of this test is that the population is normally distributed. Thus, if the p-value is less than the chosen alpha level, then the null hypothesis is rejected and there is evidence that the data tested are not from a normally distributed population. On the contrary, if the *p*-value is greater than the chosen alpha level, then the null hypothesis that the data came from a normally distributed population cannot be rejected. The test statistic is

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

where x_i: *is the i-th order statistic, i.e., the i-th-smallest number in the sample;* $\bar{x} = (x_1 + \ldots\ldots + x_n)/n$ *is the sample mean.*

The constants a_i *are given by*

$$(a_1, \ldots, a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{1/2}}$$

where $m = (m_1, \ldots, m_n)^T$ and $m_1, \ldots, m_n$ are the expected values of the order statistics of independent and identically distributed random variables sampled from the standard normal distribution, and V is the covariance matrix of those order statistics. Each feature in the feature space is checked for normality. This justifies using a multivariate normal pdf. when our feature vector is multidimensional then the *multivariate normal density pdf* is:

$$p(X) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left[-\frac{1}{2}((X-\mu)^t \Sigma^{-1}(X-\mu) \right]$$

where

X is a d dimensional feature vector; Σ is a $d \times d$ matrix called the covariance matrix, μ is the mean vector of the feature vector X i.e $\mu = E(X)$, and $(X-\mu)^t \sum^{-1} (X-\mu)$ is called square of the Mahalanobis distance.

The class conditional density for class ω_i with mean μ_i and covariance matrix Σ_i can now be defined as

$$p(X \mid \omega_i) = \frac{1}{(2\pi)^{d/2} \left| \sum_i \right|^{1/2}} \exp\left[\frac{-1}{2}((X-\mu_i)^t \sum_i{}^{-1} (X-\mu_i) \right]$$

We obtain the expression for discriminant function of class ω_i using the class conditional density of class ω_i i.e $p(X \mid \omega_i)$ as

$$g_i(X) = \ln p(X \mid \omega_i) + \ln P(\omega_i)$$

$$g_i(X) = \frac{-1}{2}\left[(X-\mu_i)^t \sum_i{}^{-1} (X-\mu_i) \right]$$

$$-\frac{d}{2} \ln 2\pi - \frac{1}{2} \ln \left| \sum_i \right| + \ln P(\omega_i)$$

This is the discriminant function expression of class ω_i. It is quadratic due to the term $\frac{-1}{2}\left[(X-\mu_i)^t \sum_i (X-\mu_i) \right]$. This implies the Bayes' classifier can take care of linearly non separable classes.

THE LOSS FUNCTION AND RISK BOUND IN INTERNATIONALITY CLASSIFICATION

Losses which may occur during the classification are described in the following manner:

λ_{11} = loss occurred when we take action for a journal in favor of class national and the true class is national.

λ_{12} = loss occurred when we take action for a journal in favor of class international and the true class is national

λ_{21} = loss occurred when we take action for a journal in favor of class international and the true class is international

λ_{22} = loss occurred when we take action for a journal in favor of class international and the true class is international

Intuitively, $\lambda_{11} < \lambda_{12}$ *and* $\lambda_{21} > \lambda_{22}$ *because the loss occurred in making a wrong decision is obviously greater than the loss occurred in making the correct one.* In the context of our problem, $\lambda_2 > \lambda_2$ *i.e the loss in classifying a national journal as international is more than the cost of classifying international journal as national journal. This is analogous to the extent of damage caused in a security system granting authentication to an unknown person compared to not providing authentication to the known/authorized person. The losses in journal classification are not homogeneous and the cost in classifying a national journal as international journal is more than the cost of classifying an international journal as national.*

If a journal is classified as national the expected risk associated with this action is defined as

$$R(\alpha_1 \mid X) = \lambda_{11}P(\omega_1 \mid X) + \lambda_{12}P(\omega_2 \mid X) \quad (5)$$

where a is the action taken in favor of class national for a journal and X is a d dimensional feature vector. OTOH, if we classify a journal as international, the expected risk due to this action is formulated as

$$R(\alpha_2 \mid X) = \lambda_{21}P(\omega_1 \mid X) + \lambda_{22}P(\omega_2 \mid X) \quad (6)$$

where α_2 is the action taken in the favor of class international for a journal. This yields the following decision rule:

If $R(\alpha_1 \mid X) < R(\alpha_2 \mid X)$ then $1 - P(\omega_1 \mid X) < 1 - P(\omega_2 \mid X)$, *implying* $P(\omega_1 \mid X) < P(\omega_2 \mid X)$; *assign journal to class national.*

However, if $R(\alpha_1 \mid X) > R(\alpha_2 \mid X)$ then $1 - P(\omega_1 \mid X) > 1 - P(\omega_2 \mid X)$ *implying* $P(\omega_1 \mid X) < 1 - P(\omega_2 \mid X)$*; assign journal to class international.* The rule is translated in the form of an algorithm.

Upper bound and lower bound of Risk under Zero-One classification cost

We present further discussions on error bounds for special cases. Let us consider zero-one classification cost i.e $\lambda_{11} = \lambda_{22} = 0$ and $\lambda_{12} = \lambda_{21} = 1$. For a two class problem such as journal internationality, we have, prior probability of the class "international journal", $P(\omega_2) = 1 - P(\omega_1)$ where $P(\omega_1)$ is the prior probability of the class "national journal". Applying zero one loss function in the context of our problem converts equation 5 to

$R(\alpha_1 \mid X) = P(\omega_2 \mid X)$ i.e. $R(\alpha_1 \mid X) = 1 - P(\omega_1 \mid X)$ and equation 6 becomes

$R(\alpha_2 \mid X) = P(\omega_1 \mid X)$ i.e $R(\alpha_2 \mid X) = 1 - P(\omega_2 \mid X)$. The Risk integral is given by

$$R(P(\omega_1) = P(\omega_1)\int_{R_2} p(X \mid \omega_1)dx + (1 - P(\omega_1))p(X \mid \omega_2))dX + \int_{R_2} (\lambda_{21}P(\omega_1)p(X \mid \omega_1).$$

Under this settings, it is now possible to present a quantitative formulation for the probability of error and compute lower and upper bounds for such error under the zero-one risk classification paradigm.

Lemma 5: Under the Bayesian decision rule the classification error in the two class problem is given as:

$$P(\text{error}) = \int P(\text{error} \mid x)p(x)dx = \int \min[P(\omega_1 \mid x),[P(\omega_2 \mid x)]p(x)dx$$

Further it can be shown that, the probability of error is bounded below and above by the following

$$\int P(\omega_1 \mid x)P(\omega_2 \mid x)p(x)dx \le P(\text{error}) \le \int 2P(\omega_1 \mid x)P(\omega_2 \mid x)p(x)dx$$

Proof:

$$P(\text{error}) = \int P(\text{error} \mid x)p(x)dx = \int \min[P(\omega_1 \mid x)],[P(\omega_2 \mid x)]p(x)dx$$

If $P(\omega_1 \mid x) \ge P(\omega_2 \mid x)$ then, $P(\text{error}) = \int P(\omega_2 \mid x)p(x)dx$. By probability theory, $0 \le P(\omega_1 \mid x) \le 1$. Hence,

$$0 \le P(\omega_1 \mid x)P(\omega_2 \mid x) \le P(\omega_2 \mid x)$$
$$0 \le \int P(\omega_1 \mid x)P(\omega_2 \mid x)p(x)dx \le \int P(\omega_2 \mid x)p(x)dx$$
$$0 \le \int P(\omega_1 \mid x)P(\omega_2 \mid x)p(x)dx \le P(\text{error})$$

Using the axiom of probability theory i.e. $P(\omega_1 \mid x) + P(\omega_2 \mid x) = 1$, we obtain the other bound

$$P(\omega_1 \mid x) = 1 - P(\omega_2 \mid x)$$
$$P(\omega_1 \mid x) \ge 1 - P(\omega_1 \mid x)$$
$$2P(\omega_1 \mid x) \ge 1$$
$$P(\omega_1 \mid x) \ge 1/2$$
$$2P(\omega_1 \mid x)P(\omega_2 \mid x) \ge P(\omega_2 \mid x)$$
$$\int 2P(\omega_1 \mid x)P(\omega_2 \mid x)p(x)dx \ge \int P(\omega_2 \mid x)p(x)dx$$
$$\int 2P(\omega_1 \mid x)P(\omega_2 \mid x)p(x)dx \ge P(\text{error})$$

Finally, the upper and lower bounds are obtained by combining inequalities as shown in Table 3.

$$\int P(\omega_1 \mid x)P(\omega_2 \mid x)p(x)dx \le P(\text{error}) \le \int 2P(\omega_1 \mid x)P(\omega_2 \mid x)p(x)dx$$

P(error)	Lower Bound	Upper Bound
3.02×10^{-10}	2.38×10^{-10}	4.76×10^{-10}

Table 3: Using prior probabilities from the data set (cf. Appendix) the bounds are computed. Evidently, the bounds of error in the journal classification problem under the Zero-One loss are reasonable.

EXPERIMENTAL SETUP

We considered 42 journals (Please see Appendix - Tables 5,6,7) which are divided into National journals and International journals from different countries, labeled by the authentic indexing service, Scimagojr. The division is done in such a way that class bias ceases to exist. Once the journals are selected, the values of the features defined in section 4.3 are scraped and organized by using algorithms written specifically for this purpose (Please see Algorithms 3 to 9). The data set consist of journals from four fields of science namely *environmental, agronomy and crop, earth and planetary* and *agricultural and biological.* We perform the task of classification by imple-

menting minimum error rate classifier. Normality of features is checked via Shapiro-Wilk test. We verify that the data set is normally distributed. As mentioned earlier, we consider four features as input variables. Multivariate normal distribution is used in the risk bounds due to multi dimensionality of feature vectors.

DATA SCRAPING ALGORITHMS

Data accumulation and preprocessing are crucial parts of any research requiring a huge and reliable data source to work upon. Depending on the data requirement, various options are available gain access to this data. For instance, one can acquire data regarding scientific journals and articles with a single click from web-sites like SciMago, or write their own scripts which retrieves data regarding the articles and journals required from the official sites of the publishers itself. Despite appearing easier, the first option of directly downloading data from 3rd party websites leads to a factor of data inconsistency or incompleteness of the data. Hence, it is a better practice to get the data from the original publisher websites

In order to meet our requirements, it was decided to write a web scraping script in Python Language to scrape information regarding ACM journals and articles directly from their digital library. This decision, even though provided the surety of accurate and complete data, provided several other hurdles along the way. A web-scraping script works by getting the page source code of a web page which for an average website contains the data presented by the browser. In such a case, getting required data from a page become the task of merely identifying the path of html tags (code components) which ultimately present the data, and letting your script crawl through them to get the data. ACM made this tedious, but not so complicated, task a horrifyingly difficult one as it was discovered that the data was being dynamically fetched when the initial web-page had loaded. This is a problem as when a web crawling script requests for a web page, only the primitive code is fetched and not the completely processed code which might be generated by a browser if there is code to be processed dynamically. Automating the task of figuring out the dynamically made requests and generating the same requests by the web-scraping script took some time effort. After successfully being able to code a script to scrape the website, we decided to let it run from a dedicated server to scrape the entire website, only to discover that they have an IP load balancer with extremely scrutinized request limit. Only after a couple of issues of a journal, the IP was blocked for 12 hours. The task of scraping one journal, let alone 40 seemed a lot more time consuming now. We had to deploy 8 independent nodes to run the script simultaneously in order to complete the task in a span of 3 weeks. Another issue discovered while scraping was there were many inconsistencies and variations in the way the data was presented by the website causing us continuously analyze and update the script to handle the anomalies as and when they were discovered. The task ended with getting information for all the ACM journals in top down fashion. The script started from the highest level, i.e. the journal and moved to the next level, discovering all the volumes and issues of it and moved to the final level of fetching data regarding all the articles for every issue.

Data cleansing and preprocessing algorithms

Scraped data needs to be processed further to make it ready for analysis. Separating useful information from a long string, deriving new metrics from the scraped metrics are few of the data preprocessing operations. We have used cosine similarity measure instead of simple string comparison operations to accommodate all small textual variations in the raw scraped data.

Cosine Similarity Metric: Similarity metrics are the class of textual based metrics resulting in a similarity or dissimilarity (distance) score between two text strings. A string metric provides a floating point number indicating the level of similarity based on plain lexicographic match. For example, similarity between the strings orange and range can be considered to be much more than the string apple and orange by using Similarity metrics.

Cosine similarity is a vector based similarity measure. Cosine of two vectors a, b can be derived by using the Euclidean dot product formula. $a.b = |a||b|\cos\theta$

Where, θ represents the angle between a and b.

Sample output-table 4

Journal name of Article: Plant Molecular Biology Journal name of Cited Article: Plant Science 247, 1-12 cosine similarity value: 0.258198889747
Journal name of Article: Theoretical and Applied Genetics Journal name of Cited Article: Theoretical and Applied Genetics 129 (3), 469-484 cosine similarity value: 0.707106781187
Journal name of Article: Agricultural and Forest Meteorology Journal name of Cited Article: American Society of Agricultural and Biological Engineers 59 (2), 555-560 cosine similarity value: 0.301511344578

Computation of derived parameters

Algorithm 3 (in Appendix II) extracts features for internationality index computation spanning all the listed journals from ACM under Engineering and Computer Science field. Features such as total citations, other-citation and international collaboration quotient are computed for each one of these journals from the accumulated data repository. We first extract all the journal names from the source: line 1. Then

extract Total Citations count and all the Articles published in each one of these journals: line 3 and 4. Further on we compute the cumulative/averaged parameter values for that journal from the various values extracted for each article: line 5 to 8. The various function calls in these lines are explained ahead in the report under respective algorithms. the average value for the International Collaboration is computed: line 11. Finally, line 12 and 13 invoke the functions to compute the SNIP and Internationality Index. Table 10 displays all the derived parameters and the Internationality index for 38 ACM journals.

Other-Citation Quotient Computation

Self-citation (Algorithm 4) is defined as a citation where the citing and the cited paper share at least one author. Other-Citation is the complement of self-citation/total citations, i.e 1 – self_citation/total citations. Algorithm 4 provides the skeleton of self-citation computation for an article in a journal. The denominator, total citations, is already computed by parsing web sources. The key to computing Other-Citations Quotient is to calculate self-citations. For this, we first scrape all the cited papers for the input article name (line 1). Then for each one of these cited papers check if it shares at least one common author name with the input article. If true then we increment the self-citation count (lines 3 and 4). By adding all the individual self-citation counts for every article in a journal, we will get the total self-citations count for a journal (line 5 in Algorithm 3).

Non-Local Influence Quotient (NLIQ) - Algorithm 5

Influence is termed as a factor which causes a paper to be cited by other papers. Non-local refers to the fact that some citations originate from different journals; that is, not from the same journal in which the cited paper is published in. Thus, Non-Local Influence Quotient (NLIQ) is defined as follows,

Let **A** be the number of citations made from articles in one journal X to articles belonging to a number of different journals. Let B be the number of citations from articles in journal X to articles in the same journal, X. Then, for a given journal, we have:

$$\text{Non - Local Influence Quotient} = \frac{A}{A+B}$$

It must be stressed that **other-citations** are uniquely different from **non-local influence**. Namely, an other-citation occurs when a paper cites another paper where no authors are in common. On the other hand, non-local influence is the number of citations made from one paper in a given journal, to a number of different journals - divided by the total number of citations.

International Collaboration Ratio : 6

International collaboration accounts for the articles that have been produced by researchers from several countries. In order to compute this parameter, we first extracted the country information of the journal and then the author affiliations for each one of the published articles in that journal. Every author's country is matched with the country of publishing journal. Ratio is calculated on the basis of weights assigned to different combination of authors affiliation and origin of the publishing journal.

Algorithm 6 shows computation of International Collaboration Ratio of a journal. Here, we look for collaboration between two or more scholars with affiliating institutes in different countries. We will pick primary/first listed affiliation while computing international collaboration. This will filter cases where a person may have multiple affiliations. We scrape the data of multiple affiliations (Algorithm 7) from the websites but consider only the primary (first listed) institute in the computation of international collaboration.

Computation of the internationality weight of an article ased on the combination, deduce the weight of the article from a predefined values for a given combination. We form a set of all the different countries that the authors belong to. Another set of all the authors. If all the authors don't belong to the same country(mod(Set of different countries) != 1) then wt(article) = mod(Set of different countries)/mod(Set of all the authors) otherwise wt(article)=0.

Algorithm 7 illustrates steps to fetch author affiliations of an article. Data from an article's **url** is scraped to obtain author name and respective affiliations. We extract all the affiliations in case of multiple affiliations for an author. results of one such article are in the as shown below

To extract the affiliation from the entire string, we first split the string into list of all its words and remove the author name. Onwards convert it into lower case and replace all "underscore" with "blank space". (note-The *stringswithlength* > 75, have the department name along with institute name), af-filiation_words might contains all the words synonym with "university" and "institute" in different languages, like "uni-versidade","universitat"etc.

Hence we extract the strings before and after this and check if it is a institution using Google Geo location API as per algorithm 9. Algorithm 9 uses the output from the API to determine the country and city information of the institution.[1]

RESULTS OF MINIMUM ERROR RATE CLASSIFIER

Let us define frequently used metrics to judge the efficacy of the classifier. We then proceed to report the results in terms of those metrics.

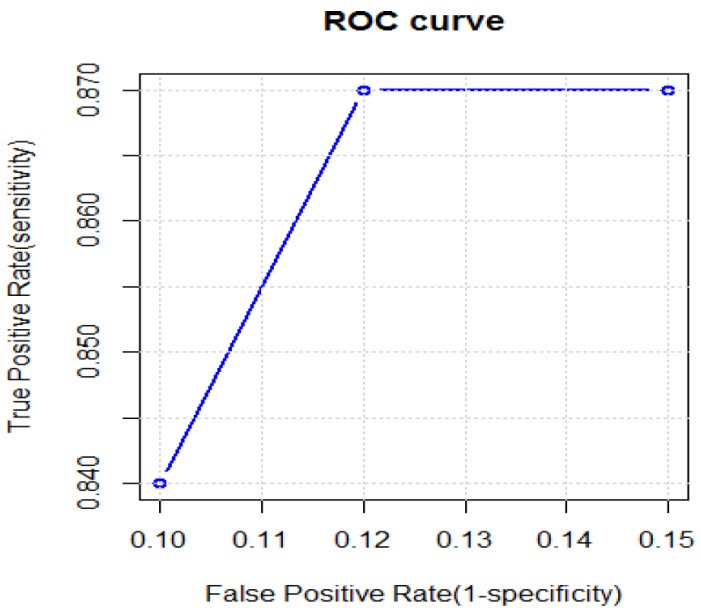

Figure 3: ROC curve.

- ROC curve In a Receiver Operating Characteristic (ROC) ROC curve curve the true positive rate (Sensitivity) is plotted in function of the false positive rate (1-Specificity) for different cut-off points as shown in Figure 3 Each point on the ROC plot represents a sensitivity/specificity pair corresponding to a particular decision threshold. A test with perfect discrimination (no overlap in the two distributions) has a ROC plot that passes through the upper left corner (100 % sensitivity, 100[1] % specificity). Therefore the closer the ROC plot is to the upper left corner, the higher the overall accuracy of the test,[16] properties. An ROC curve demonstrates several things:

1. It shows the trade off between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).

2. The closer the curve follows the upper-left border of the ROC space, the more accurate the test is.

3. The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test will be.

4. The area under the curve is a measure of accuracy.

- **Sensitivity:** Sensitivity is also referred to as the true positive (recognition) rate (i.e., the proportion of positive tuples that are correctly identified).
- **Specificity:** Specificity is the true negative rate (i.e., the proportion of negative tuples that are correctly identified).

The tables show constant trend in accuracy, sensitivity and specificity over varied sizes of train and test data. This is a testimony for the stability of the classifier.

The results found indicate various remedial metrics values and classify journals as national and international. Two journals were found to be tagged as "international" but our classifier labeled it as "national". On further investigation it was found that "internationality" is simply a tag based on country belongingness and does not indicate any measure to the "quality"of work done or collaboration effort made .The results achieved raised two important pointers that further solidify our thought process about the existing doctrine.[2]

- Firstly, our metrics are much effective in classifying journals based on the true aspects of the quality of work done. This is extremely motivating for the research community because finally "quality " gets prioritized over other irrelevant factors.[1] *et. al*. calculated the 'Internationality score' which is convex combination of Journal Influence Score (JIS) and Journal Internationality Modeling Index (JIMI). To calculate JIMI, the following features are used:
  - International Collaboration Ratio
  - Source-Normalized Impact per Paper (SNIP)
  - Other-Citation Quotient
  - Non-Local Influence Quotient (NLIQ)

  The authors observed that the internationality scores are ranging from 0-1, too granular to classify into two groups. The complexity of the classification problem can be further questioned as the granularity in interna-tionality score indicates a demand for a better classification doctrine that can do justice to measuring the " Internationality" tag with qualitative parameters.
- Will a multi-class classification method be a solution to the problems highlighted? The next section further explores this aspect of the solution.

DISCUSSION

It is observed from the previous section that " an international tag" does not make a journal international. The perception about "International Journal of abcd... " being predatory is so strong in the academic community that data evidence is often not required. Indeed, some countries do have cottage industries built around neighborhood publications. These publication houses produce hundreds of so-called "international Journals" with ISSN and few editorial board members from different countries. The two journals in Table 4 despite being labeled

[1]: Source of scraped data : Additional file

[2]: Source of scraped data : Additional file

"International" a priori turn out to be false positives. There may be two reasons behind this. Either, the scheme (minimum error rate) is not good enough i.e. the deficiency in the machine classification approach is not able to classify the journals appropriately or the bina-rization doctrine based on those supposedly "golden rules" is flawed. The doctrine may not capture the shades of gray between the two classes and may well overlook the granularity present in the manual labeling/ classification of journals. We scrutinize the perception further by including another set of journals chosen carefully from one particular country with distinct labels, international and national attached to all of them (list provided in Tables 7-11 in Appendix I). This is accomplished in the following manner:

- The established theory in the manuscript (ref. sections 4, 5 and 6) is tested on the new set by assuming the efficacy of the method proposed i.e. the minimum error rate classifier.
- We have chosen the new set by collecting names of journals with the tag " International" with other para-phernalia such as ISSN number, some members of the editorial board affiliated to other countries different from the country of publication, articles authored by people from "foreign countries" etc.
- Since the accuracy (theoretical and numerical along with hard error bounds) of our method is beyond reasonable doubt, we subject the method to testing journals with "International" and "National" tags.
- We show that, empirically, some of the so called Interna-tional journals do not belong to the "International" class and some of the "National" journals are actually interna-tional (Tables 7- 11, Appendix I). The misclassification is identified by the discrepancy between original label associ-ated with the journals and true label computed by our method.
- This bolsters our assertion that it is too superficial to assume two classes of journals based on internationality, National and International, namely.

[3]For the sake of clarity, let us investigate the efficacy of the classification approach adopted by the authors.

We investigate an important error bound for a multi-class classification. We intend to show that the error bound for the minimum error rate classifier in a binary classification is hard to improve when compared to multi-class (greater than 2) discrimination schemes. It is known that, $\sum_{i=1}^{c} P(\omega_i \mid x) = 1;$

Let i = 1,2, ... c be the number of classes and $\omega_m ax(c)$ be the state for which

$$p(\omega_{\max} \mid x) \geq p(\omega_i \mid x)$$

It follows clearly that, $P(\omega_{\max} \mid c) \geq \frac{1}{c}$

The probability of error $= 1 - \int P(\omega_m \mid x) p(x) dx$

$$\leq 1 - \frac{1}{c} \int p(x) dx \leq 1 - \frac{1}{c}$$

Next, we consider the different cases corresponding to the number of classes.

- **Case1: The trivial case,** c = 1

  The Probability of error is trivially *0* since there exists only one class and hence there is no risk of misclassification! This is justified by the following: $p(\text{error}) \leq 1 - \frac{1}{c} = 0$ implying *p(error) = 0.*

- **Case 2: The binary class problem,** c = 2

  This is a binary classification problem. Let us prove a lemma which shall show that the probability of error is the least when the classification problem is binary.

  Lemma: The probability of error is a monotonically increasing function.

  Proof of Lemma: If the number of classes on the LHS is greater than the number of classes on the RHS i.e. if $e > d$ then $\frac{1}{e} < \frac{1}{d}$; It follows that $\frac{-1}{c} > \frac{-1}{d}$; hence $1 - \frac{1}{c} > 1 - \frac{1}{d}$

  Therefore, $p(\text{error})_{\text{classes} = e} > p(\text{error})_{\text{classes} = d}$

Since c = *1* is the trivial case, c = *2,* the binary classification problem is the smallest of all non-trivial c values. Therefore, the probability of error, p(error) w.r.t any value of $c > 2$ can't be lesser than the probability of error, $p(\text{error})_{c=2}$. Hence the proof. We conclude that, *under the minimum error rate classifier scheme, binary classification produces the best error bound. We demonstrate this fact visually.*

The graph shows that as 'c', the number of classes, increases the probability of error also increases. It is evident from Figure 4 that binary classification problem has the least error bound under this scheme. Hence as far as "internationality"is con-cerned, the minimum error rate based Bayes' classification method is difficult to improve. But the question still remains that what if this method fails to support the binarization doc-trine? It is evident that the scheme is not able to support the doctrine, as observed from Section 5. The conundrum before us is the following:

- Is the method at fault?
- Is the doctrine flawed?

As demonstrated above via analytical and graphical methods, the error rate classifier method gives the best error bound when

[3]: Source of scraped data : Appendix I

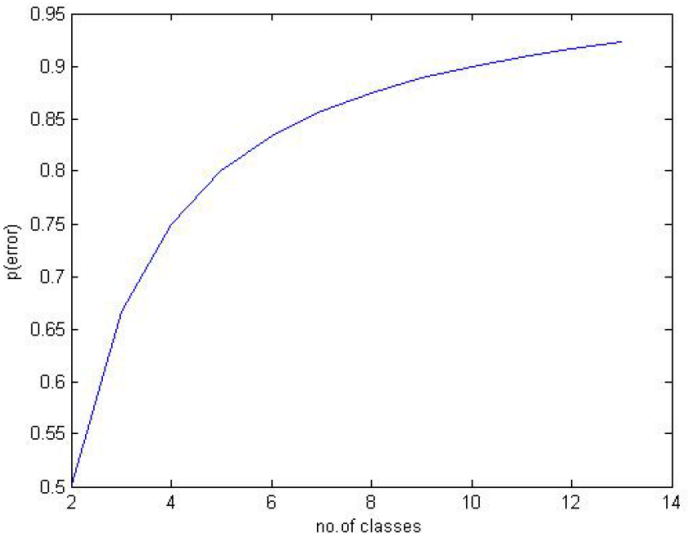

Figure 4: Probability of Error, *P*(error) vs. number of classes, c : The error bound is minimum at *c* = 2 and increases as the number of classes increase. The bound is asymptotically stable after a certain number of classes. The probability of error is a monotonically increasing function.

there are only two classes. Thus, any better method of training the machine so that the observed outcome may reverse is a distant possibility. By the theory of elimination, this directs us to question the credibility of the doctrine! Therefore, let us now explore the suitability of binarization doctrine. The authors pose an interesting question for the research community: "if impact factor and influence has levels then why don't we have levels for "internationality"? Why do we still follow the binarization doctrine? And, if we indeed, need to stick to the doctrine proposed by the Scientometric community, should it really be based on some questionable parameters such as ISSN #, Country of Publication etc ? Should it rather not be based on international influence (the spread and and the rate), international readership, international collaboration, number of international subscriptions and similar measures which could be quantified and normalized across disciplines? We believe that both the doctrine and the governing parameters behind the doctrine are not beyond reasonable doubt!

Currently, we assume four remedial metrics NLIQ, ICQ, OCQ and h-index. There are many more metrics that can be given as input and the model has to scale up accordingly. NLIQ may vary widely across domains and this may influence the results of some journals more compared to others. The challenge lies in correctly identifying classification of journals for categorization and count as there is always some overlap across domains.

One may question the authenticity of the metrics proposed and formulated by the authors of this manuscript. To our defense, we state that these metrics have already been peer reviewed for correctness and relevance and published in Scientometrics.[1] It is not unusual to raise doubts about the values of the features/metrics used in the classifier scheme. These may well have been wrong, one might argue, thus affecting the performance of the classifier and outcome. We state categorically that accuracy of these feature values has been cross-checked with authentic data collection sites (e.g. www.scimagojr.com and Google scholar). This should dispel any doubt regarding the authenticity of the data scraped and used. This leaves the doctrine, the common wisdom and practices open to rigorous scrutiny. Section 5-8 present a strong case for investigating the journal internationality problem

Table 1: Result Of Classification (70% Trainingdata)

Measures	fold-1	fold-2	fold-3	fold-4	fold-5	fold-6	fold-7	fold-8	fold-9	fold-10	Average
accuracy	0.85	0.85	0.92	0.77	0.77	1	0.85	0.92	1	0.85	0.88
sensitivity	1	0.86	1	0.86	0.57	1	0.86	0.86	1	0.71	0.87
specificity	0.67	0.83	0.83	0.67	1	1	1	0.83	1	1	0.88

Table 2: Result Of Classification (80% Training data)

Measures	fold-1	fold-2	fold-3	fold-4	fold-5	fold-6	fold-7	fold-8	fold-9	fold-10	Average
accuracy	0.89	1	0.89	0.89	0.78	0.89	0.89	0.78	0.67	1	0.87
sensitivity	0.8	1	0.8	1	0.8	1	0.8	0.6	0.6	1	0.84
specificity	1	1	1	0.75	0.75	0.75	1	1	0.75	1	0.9

Table 3: The journals which were incorrectly labeled in the catalog of International Journals: The discriminating feature values are reported.

S.no	Journal's Name	OCQ	NLIQ	H-INDEX	IC
1.	Nongye Jixie Xuebao/Transactions of the Chinese Societyof Agricultural Machinery	0.809524	0.928571	0.141243	0.102177
2.	Bragantia	0.470588	0.941176	0.112994	0.093642

Table 4: Result Of Classification using KNN

Training Sample(in %)	Model 1	Model 2
60	79.14729%	80.57142%
80	82.4444%	83.88888%

Model 1: Ungranulated FVs + KNN Algorithm

Model 2: Class Supportive Granulated FVs + KNN Algorithm

where FVs: Feature Vectors and K=5

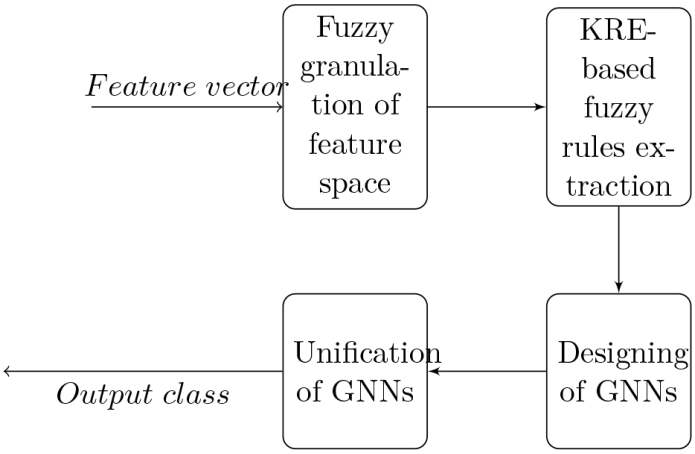

Figure 5: Block diagram of model of Unified GNN.

from a different perspective. We note that the average accuracy of classification under the binary scheme is 87% and it is difficult to improve upon that statistic. Therefore, instead of attempting to classify journals into National and International, we propose a new paradigm. We call this "granules/levels" of journal internationality. Under this paradigm, the internationality of journals is divided into three levels, high, medium and low, thereby imparting the subtle angle the problem deserves to have. We proceed to the methodology of such classification in the next section by iterating our assertion that binary classification of journal in-ternationality does not do justice to the complexity of the problem. We show that the method achieves better accuracy in new classification schemes. This is a testimony of empirical evidence that our proposed scheme is fairer and measured compared to the binarization doctrine. This, we believe, is a disruptive change, in the study of internationality of perreviewed journals.

INTERNATIONALITY CLASSIFICATION BASED ON UNIFIED GRANULAR NEURAL NETWORKS

A conventional Neural Network, though a powerful classification tool, becomes inefficient and complex when it has to

Table 4:List of "International Journals" as labeled by indexing services, Source: www.scimagojr.com; the indicators, OCQ, NLIQ,HIndex and IC havebeen computed by the algorithms described in section 5.3 Table V: List of "National Journals" as labeled by indexing services, Source: www.scimagojr.com; the indicators, OCQ, NLIQ, H-Index and IC havebeen computed by the algorithms described in section 5.3

S.No	Journal's Name	OCQ	NLIQ	HIndex	IC	
1.	Nature Geoscience	0.764923402	0.995773904	111	42.51	Earth and Planetary
2.	Earth and Planetary Science Letters	0.752840909	0.992897727	177	53.89333	Earth and Plantery
3.	Journal of Advances in Modeling Earth Systems	0.72859116	0.989640884	18	30.8	Earth and Planetary
4.	Earth System Science Data	0.705671642	0.988656716	12	37.753333	Earth and Plantery
5.	Saudi Journal of Biological Sciences	0.930795848	0.975778547	14	24.826667	agricultural and biological
6.	Journal of Agricultural and Environmental Ethics	0.840172786	0.967602592	32	12.49333	agricultural and biological
7.	Brazilian Journal of Biology	0.735849057	0.952830189	35	5.4633333	agricultural and biological
8.	Nongye Jixie Xuebao(Transactions of the Chinese Societyof Agricultural Machinery)	0.80952381	0.928571429	25	5.506667	agricultural and biological
9.	Open Life Sciences	0.729468599	0.956521739	21	18.623333	agricultural and biological
10.	Journal of Biological Systems	0.684210526	0.921052632	22	20.976667	agricultural and biological
11.	Biology Direct	0.866081229	0.986827662	45	23.16	agricultural and biological
12.	Bragantia	0.470588235	0.941176471	20	5.046667	agricultural and biological
13.	Plant Molecular Biology	0.892265193	0.987569061	137	30.036667	agronomyand crop
14.	Theoretical and Applied Genetics	0.802884615	0.990384615	144	41.586667	agronomyand crop
15.	Algal Research	0.861931365	0.984038308	20	23.58	agronomyand crop
16.	Radiocarbon	0.656147272	0.996055227	65	44.113333	earth and plantery
17.	Global Change Biology Bioenergy	0.885648504	0.990023024	167	51.74333	environmental
18.	Agricultural and Forest Meteorology	0.735238095	0.993333333	116	46.29333	agronomyand crop
19.	Molecular Plant Pathology	0.78406326	0.992092457	72	34.37333	agronomyand crop
20.	Molecular Plant-Microbe Interactions	0.839805825	0.991504854	118	49.27333	environmental
21.	Earth System Dynamics	0.781055901	0.976708075	13	50.87	Earth and Planetary
22.	Geoscientific Model Development	0.615015974	0.992811502	32	47.64667	Earth and Planetary

Table 5: List of "National Journals" as labeled by indexing services, Source: www.scimagojr.com; the

S.No	Journal's Name	OCQ	NLIQ	H-index	IC	
1.	Journal of Plant Biochemistry and Biotechnology	0.726530612	0.955102041	20	11.216667	agronomyand crop
2.	Indian Journal of Agricultural Sciences	0.578947368	0.947368421	218	6.21	agronomyand crop
3.	Indian Journal of Agronomy	0.785714286	0.904761905	16	3.06	agronomyand crop
4.	Research Journal of Chemistry and Environment	0.761904762	0.940476191	10	8.03333	environmental
5.	Journal of Biopesticides	0.933962264	0.952830189	9	7.416667	agronomyand crop
6.	Range Management and Agroforestry	1	0.75	2	0.793333	agronomyand crop
7.	Colourage	0.6	0.6	20	0.67	environmental
8.	International Journal of Agricultural and Statistical Sciences	0.4	0.88	3	3.843333	agricultural and biological
9.	Disaster Advances	0.756521739	0.939130435	11	7.74	earth and planetary
10.	Nature Environment and Pollution Technology	0.555555556	0.833333333	5	12.03	environmental
11.	Journal of Agrometeorology	0.285714286	0.857142857	5	2.82	agronomyand crop
12.	Journal of Tropical Agriculture	0.157894737	0.947368421	9	4.4	agronomyand crop
13.	Legume Research	0.214285714	0.857142857	6	4.723333	agronomyand crop
14.	Annals of Biology	0.5	0.7	4	0	Agricultural and biological
15.	Journal of the Indian Society of Remote Sensing	0.774193548	0.949308756	21	12.836667	earth and planetary
16.	Annals of Agri Bio Research	0.25	0.75	3	0	agricultural and biological
17.	Sugar Technology	0.804878049	0.93902439	15	10.37	agronomyand crop
18.	Allelopathy Journal	0.724137931	0.931034483	22	11.88667	agronomyand crop
19.	Research on Crops.	1	0.8	4	9.83333	agronomyand crop
20.	Indian Journal of Agricultural Research	0.76	0.84	3	6.173333	agronomyand crop

deal with datasets with a large number of features and samples. Moreover, a large set of features increases computational complexity and thus makes the network impractical to use for an online data processing task. Though, it provides satisfactory results, the network parameters and their interconnections are difficult to interpret which leads to difficulty in managing it. To achieve better workability and efficiency, the neural network is combined with the concepts of fuzzy logic, thus creating a new paradigm called neuro-fuzzy network paradigm. This integration of fuzzy set theory with neural network enables the classification system to solve complex, real-life decision-making problems. Classification task is simplified by granularizing the input data into smaller granules or clusters. The process of computation of granules is accomplished by using the concept of fuzzy set theory and the process is called fuzzy information granulation. The architecture of the granular neural network, i.e. weights between nodes and node-to-node connectivity is derived using a set of rules extracted from the network. The advantages of GNN is to make the neural network structure transparent and thus opens the way to customize the network based on the classification task.

Unified Granular Neural Network (UGNN)[17] combines the effect of different GNN's and unifies them to achieve even higher accuracy during classification. UGNN model has 4 phases of operation. The first phase uses the input feature

Table 6: Journals and respective metrics: the indicators, OCQ, NLIQ, SNIP, JIMI and ICR have been computed by the algorithms described in section 5.3

Journal name	NLIQ	ICR	OCQ	SNIP	JIMI Index
ACM Computing Surveys	0.98	0.11	0.98	1.71	0.83
Journal of the ACM	0.96	0.13	0.97	0.78	0.77
Journal of Data and Information Quality	0.83	0.11	0.89	0.17	0.58
Journal of Experimental Algorithmics	0.9	0.16	0.83	0.2	0.65
ACM Journal on Emerging Technologies in Computing Systems	0.8	0.11	0.77	0.19	0.57
Journal on Computing and Cultural Heritage	0.78	0.15	0.8	0.22	0.58
ACM Transactions on Autonomous and Adaptive Systems	0.93	0.18	0.82	0.37	0.71
ACM Transactions on Accessible Computing	0.85	0.11	0.87	0.41	0.65
ACM Transactions on Architecture and Code Optimization	0.9	0.15	0.87	0.56	0.72
ACM Transactions on Algorithms	0.95	0.22	0.82	0.5	0.76
ACM Transactions on Asian and Low-Resource Language Information Processing	1	0.13	0.83	0.15	0.66
ACM Transactions on Applied Perception	0.9	0.15	0.88	0.24	0.66
ACM Transactions on Economics and Computation	1	0.16	0.64	0.45	0.73
ACM Transactions on Embedded Computing Systems	0.93	0.16	0.87	0.33	0.7
ACM Transactions on Interactive Intelligent Systems	1	0.18	0.73	0.49	0.76
ACM Transactions on Information and System Security	0.96	0.14	0.94	0.45	0.73
ACM Transactions on Intelligent Systems and Technology	0.97	0.17	0.88	0.76	0.79
ACM Transactions on Knowledge Discovery from Data	0.97	0.15	0.9	0.43	0.74
ACM Transactions on Management Information Systems	0.64	0.17	0.93	0.33	0.55
ACM Transactions on Computing Education	0.9	0.12	0.91	1.02	0.75
ACM Transactions on Computer-Human Interaction	0.96	0.16	0.93	1.37	0.83
ACM Transactions on Computer Systems	0.97	0.12	0.97	1	0.79
ACM Transactions on Computation Theory	0.93	0.14	0.76	0.21	0.65
ACM Transactions on Design Automation of Electronic Systems	0.9	0.15	0.87	0.33	0.68
ACM Transactions on Database Systems	0.93	0.16	0.95	0.39	0.72
ACM Transactions on Graphics	0.73	0.1	0.92	2.09	0.68
ACM Transactions on Information Systems	0.96	0.14	0.95	0.65	0.77
ACM Transactions on Internet Technology	0.98	0.16	0.92	0.31	0.72
ACM Transactions on Modeling and Computer Simulations	0.9	0.16	0.91	0.29	0.68
ACM Transactions on Mathematical Software	0.96	0.17	0.88	0.52	0.76
IEEE/ACM Transactions on Networking	0.9	0.2	0.96	0.42	0.72
ACM Transactions on Storage	1	0.09	0.73	0.14	0.63
ACM Transactions on Programming Languages and Systems	0.95	0.14	0.95	0.65	0.76
ACM Transactions on Storage	0.89	0.12	0.92	0.53	0.69
ACM Transactions on Software Engineering and Methodology	0.96	0.16	0.8	0.91	0.79
ACM Transactions on Sensor Networks	0.9	0.18	0.93	0.48	0.72
ACM Transactions on Spatial Algorithms and Systems	1	0.7	0	0.05	0
ACM Transactions on the Web	0.97	0.16	0.89	0.5	0.75

Table 7: The new set of International and National journals: The proposed binary classification scheme brings out the errors in labeling the class of many of the listed journals. This is due to the lack of granularity in labeling journals based on internationality. The binary labeling is superficial and incorrect. I stands for International and N stands for National Journals. The mismatch in labels is evident.

Journal name	Original Label	True Label
1.International Journal of Applied Agricultural Research	I	N
2.International Journal of Agriculture FoodScience and Technology	I	N
3.The International Journal of Horticulture and Crop Science Research(IJHCSR)	I	N
4.International Journal of Biotechnology and Biochemistry (IJBB)	I	N
5.International Journal of Molecular Genetics (IJMG)	I	I
6.International Journal of Fisheries and Aquaculture Sciences	I	N
7.Global Journal of Applied Agricultural Research	I	I
8.International Journal of Agricultural Economics and Management (IJAEM)	I	I
9.The Journal of Computational Intelligence in Bio-informatics (JCIB)	I	I
10.Advances in Computational Sciences and Technology	I	I
11.International Journal of Computational Intelligence Research (IJCIR)	I	N
12.Journal of Computer Science and Applications	N	N
13.International Journal of Information and Computation Technology	I	N
14.International Journal of Software Engineering	I	I
15.Current Development in Artificial Intelligence	I	N
16.Advances in Wireless and Mobile Communications	I	N
17.International Journal of Networking and Computer Engineering (IJNCE)	I	N
18.International Journal of Wireless Networks and Communications	I	N
19.Mathematical Modeling and Applied Computing	I	I
20.International Journal of Wireless Communication and Simulation	I	I
21.International Journal of Information Science and Education (IJISE)	I	N
22.International Journal of Computer and Internet Security	I	N
23.International Journal of Information Sciences and Application (IJISA)	I	N
24.International Journal of Advanced Computer Science and Technology (IJACST)	I	N
25.International Journal of Networks and Applications [IJN &A]	I	N
26.Global Journal of Computational Intelligence Research(GJCIR)	I	N
27.Advances in Computational Sciences and Technology	I	N
28.International Journal of Computational Physical Sciences [IJCPS]	I	N
29.International Journal of Nanoscience and Nanotechnology [IJNN]	I	I

vector and granularize it by fuzzy granulation techniques. The key idea behind the technique is to generate group of fuzzy granules for enhancing the transparency of the input. The fuzzy granulation can be Class-Supportive (CS) and Non-Class Supportive(NCS). CS granulation is used when data set for classification has overlapping classes. In the second phase, the informative fuzzy granules are used to train the network using back propagation algorithm. The trained network is utilized for extraction of rules using Kasabov Rules Extraction(KRE) method. GNN's are derived using these extracted rules. These GNN's are not fully connected and the node-to-node connection is based on the derived rule. The fourth phase combines the GNN's and the final class label of the input pattern is obtained via consensus decisions of

Table 8: The new set of International and National journals:

Journal name	Original Label	True Label
30.International Journal of Chemistry and Applications	I	N
31.International Journal of Physics and Applications	I	N
32.International Journal of Applied Engineering Research	I	I
33.International Journal of Dynamics of Fluids [IJDF]	I	N
34.International Journal of Pure and Applied Physics	I	I
35.International Journal of Applied Chemistry	I	N
36.International Journal of Computational Intelligence Research [IJCIR]	I	N
37.International Journal of Semiconductor Science and Technology	I	N
38.International Journal of Librarian-ship and Administration	I	I
39.Mathematics Applied in Science and Technology	I	N
40.International Journal of Petroleum Science and Technology	I	N
41.Advances in Aerospace Science and Applications	I	N
42.International Journal of Materials Physics [IJMP]	I	I
43.Global Journal of Pure and Applied Mathematics	I	I
44.Advances in Theoretical and Applied Mathematics	I	I
45.Advances in Fuzzy Mathematics [AFM]	I	N
46.Advances in Dynamical Systems and Applications	I	I
47.International Journal of Difference Equations [IJDE]	I	I
48.Advances in Applied Mathematical Analysis [AAMA]	I	N
49.Journal of Wavelet Theory and Applications [JWTA]	I	N
50.International Journal of Statistics and Systems [IJSS]	I	N
51.Advances in Algebra [AA]	I	I
52.International Journal of Pure and Applied Mathematical Sciences [IJPAMS]	I	I
53.International Journal of Applied Mathematical Sciences [JAMS]	I	N
54.Global Journal of Mathematics and Mathematical Sciences [GJMMS]	I	N
55.International Journal of Theoretical and Applied Computer Sciences [IJTACS]	I	N
56.Communication in Applied Geometry [CAG]	I	N
57.Advances in Applied Mathematical Biosciences	I	I
58.Mathematics Applied in Science and Technology	I	N
59.Mathematical Modeling and Applied Computing	I	I
60.International Journal of Mathematical Education	I	N
61.International Journal of Computational and Applied Mathematics [IJCAM]	I	I
62.International Journal of Computational Science and Mathematics [IJCSM]	I	N
63.International Journal of Mathematics Research[IJMR]	I	N
64.Communication in differential and Difference Equation [CDDE]	I	I
65.Global Journal of Mathematical Science: Theory and Practical [GJMS]	I	N
66.International Journal of Statistics and Analysis [IJSA]	I	I
67.Global Journal of Dynamical System and Applications	I	N
68.Global Journal of Computational Science and Mathematics (GJCSM)	I	N
69.Global Journal of Theoretical and Applied Mathematical Sciences	I	N

individual GNNs'. The 4 phases of UGNN operations are shown in Figure 5

In the current classification scenario, we have classified 'internationality' on a scale of low, medium and high. The input data is a feature vector comprising of IC, SNIP, Impact Factor, Other citation quotient and H-index. The data set is divided into two parts, one for training and the other for testing. We have used training set to train the network and testing set to validate the model. In the current investigation, 2 sets of data have been taken, 60% and 80% for training and remaining 40% and 20% for testing. To increase the performance of the

Table 9: The new set of International and National journals:

Journal name	Original Label	True Label
70.International Journal of Applied Mathematics and Mechanics (IJAMM)	I	N
71.International Journal of Fuzzy Mathematics and Systems (IJFMS)	I	I
72.Global Journal of Difference Equations [GJDE]	I	I
73.International Journal of Applied Environmental Sciences (IJAES)	I	N
74.International Journal of Environmental Engineering and Management [IJEEM]	I	N
75.International Journal of Environmental Sci. Development and Monitoring [IJESDM]	I	N
76.International Journal of Oceans and Oceanography	I	N
77.International Journal of Lakes and Rivers [IJLR]	I	N
78.Global Journal of Applied Environmental Sciences	I	I
79.International Journal of Environmental Researchand Development [IJERD]	I	I
80.International Journal of Information and Computation Technology [IJICT]	I	N
81.Advances in Computational Sciences and Technology	I	N
82.International Journal of Information Science and Education [IJISE]	I	I
83.International Journal of Information Science and Application [IJISA]	I	N
84.Global Journal of Business Management and Information Technology [GJBMIT]	I	I
85.International Journal of Operations Management and Information Tech. [IJOMIT]	I	I
86.International Journal of Knowledge Management and Information Tech. [IJKMIT]	I	N
87.International Journal of Information Technology and Library Science [IJITLS]	I	I
88.International Journal of Education and Information Studies [IJEIS]	I	N
89.International Journal of Applied Engineering Research	I	I
90.International Journal of Engineering Studies [IJES]	I	N
91.Advances in Computational Sciences and Technology	I	N
92.Advances in Wireless and Mobile Communications	I	N
93.International Journal of Dynamics of Fluids [IJDF]	I	N
94.International Journal of Materials Science [IJoMS]	I	N
95.International Journal of Mechanics and Solids	I	I
96.International Journal of Nanotechnology and Application [IJNA]	I	I
97.International Journal of Theoretical and Applied Mechanics [IJTAM]	I	I
98.International Journal of Semiconductor Science and Technology [IJSST]	I	N
99.International Journal of Engineering Researchand Technology [IJERT]	I	N

classification model, 10-fold cross-validation is also used. This data set is run on 6 different models. Model 1 and Model 2 uses KNN classification with ungranulated and CS granulated feature vectors respectively. Model 3 and Model 4 uses Back propagation algorithm with ungranulated and CS granulated feature vectors respectively. Model 5 and Model 6 uses GNN and UGNN with CS granulated feature vectors as input to the models. The results are as shown:

UGNN, when applied on class supportive granulated feature vector for 60% and 80% training samples give 94.3% and 96.1% accuracy respectively. The results of classification are shown in Table 8.

CONCLUSION

Classification of journals using the existing parameters has been an open problem in the field of Scientometrics since long. The authors of this paper attempt to break the glass ceiling of classifying journals as "international" and "national" and also dispel the binarization doctrine using adequate proofs and results. It is observed during the course of this work that Classifying journals cannot be a binary classification problem because of the granularity it deals with. The approach adopted was to question the Binarization doctrine. To begin with, a lot of research went into finding out the existing mechanism of classification of journals. Subsequently, a huge lacunae is discovered as no specific parameters or metrics could quantify "internationality" at journal level. This thought evoked the process of adding remedial metrics like NLIQ, ICQ, H-index

Table 10: The new set of International and National journals

Journal name	Original Label	True Label
100.International Journal of Mechanics and Thermodynamics [IJMT]	I	N
101.International Journal of Civil Engineering Research[IJCER]	I	I
102.International Journal of Chemical Engineering Research [IJChER]	I	I
103.International Journal of Industrial and Production Engineering and Tech. [IJIPET]	I	N
104.International Journal of Instrumentation Science and Engineering [JISE]	I	N
105.Advance in Applied Computational Mechanics [AACM]	I	I
106.International Journal of Fluids Engineering [IJFE]	I	N
107.International Journal of Mechanics Structural [IJMS]	I	N
108.International Journal of Civil Mechanical Engineering [IJCME]	I	N
109.International Journal of Mechanical Engineering and Research [IJMER]	I	N
110.International Journal of Mechanical and Material Sciences Research[IJMMSR]	I	N
111.International Journal of Material Sciences and Technology [IJMST]	I	I
112.International Journal of Computational Physical Sciences [IJCPS]	I	I
113.International Journal of Nanoscience and Nanotechnology [IJNN]	I	N
114.International Journal of Chemistry and Applications [IJCA]	I	N
115.International Journal of Physics and Applications [IJPA]	I	I
116.International Journal of Pure and Applied Physics [IJPAP]	I	N
117.International Journal of Applied Chemistry [IJAC]	I	N
118.International Journal of Computational Intelligence Research [IJCIR]	I	N
119.International Journal of Advanced Mechanical Engineering (IJAME)	I	I
120.International Journal of Applied Physics [IJAP]	I	I
121.International Journal of Materials Physics [IJMP]	I	N
122.International Journal of Advanced Materials Sciences (IJAMS)	I	N
123.International Journal of Civil Engineering and Applications (IJCEA)	I	N
124.International Journal of Engineering and Manufacturing Science (IJEMS)	I	N
125.International Review of Applied Engineering Research[IRAER]	I	N
126.Global Journal of Academic Librarianship [GJAL]	I	N
127.International Journal of Librarianship and Administration [IJLA]	I	I
128.International Journal of Library Automation, Networking and Consortia [IJLANC]	I	N
129.International Journal of Information Technology and Library Sciences [IJITLS]	I	N
130.International Journal of Digital Libraries and Knowledge Management [IJDLKM]	I	I

and OC. Exclusive Scraping algorithms are written to find metric information from the web. This required scrutinizing of web pages at journal level and even further. Data acquisition of journal names, origin, article names, author names and their affiliations were the first few steps in tailoring the parameters of NLIQ, ICQ and IC. The values of these indicators are successfully implemented and cross validated in earlier work.[1] Data acquisition and Processing are therefore significant contributions of the paper and results have been added in appendix.

Once the data set was ready, the next challenge was how to attack the premises of "internationality" as binary classification problem owing to its granularity. The entire paper works on two tracks that is "the problem" and "method adopted". Bayesian classification was the adopted approach to solve this 2- class problem where linearly separable and Shapiro-wilk test were applied on the data set of 42 journals, 20 national and 22 international to prove that the samples $x_1, x_2 \ldots x_n$ are from a normally distributed population. Further, four remedial metrics (NLIQ, ICQ, OC and H-index) were given as input to the feature vector and we applied multivariate normal distribution. Finally, the minimum error rate classifier was used to classify the journals on the basis of values of these features. The classifier shows 80%-90% accuracy. Also, the probability of error in two class problem is the least making it the most suitable approach to be followed. This validated the first part of the problem. Now, few interesting observations were made while classifying journals (listed in table 4). where frequently listed as "national" by our classifier but actually they are "international" by name. The doctrine of "binarization" was first ac-

Table 11: The new set of National journals.

Journal name	Original Label	True Label
1.Indian Journal of Medical Research	N	I
2.Indian Journal of Experimental Biology	N	I
3.Indian Journal of Pharmacology	N	I
4.Journal of Chemical Sciences	N	I
5.Indian Journal of Physiology and Pharmacology	N	I
6.Indian Journal of Pediatrics	N	I
7.Indian Journal of Ophthalmology	N	I
8.Indian Journal of Chemistry -Section B Organic and Medicinal Chemistry	N	I
9.Pramana -Journal of Physics	N	I
10.Indian Journal of Gastroenterology	N	I
11.Indian Journal of Medical Sciences	N	I
12.Indian Journal of Chemistry -Section A Inorganic, Physical, Theoretical and Analytical Chemistry	N	I
13.Indian Journal of Pure and Applied Physics	N	I
14.Indian Journal of Biochemistry and Biophysics	N	I
15.Indian Journal of Pure and Applied Mathematics	N	I
16.Indian Journal of Clinical Biochemistry	N	I
17.Indian Journal of Microbiology	N	I
18.Indian Journal of Biotechnology	N	I
19.Sankhya: The Indian Journal of Statistics	N	I
20.Indian Journal of Psychiatry	N	N
21.Indian Journal of Agricultural Sciences	N	N
22.Indian Journal of Orthopedics	N	N
23.Indian journal of public health	N	N
24.Indian Journal of Plastic Surgery	N	N
25.Indian Journal of Radio and Space Physics	N	N
26.Indian Journal of Human Genetics	N	N
27.Indian Journal of Agricultural Economics	N	N
28.Indian Journal of Geo-sciences	N	N
29.Indian Journal of Labor Economics	N	N
30.Indian Geo-technical Journal	N	N
31.Indian Journal of Earth Sciences	N	N
32.Indian Journal of Mathematics	N	N
33.Journal of the Indian Mathematical Society	N	N

cepted and even with reasonably good accuracy and modified features, contradictory results were observed. Hence it validates our assertion that since "the method" adopted was correct with results to prove it, the problem was with the "binary doctrine"itself i.e "the problem".

The authors in[13] have devised an explicit scoring scheme by exploiting models in production economics. The paper didn't propose any classification scheme but when the internationality scores of more than 200 journals were plotted using a single-parameter histogram, it was observed that the scores do not fit into a unambiguous binary discrimination. This was the empirical evidence that journal international-ity is too simple to be disseminated as a binary classification problem. This lays the foundation of the present manuscript. The authors humbly put across all data, experimental models and results for further judgment to the academic community. We firmly believe that "internationality" cannot be a binary classification problem and there must be a more granular understanding to it. Since parameters like "Influence" and "impact" are measured in levels of low and high, a novel approach to classify "internationality" similarly can be explored. Our Argument in a nutshell is the following: Instead of categorizing journals in two classes, multi class discrimination is a more reasonable and scientific approach so that young researchers know the

difference. Internationality, as a concept must relate to quality and influence of journals and help people refrain from submitting to so-called "International" journals! Moreover, frequent misclassification of journals based on internationality could be avoided if binary doctrine is rejected altogether. We believe the survey of 205 carefully chosen journals is good enough to prove our case! The new set of International and National journals makes The proposed binary classification scheme bring out the errors in labeling the class of many of the listed journals. This is due to the lack of granularity in labeling journals based on internationality. The binary labeling is superficial and incorrect. **I** stands for International and **N** stands for National Journals (Appendix I - Table 9). The authors conclude by advocating the need for a multi-class problem instead and propose to solve it in future.

Future research directions may also include promising comparative studies. It would mean different classification methods to be applied to the same internationality classification problem including discovery of superior method(s) under specific parameter settings, identifying those parameters and the reasons behind the superiority of certain methods (the research issue). Comparative analysis of different classification algorithms may be performed in future lending further credence to the theory proposed in the paper. Consistency of the proposed model may be sought from empirical results and comparisons derived from other binary classification algorithms such as Support Vector Machines (SVM), Linear Discriminant Analysis (LDA), Artificial Neural Networks (ANN). Random Forest (RF) etc. The comparison shall cross-validate the main method adopted in the manuscript.

REFERENCES

1. Ginde, G. (2016). Visualisation of massive data from scholarly Article and Journal Database A Novel Scheme. CoRR abs/1611.01152
2. G. Buchandiran, *An Exploratory Study of Indian Science and Technology Publication Output,* Department of Library and Information Science, Loyola Institute of Technology Chennai, ISSN 1522-0222, 2011.
3. Chia-Lin Changa, Michael McAleer, Les Oxley, *Coercive journal self citations, impact factor, Journal Influence and Article Influence,* Mathematics and Computers in Simulation, 93, 190197, 2013.
4. Gunther K. H. Zupanc, *Impact beyond the impact factor,* J Comp Physiology A, 200, 113116, 2014.
5. Ludo Waltman, Nees Jan van Eck, Thed N. van Leeuwen, Martijn S. Visser, *Some modifications to the SNIP journal impact indicator,* Journal of Informetrics, 7, 272-2, 2013.
6. Waltman L, Van Eck, NJ, Visser MS, Wouters P. The elephant in the room: The problem of quantifying productivity in evaluative sciento-metrics. Journal of Informetrics. 2016;10(2):671-674. doi:10.1016/j.joi.2015.12.008.
7. Liping Yu, Yuqing Chen, Yuntao Pan, Yishan Wu; *Research on the evaluation of academic journals based on structural equation modeling,* Journal of Informetrics, 3(4), 304-311, 2009.
8. Chiang Kao, *The Authorship and Internationality ofIn-dustrial Engineering Journals,* Scientometrics, 80 (3), 123-136, 2009.
9. Bora K, Saha S, Agrawal S, Safonova M, Routh SMN, Anand. *CD-HPF: New* Habitability Score Via Data Analytic Modeling, *Astronomy and Computing,* 17(2), 2016.
10. Gose, Earl, Richard Johnsonbaugh, Steve Jost, Pattern recognition and image analysis, *PHI,* 1997
11. Hart, Peter E and Stork, David G and Duda, Richard O, Pattern classification, *John Willey and Sons,* 2001
12. Snehanshu Saha's Machine Learning Page: https://sites.google.com/pes.edu/snehanshu/home, 2017.
13. Gouri, G., Saha, S., Mathur, A., Venkatagiri, S., Vadakkepat, S., Narasimhamurthy, A., Daya Sagar, B.S. (June 2016). ScientoBASE: A Framework and Model for Computing Scholastic Indicators of non local influence of Journals via Native Data Acquisition algorithms. Journal Of Scientometrics. 1-51. doi:10.1007/s11192-016-2006-2 http://link.springer.com/article/10.1007/s11192-016-2006-2
14. Scimago journal ranking and classification site: www.scimagojr.com, *accessed on 30/06/2017.*
15. Ginde, G., Saha, S., Balasubramaniam, Chitra., R.S, Harsha., Mathur, A., Daya Sagar, B. S., Narsimhamurthy, A. (August 2015). Mining massive databases for computation of scholastic indices - Model and Quantify internationality and influence diffusion of peer-reviewed journals. Proceedings of the Fourth National Conference of Institute of Scientometrics, SIoT.
16. Mark H. Zweig and Gregory Campbell; *Receiver-Operating Characteristic (ROC) Plots: A Fundamental Evaluation Tool in Clinical Medicine,* CLIN. CHEM. 1993; 39(4):561-77,.
17. D Arun Kumar, K Meher SK, Kanhar D, Kumari KP, "Unified granular neural networks for pattern classification" *Neurocomputing, vol.* 216,pp. 109-125,2016

Appendix 1:Result Of Classification (60% Training data).

Measures	fold-1	fold-2	fold-3	fold-4	fold-5	fold-6	fold-7	fold-8	fold-9	fold10	Average
accuracy	0.82	0.88	0.88	0.88	0.76	0.88	0.82	0.82	0.82	1	0.86
sensitivity	0.89	0.89	1	0.78	0.89	0.78	0.67	0.89	0.89	1	0.87
specificity	0.75	0.87	0.75	1	0.62	1	1	0.75	0.75	1	0.85

Algorithm 1: Minimum Error Rate classifier.

Input: d dimensional feature Vector X
Output: Class Label
Step 1: Train classifier with training samples.
Step 2: for $(k := 1; k = no.of classes;$ k++) do calculate mean vector μ_i
calculate covariance matrix $\sum_i$
calculate class conditional density $p(X \mid \omega_i)$
calculate prior probability $P(\omega_i)$
calculate posteriori probability $P(\omega_i \mid X) == \frac{p(X|\omega_i)P(\omega_i)}{p(X)}$
calculate expected loss or risk function $R(\alpha_i \mid X)$=1-$P(\omega_i \mid X)$
calculate discriminant function $g_i(X)$=$P(\omega_i \mid X)$
Step 3: calculate error bound
$\int P(\omega_1 \mid x)P(\omega_2 \mid x)p(x)dx \leq$ P(error) $\leq \int 2P(\omega_1 \mid x)P(\omega_2 \mid x)p(x)dx$
Step 4: if $g_1(X) > g_2(X)$ then signify that
$P(\omega_1 \mid X) > P(\omega_2 \mid X) \Rightarrow R(\alpha_i \mid X) < R(\alpha_i \mid X)$,
Assign the sample to international class
Step 5: else signify that
$P(\omega_1 \mid X) < P(\omega_2 \mid X) \Rightarrow R(\alpha_i \mid X) > R(\alpha_i \mid X)$,
Assign the sample to national class
Step 6: return class label

Algorithm 2: Scraping.

Input: Scrape$(journal, result)$
  $volumes \leftarrow journal[volumes]$
  for $volume \in volumes$ **do**
  $issues \leftarrow volumes[issues]$
  for $issue \in issues$ **do**
  $articles \leftarrow issues[articles]$
  for $article \in articles$ **do**
  $result[volume][issue][article][title] \leftarrow article[title]$
  $result[volume][issue][article][authors] \leftarrow article[authors]$
  $result[volume][issue][article][citations] \leftarrow article[citations]$
  end for
  end for
  end for
  return *result*
  end procedure

Algorithm 3: Driver Algo: Algorithm to extract various features and to compute Internationality Index of Journals: collect_genealogy_network_id(). Output of this algorithm consist of Ids of authors in genealogy network of an input author.

Input: Scraped data repository from ACM
Output: Features such as International Collaboration Ratio, SNIP, Other-Citations and Internationality Index
$JNames[] = Fetch\_Journal\_Names\_from\_Scraped\_Repository(ACM)$
for every journal: $JNames[i]$ do
    TotalCites = Get the totalcites value
    Get all the published articles/papers: $X[]$
    for every article: $X[i]$ do
        $JNames[i]$.Selfcites+=compute_SelfCitations($X[i]$)
    end for
    $x_1 = 1 - JNames[i].Selfcites/TotalCites$; compute other-citation-quotient
    $x_2 = compute\_Intl\_Collaboration\_Ratio(JNames[i])/100$; compute International Collaboration Ratio
    $x_3 = compute\_SNIP(JNames[i])/MaxSNIP$; compute SNIP
    $x_4 = compute\_NonLocalIQ(JNames[i])$; compute NLIQ
    $Internationlity\_index = StochasticCobbDouglasModel(JNames[i], x_1, x_2, x_3, x_4)$;compute JIMI ▷ refer section 6.2 for Stochastic Cobb-Douglas Model
end for

Algorithm 4: Self Citation Count

Input: Article/paper name (P) from Google Scholar
Output: self-citation count for article / paper (P)
Get all cited Papers for article/paper(P): $citedBy[]$
for Every cited paper: $citedBy[i]$ do
if $P.author\_name$ IN $citedBy[i].author\_names$ then
$IncrByOne(P.SelfCitationCount)$
end if
end for
return $SelfCitationCount$

Algorithm 5: compute_NonLocalIQ(): Algorithm to calculate Non-Local Influence Quotient

Input: $journal\_name, citation\_database$
Output: $NLIQ\ of journal\_name$
$A \leftarrow 0$ ▷ external citation count
$B \leftarrow 0$ ▷ internal citation count
$J\_articles \leftarrow [\ ]$ ▷ used to store articles in a journal
$count \leftarrow 0$
for each $article \in citation\_database$ **do** ▷ get all articles in a journal
    if $article[journal] = journal\_name$ **then**
        $J\_articles[count++] \leftarrow article$
    end if
end for
for each $article \in J\_articles$ **do** ▷ get count of internal, external cites
    for each $reference \in article[references]$ **do**
        if $reference \in ARTICLE\_TYPE$ **then** ▷ reference is an article
            if $reference[journal]$!= $journal\_name$ **then**
                $A \leftarrow A + 1$
            else
                $B \leftarrow B + 1$
            end if
        end if
    end for
end for
$NLIQ \leftarrow A\ /\ (A + B)$
return $NLIQ$

Algorithm 6: Intl_Collaboration_Ratio(JNames[i]): Algorithm to compute international collaboration ratio of a Journal

Input: Journal Name: J
URL to all the articles in that Journal : $J.all\_articles\_url[]$
Country information of the Journal: $J.contryName$
Output: %international collaboration ratio of Journal: J ▷ Compute the internationality weight of an article Based on the combination (eg: out of 5 authors 2 are from same rest from other) deduce the weight of the article from a predefined values for a given combination, Eg: For all authors from different countries weight=1, For all authors from same country weight = 0, For n/2 authors from one country and n/2 from others weight=0.5
$authAffs = []$
for Every article in $J.all\_articles\_url[i]$ **do**
    $Authors\_Affiliation \leftarrow Fetch\_Author\_Affiliations(article)$ ▷ Algorithm 7
    $authAffs.append(read\_author\_name\_and\_first\_affiliationinformation(Author\_Affiliation))$ ▷ Generate 2D array of i:author name, j:country name
    $iNtrNationality\_wt[i] = compute\_wt(article)$
end for
$J.iNtrNational[][]$ ▷ Create one big matrix for a journal where i:country names, j:author names
for every i in $authAffs$ **do**
    if $Country\_of(i['Affiliation']) == J.countryName)$ **then** ▷ if author's country same as Journal's country then make entry = 0
        $J.iNtrNational[Country\_of(i['Affiliation'])][i['Author']] = 0$
    else
        $J.iNtrNational[Country\_of(i['Affiliation'])][i['Author']] = 1$
    end if
end for
x = Ratio of(Number of 0's and Number of 1's in $J.iNtrNational[][]$)
y = cumulative weights($iNtrNationality\_wt[i]$)
return $(\%international\_collaboration = \alpha x + (1-\alpha)y)$ ▷ α is a weight deduced from cross correlation

Algorithm 7: Fetch_Author_Affiliations(article): Algorithm to fetch author affiliations information for the article

Sample Output of Algorithm 6:

Input: Author name and Country name from affiliation information
Intermediate Sets: [u'lei yang', u'pedro v sander', u'jason lawrence'] [u'Hong Kong', u'Honduras']
Mod of country set: 2 Mod of author set: 3
IC = 0.666666666667
Intermediate Sets: [u'jaewon kim', u'roarke horstmeyer'] [u'New Zealand']
Mod of country set: 1 Mod of author set: 2
IC = 0
Intermediate Sets: [u'jing dong', u'reza curtmola', u'cristina nita-rotaru'] [u'Mali', u'United States', u'Iceland']
Mod of country set: 3 Mod of author set: 3
IC = 1.0

```
Input: Link to the article from algorithm 5: article_URL
Output: Author names and respective Affiliations
authors[] ⟵ scraped_author_names(article_URL)
list = []                                                    ▷ list of dictionaries
for every author in authors[] do
    dictionary_element = {'Author' : author}
    count = 1
    for every affiliation of author do
        dictionary_element.update{'count' :affiliation}
        count = count + 1                                    ▷ First, Second, Third Affiliations
    end for
    list.append(dictionary_element)
end for
return list
```

```
Input: University string
Output: Affiliation of University
    Convert input string into lowercase
    Break the string into tokens and store in array list
    author = list[0]
    affiliation = list[1]
    author=line1[0].replace("","")
    affiliation=line1[1].replace("","")
    if 'university' or 'institute' in author then
    affiliation = author
    end if
    if length(inputstring) ≤ 75 then
    for i in list_of_words do
    if i in affiliation then
    new_affiliation = affiiation
    if 'in ' in new_affiliation then
    loc = new_affiliation.index('in ')
    new_affiliation = new_affiliation[0:loc]
    end if
    break
    else
    new_affiliation = 'NULL'
    end if
    end for
    print new_affiliation
    end if
    if length(inputstring) > 75 then
    for word in input string do
    if length(word) == 9 or word == 'university' or word == 'institute' or word == instituto' or word
            == 'universidade' or word == 'universitario' or word == 'universitat'then
    location = list.index('word')
    prefix_string = list[0:location]
    if 'department' in prefix_string then
    postfix_string = list[location:length(inputstring)]
    affiliation = postfix_string
    end if
    end if
    end for
    print affiliation
    end if
```

Algorithm 9: City and Country extraction for Affiliation

```
Input:University string
Output:City and Country of University
  temp_university =university_name_string
  ToAPI(temp_university)
  function ToAPI temp_university
  if "university_of" in temp_university then
  next_word = word after "university_of"
  temp_university = temp_university + next_word
  end if
  Send temp_university to Google Maps API
  Parse through Response.json to obtain City and Country
  end function
```

Sample Output of Algorithm 8

1) **Input** : mario_mezzanzanica department_of_statistics_and_quantitative_methods _crisp_research_centre_ university_of_milano-bicocca_italy
   Output : university of milano-bicocca italy
2) **Input** : m_kaiser university_of_augsburg
   Output : university of augsburg
3) **Input** : xiao-bai_li university_of_massachusetts_lowell
   Output : university of massachusetts lowell
4) **Input** : stuart_e_madnick massachusetts_institute_of_technology
   Output : massachusetts institute of technology
5) **Input** : stuart_e_madnick yang_w_lee
   Output : NULL

Sample Output from algorithm 9

1) **Input** : university of Milano-bicocca Italy
   Output : City:Milan Country:Italy
2) **Input** : university of Augsburg
   Output : City:Augsburg Country:Germany
3) **Input** : university of Massachusetts Lowell
   Output: City:Lowell Country:United States