ABSTRACT

ASSESSMENT OF SOCIETAL IMPACT OF RESEARCH

Harish Varma Siravuri, M.S.
Department of Computer Science
Northern Illinois University, 2018
Hamed Alhoori, Director

The unprecedented growth of scholarly literature published every year has affected many aspects of our lives. Despite the extensive studies of scholarly impact, there are broader impacts across society that remain underexplored. This thesis aims to predict the societal impact of research using information from a wide range of sources not limited to academic sources like citations. It identifies factors best suited to recognize scientific works that are most likely to be of interest to society. This has been achieved by building machine learning models that use three indicators of online attention: (1) whether a research article will be cited in public policy and the number of citations it is likely to receive (2) if a research article will be found newsworthy and the number of mentions it is likely to receive (3) public understanding of the research paper. This research also explores new approaches to measure the general public's understanding of scientific outcomes thereby enabling more accurate measurements of scientific literacy. Models were used to study relationships between public understanding of scientific outcomes and textual features extracted from scholarly text like average word length and average sentence length that are indicative of the text complexity.

NORTHERN ILLINOIS UNIVERSITY
DE KALB, ILLINOIS

MAY 2018

ASSESSMENT OF SOCIETAL IMPACT OF RESEARCH

BY

HARISH VARMA SIRAVURI

A THESIS SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE

MASTER OF SCIENCE

DEPARTMENT OF COMPUTER SCIENCE

Thesis Director:
Hamed Alhoori

ACKNOWLEDGEMENTS

Foremost, I would like to express my sincere gratitude to my advisor Dr. Hamed Alhoori for the continuous support of my M.S. study and research, for his patience, motivation, enthusiasm and immense knowledge. His guidance helped me with my research and my thesis. I cannot imagine having a better advisor and mentor for my research. Besides my advisor, I would like to thank the rest of my thesis committee: Dr. Kirk Duffin and Dr. Reva Freedman, for their encouragement and insightful comments.

I also thank my family for always supporting me and believing in me. Last but not the least, I thank my fellow labmates in the DATA Lab for the stimulating discussions, and for all the fun we have had in the last two years.

DEDICATION

This thesis is dedicated to my family.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The extensive growth in scientific literature has necessitated development of ways to assess societal impact of research. Traditional methods focusing on academic impact are insufficient to assess societal impact. Citation analysis limits itself by not accounting for other sources through which research receives attention [1]. While citations measure research impact within the world of academia, altmetrics (alternative metrics) make it possible to measure other forms of attention that research receives and its societal impact [2, 3]. This study uses altmetrics to investigate the potential impact of a given research output as measured by the number of citations it receives from public policy documents, the number of mentions it receives from news articles, and how well readers understand the research. Predicting the number of policy citations, news mentions and estimating public understanding of research will enable the identification of potential high impact work in its early stages. To this end, we have built machine learning models that investigate the possible existence of relationships between these three indicators of impact and altmetric features. Classifiers were used to predict whether a research output is likely to be cited by public policy documents and/or receive coverage from news outlets based on the attention it generates on a wide range of online platforms. The classifiers were evaluated based on their accuracy, precision and recall values. Regression models were also built and used to predict the extent of press coverage and public policy citations a scholarly text is likely to receive. Additionally, regression models were used to study any possible relationships between features extracted from the abstract sections of scholarly texts and the public understanding of the scholarly text. The public understanding of the scholarly text was estimated by calculating the semantic

cosine similarity between scholarly text and the textual content posted by the readers about it online. The regression models were evaluated based on their R^2 (coefficient of determination) values and Mean Squared Errors. The outcome of this study is an ensemble of models that together can be used to assess the potential relative societal impact of any research output.

1.1 Background

The massive increase in research being published every year [4, 5] and the demand for public resources to support that research beget the development of methods to evaluate the impact of research. It has become a matter of great importance to researchers to provide evidence that their work is likely to have a positive impact on society [6, 7]. Traditional techniques used to assess research do a good job of evaluating the academic impact of research [8, 9, 10] but pay little attention to measuring the holistic impact of scholarly outcomes and have several problems and limitations [11, 12, 13]. There exists a clear distinction between academic or scholarly impact and societal impact. Academic impact is an indicator of the contribution the research makes to its field in academia whereas societal impact is a broader term that applies to the world beyond academia. Academically brilliant research may have little or no direct impact on society and therefore positive scholarly impact might not necessarily always translate to equivalent societal impact. Research communities are looking for ways to measure the broader impact of research [14, 15]. In recent years, more attention has been given to the societal impact of research [16, 17, 18]. Earlier research into assessing ways of evaluating societal impact have found the traditional metrics like impact factors to be inadequate [19]. The quest to solve the challenges posed by the use of traditional techniques has led to the development of interesting alternatives like the Integrated Impact

Indicator (I3) [20]. Harzing and Van Der Wal [21] proposed the use of the Google Scholar h-index to assess journal impact in economics and business. The study compared the Google Scholar h-index with the ISI Journal Impact Factor and found the h-index to be a better indicator of impact because of the Journal Impact Factor's sensitivity to individual highly cited papers. There are problems with the h-index too and modifications have been proposed, such as the e-index [22]. These alternatives, despite addressing several challenges faced by traditional metrics, lack in evaluating societal impact of research.

1.2 Related Work

Citation analysis has been found to be inadequate at measuring societal impact [23] and various techniques have been developed to address this problem [16]. Primarily, three methods have been used to assess societal impact: econometric studies, surveys, and case studies [24]. A number of countries have adopted or are in the process of adopting a host of systems to evaluate the societal impact of research. Netherlands is believed to have one of the most developed examples of impact evaluation [25]. The focus primarily is on the economic impact of research. The Dutch model was considered robust enough to be practiced in other countries too [26]. According to van der Meulen and Rip [26], in the Netherlands more than 80% of documents from evaluation processes involved a societal impact assessment. The Standard Evaluation Protocols (SEP) in the Netherlands are laid down based on expert assessments [27]. SEP 2015 to 2021 [28] is now the fifth protocol which states relevance to society as one of the three criteria for assessment of research. The ERiC project [29] was also launched – a partnership between the Netherlands Association of Universities of Applied Sciences, the Royal Netherlands Academy of Arts and Sciences (KNAW), the Netherlands Organisation for Scientific Research (NWO), the Association

of Universities in the Netherlands (VSNU) and the Rathenau Institutes Science System Assessment department, that aims to evaluate the societal relevance of research. Other projects have also been initiated in recent years which aspire to improve the assessment of societal impact of research in certain fields. A study conducted at the Leiden University Medical Center to develop ways to assess societal impact demonstrated that the correlation between societal quality and scientific quality is weak. It demonstrated that:

> ...high scientific quality of research groups is not necessarily related to communication with society, and that in order to increase societal quality of research groups, additional activities are needed. Therefore societal quality is not simply the consequence of high scientific quality. Obviously, in a university medical centre, scientific quality prevails, and is a prerequisite, which cannot be replaced by aiming instead for high societal quality. [27]

Spaapen et al. [30] proposed the Research Embedment and Performance Profile (REPP) which represented a host of indicators relating to research using five dimensions of research. They are: science and certified knowledge, education and training, innovation and professionals, public policy and societal issues, and collaboration and visibility. The United Kingdom has replaced the previous national evaluation system Research Assessment Exercise (RAE) with the Research Excellence Framework (REF) in which, "the impact element will include all kinds of social, economic and cultural benefits and impacts beyond academia, arising from excellent research" [31]. This approach rests on the method developed for the Australian Research Quality Framework (RQF) which was recommended as the best practice by Grant et al. [32] in their report to the Higher Education Funding Council for England. The Australian government in 2007 replaced RQF with Excellence in Research for Australia (ERA) under which research was assessed by evaluation committees based on different indicators of research quality, volume, application, activity and recognition [33]. The tool published by the Danish Council for Research Policy also used indicators for societal significance in addition to quality-related indicators like citation counts to assess the quality of Danish research

[34]. A consortium of five Finnish public research organizations proposed five dimensions of socioeconomic impact of research [35]. They are: (a) impact on economy; (b) impact on knowledge, expertise, human capital, and management; (c) impact on networking and social capital; (d) impact on decision making and public discourse; and (e) impact on social and physical environment. The U.S. National Science Foundation (NSF) too asked reviewers in the late 1990s to evaluate applications based not only on the intellectual merit but also the broader impact [36, 37] defined as follows:

> How well does the activity advance discovery and understanding while promoting teaching, training, and learning? How well does the proposed activity broaden the participation of underrepresented groups (e.g., gender, ethnicity, disability, geographic, etc.)? To what extent will it enhance the infrastructure for research and education, such as facilities, instrumentation, networks and partnerships? Will the results be disseminated broadly to enhance scientific and technological understanding? What may be the benefits of the proposed activity to society? [38]

Most studies focused on assessing societal impact traditionally focus on the economic part of societal impact. This study instead focuses on societal impact from three different angles.

1.3 Dataset

The entire data used in this project has been provided by altmetric.com. The database dump consists of over 5 million articles that have been tracked by altmetric. In addition to information about scholarly citations and public policy citations, the dataset includes activity on social media platforms such as Facebook, Twitter, Reddit, Google+ and Weibo; data from online reference managers such as Mendeley, Citeulike and Connotea; Wikipedia; news outlets; blogs; and YouTube collectively known as altmetrics. Various studies have been conducted that make use of altmetrics [2, 39, 40, 41, 42]. In this study, we use altmetrics to assess and predict the societal impact of scholarly articles.

1.4 Objectives

The main objective of this thesis is to predict societal impact of a scholarly article. This objective has been divided into 5 smaller objectives that together accomplish the aim of this study. They are:

1. Predict whether a scholarly article is likely to be cited in public policy documents.
2. Predict the number of policy citations a scholarly article is likely to receive.
3. Predict whether a scholarly article is likely to be found newsworthy.
4. Predict the number of news mentions a scholarly article is likely to receive from online news outlets.
5. Estimate the public comprehension of a scholarly article represented by the semantic cosine similarity between text from the article and text posted by readers about it online.

1.5 Methodology

To achieve the objectives mentioned earlier, We made use of machine learning models. Classification and regression models were built using Scikit-learn [43] to this end.

In chapters 2 and 3, classifiers were used to predict if a scholarly article is likely to be cited by policy documents or mentioned in news articles and the regression models used to predict the number of policy citations and news mentions likely to be received. Three classifiers were used primarily – Multinomial Naive Bayes, Random Forest and Support Vector Machine. Random Forest and Support Vector Machine were chosen owing to their good performance

on real world classification problems [44]. The classifiers were evaluated based on their accuracy, precision and recall values. These metrics are derived from a confusion matrix as shown in Figure 1.1.

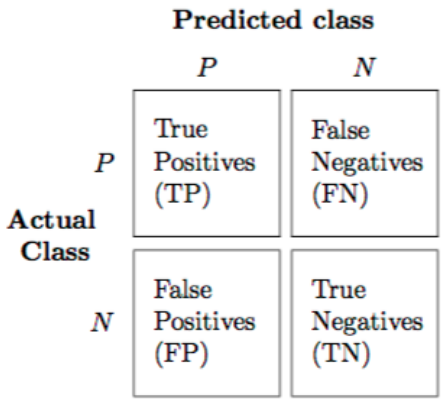

Figure 1.1: Confusion Matrix

Accuracy is the proportion of true results (both true positives and true negatives) among the total number of cases examined.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Precision is the fraction of the documents retrieved that are relevant to the user's information need.

$$Precision = \frac{TP}{TP + FP}$$

Recall is the fraction of the documents that are relevant to the query that are successfully retrieved.

$$Recall = \frac{TP}{TP + FN}$$

F1 Score is a measure of the test's accuracy that considers both precision and recall values.

$$F1 = \frac{2TP}{2TP + FP + FN}$$

The regression models were evaluated based on their coefficients of determination and Mean Squared Errors.

Coefficient of determination (R^2) is the proportion of the variance in the dependent variable that can be predicted from the independent variables. Given a data set with n values marked y_1,...,y_n (collectively known as y_i), each associated with a predicted value f_1,...,f_n (collectively known as f_i),

$$R^2 \equiv 1 - \frac{SS_{res}}{SS_{tot}}$$

Here, SS_{tot} represents the total sum of squares:

$$SS_{tot} = \sum_i (y_i - \overline{y})^2$$

where $\overline{y}$ is the mean of the observed data:

$$\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

SS_{res} represents the sum of squares of residuals:

$$SS_{res} = \sum_i (y_i - f_i)^2$$

The Mean Squared Error (MSE) of the model measures the average of the squares of errors in the predictions made by the model. For a vector of observed values Y and a vector of predicted values $\hat{Y}$ with n values each,

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

In chapter 4, regression models were also used to predict the extent to which the public understood a scholarly article as represented by the semantic similarity between scholarly text and text posted by the public online. The semantic similarity was represented by semantic cosine similarity [45] between the two bodies of text. The regression models used text complexity features as predictors to predict the semantic similarity.

The overall result was an ensemble of models that together can be used to assess the societal impact of a scholarly article based on the attention it receives from the world outside academia.

CHAPTER 2

PUBLIC POLICY

2.1 Introduction

The number of citations in public policy documents was chosen as the first measure of societal impact. Policy documents play a vital role in generating demand for scientific innovation [46]. By their very nature, they impact large sections of society [15]. Consequently, the research that provides the evidence upon which the policy is based indirectly impacts the same sections of the society. Therefore citations in policy could be considered a critical indicator of the societal impact of research. They also support the credibility of the author cited and the policy document itself [47].

Though research is currently underused in policy-making [48], evidence-based policy making is being encouraged in all areas of public service. Winterfeldt [49] presented a framework to bridge the gap between research and policy making. The ability to predict the likelihood of research output being essential to public policies in the future would help governments and other funding agencies in their pursuit to allocate available resources efficiently. In this study, altmetric data has been used to predict policy citations. The relationship between altmetrics and public policy citations has been studied earlier [50]. It was observed that altmetric data proved better compared to academic citations at predicting public policy citations.

Parts of this chapter were previously published [51]. My contributions consisted of (i) extracting data about documents not cited in public policy documents from a database

dump with information about scholarly articles, (ii) building the support vector machines, (iii) calculating the Gini importance of each feature.

2.2 Data

Initial analysis of the altmetric dataset showed that of the over 5 million articles, 89,350 had been cited in at least one policy document whereas 5,097,207 had not been included in a document of this kind. To create a balanced dataset for further analysis, along with the 89,350 articles that had been cited in a policy document, we randomly chose another 89,350 articles that had not been cited in a policy document. The result was a balanced dataset with approximately 180,000 records, half of which had been cited in policy documents.

The resulting dataset had a very rich set of features for each article. The total attention received by scholarly articles that have been cited in public policy and articles that have not been cited in public policy from different online sources are shown in Figure 2.1 and 2.2 respectively. In our analysis, we considered only features related to online attention. The data used consisted of mention counts on various online sources including reference managers, mainstream news outlets, blogs, peer-review platforms (e.g., PubPeer and Publons), social media, public policy documents, and Wikipedia. We used mention counts on Twitter, Facebook, Reddit, Mendeley, Google+, Wikipedia, Weibo, mainstream news outlets, blogs, videos, and peer review sites as features to build the classifiers. Yet, we left a few sources out of our account, including Connotea, which was discontinued in 2013, and Pinterest and Stackoverflow, which together contributed to less than 1% of the articles in the sample. We replaced the policy citation count with a binary class label denoting whether a given article had been cited in a policy document for classification.

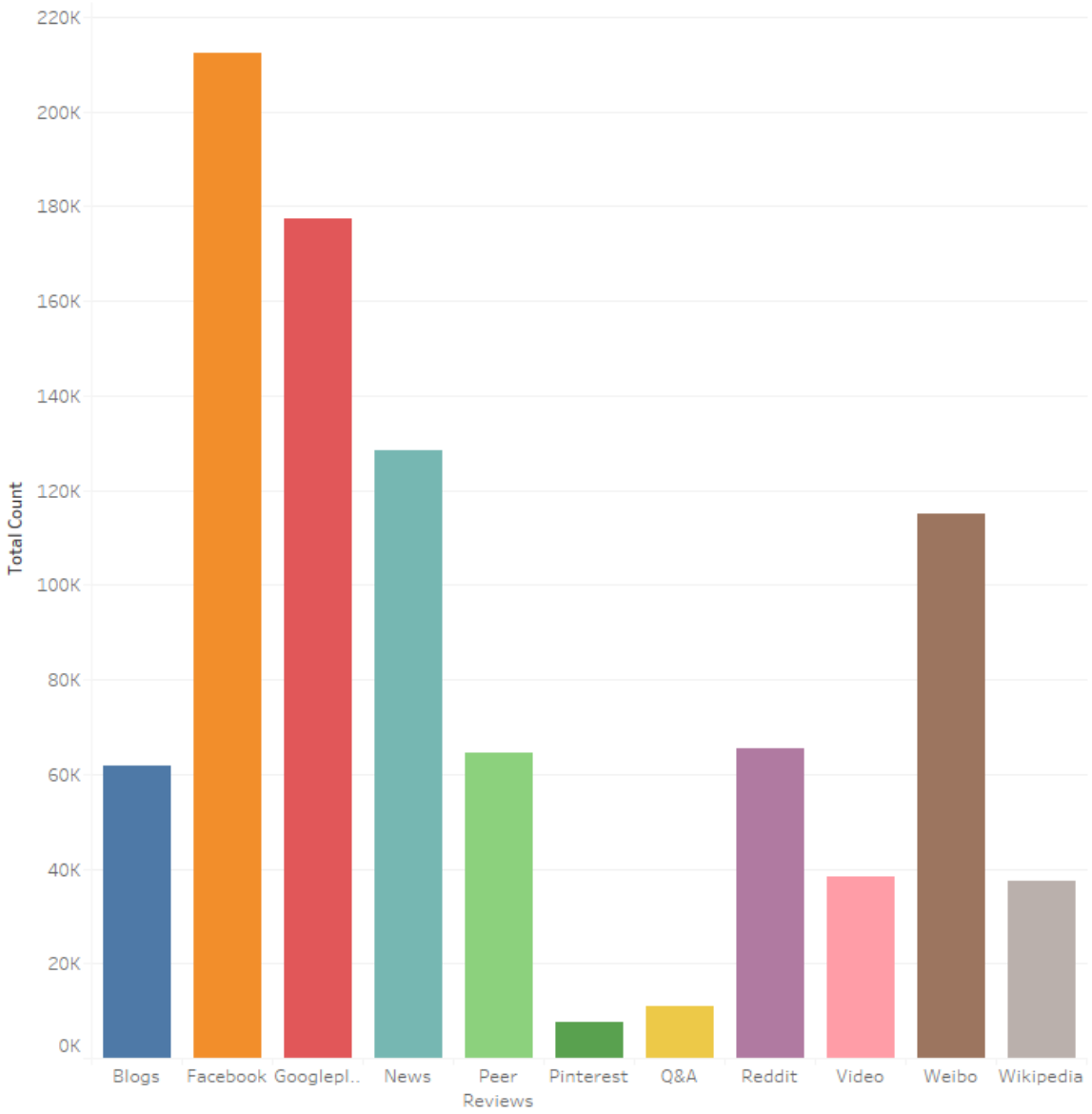

Figure 2.1: Online attention received by scholarly articles that have been cited in public policy documents

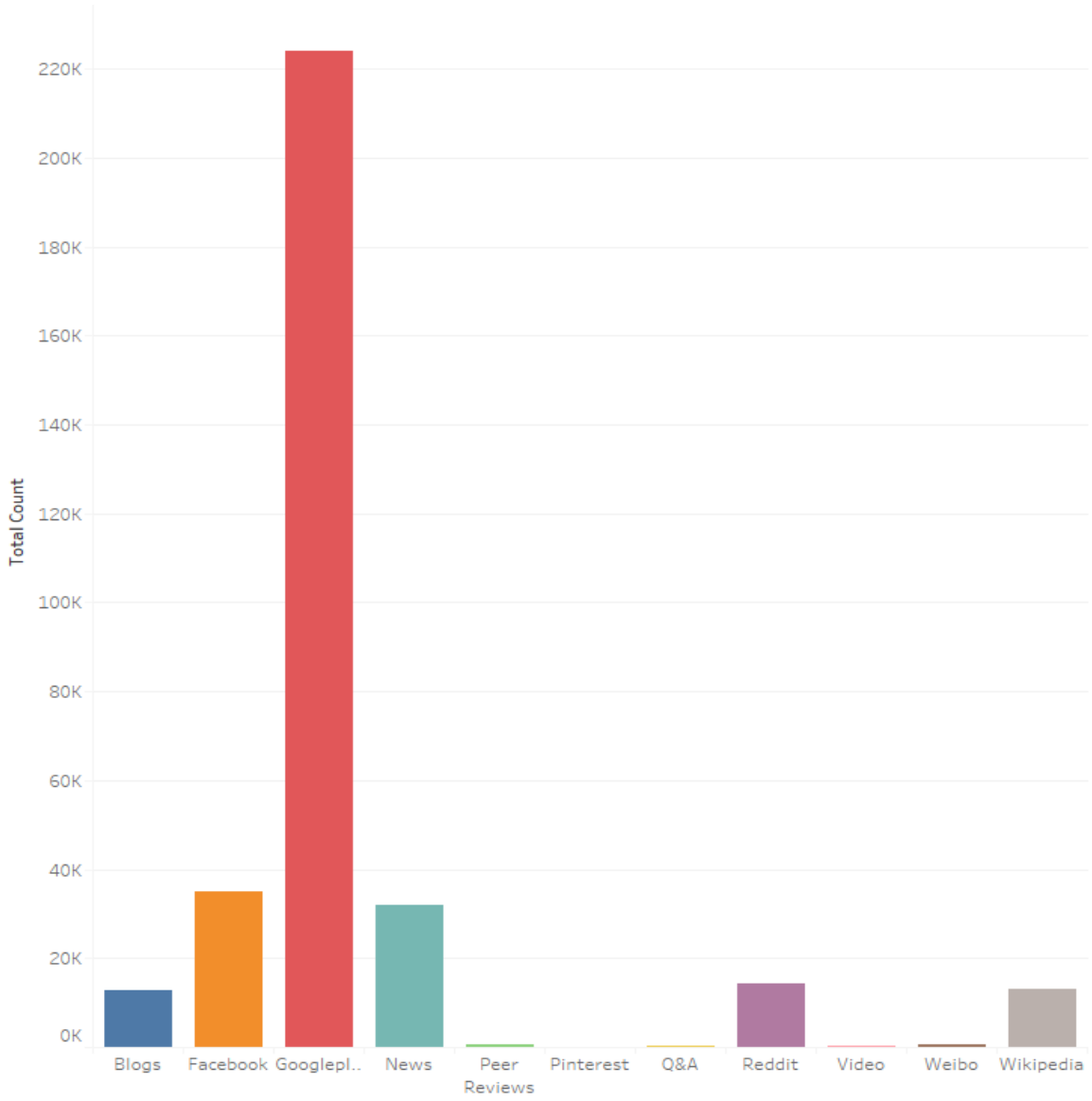

Figure 2.2: Online attention received by scholarly articles that have not been cited in public policy documents

2.2.1 Features

The following features were used as predictors in the classification and regression models in this chapter.

1. Peer Review - Number of peer reviews the article has received.
2. Google+ - Number of posts on the social media platform Google+ about the article.
3. Reddit - Number of Reddit threads that talk about the article.
4. Video - Number of YouTube videos on the article.
5. Twitter - Number of tweets that mention the article.
6. Weibo - Number of posts on Weibo about the article.
7. Mendeley - Number of readers on Mendeley who read the scholarly article.
8. Wikipedia - Number of Wikipedia pages that mention the article.
9. Blogs - Number of blogs that discuss the article.
10. Facebook - Number of posts on Facebook that mention the article.
11. News - Number of online news articles on the scholarly article.

2.3 Methods

We used binary classification to predict if a scholarly article is likely to be cited in public policy documents as described in Section 2.2.1. To predict the number of policy citations, we built regression models as described in Section 2.2.2.

2.3.1 Classification

To predict the likelihood of a research article being cited in a policy document, We implemented three classifiers: the Multinomial Naive Bayes classifier, the Random Forest classifier with the number of trees set at 100, and a C-Support Vector Machine with the Radial Basis Function (RBF) kernel. We then divided the entire dataset into training and test sets comprising 70% and 30% of the entire dataset, respectively. We trained the models using 10-fold cross-validation technique and evaluated them based on accuracy, precision, recall, and F1-measure metrics. The entire process has been depicted in Figure 2.3.

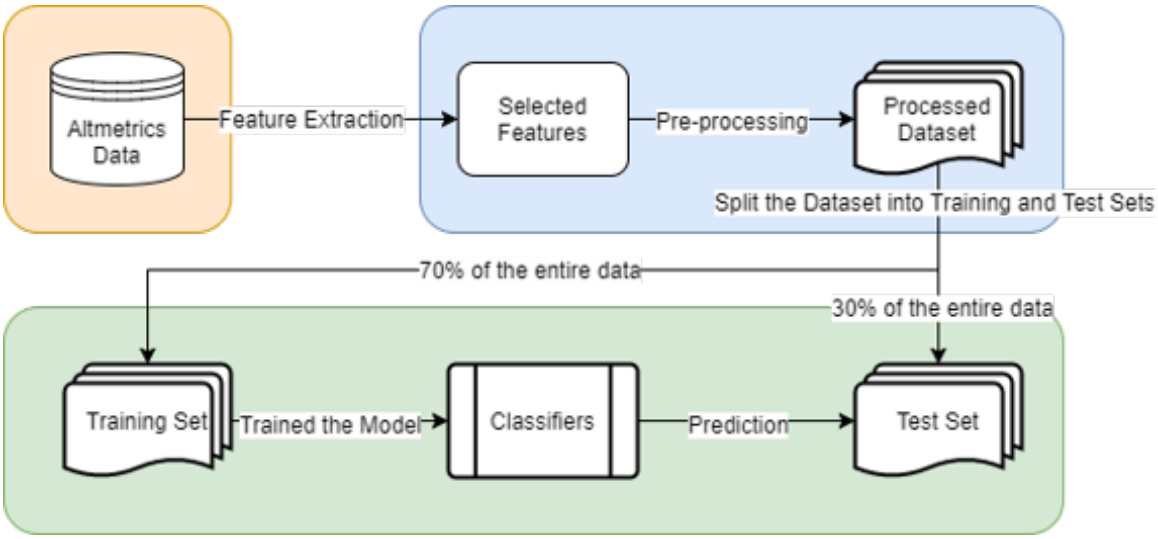

Figure 2.3: The classification process

With the classification models built, we also calculated the weight for each feature to determine the significance of each in making the final prediction as measured by its Gini importance [52]. Given that feature weights in the case of a Support Vector Machine can be determined only for linear kernels, we ranked the features based on their relevance for only the Random Forest and Multinomial Naive Bayes classifiers. We ranked the features

in regard to their importance to the Random Forest classifier from most to least important based on the Gini index. The importance of each feature with respect to the Multinomial Naive Bayes model has been represented by their coefficients.

2.3.2 Regression

To predict the number of public policy citations a scholarly article is likely to receive, we built regression models using the same features used for classification. The target variable used was the actual number of policy citations instead of the binary variable used for classification. The models were evaluated based on their coefficients of determination (R^2) and their Mean Squared Errors. The entire process has been depicted in Figure 2.4.

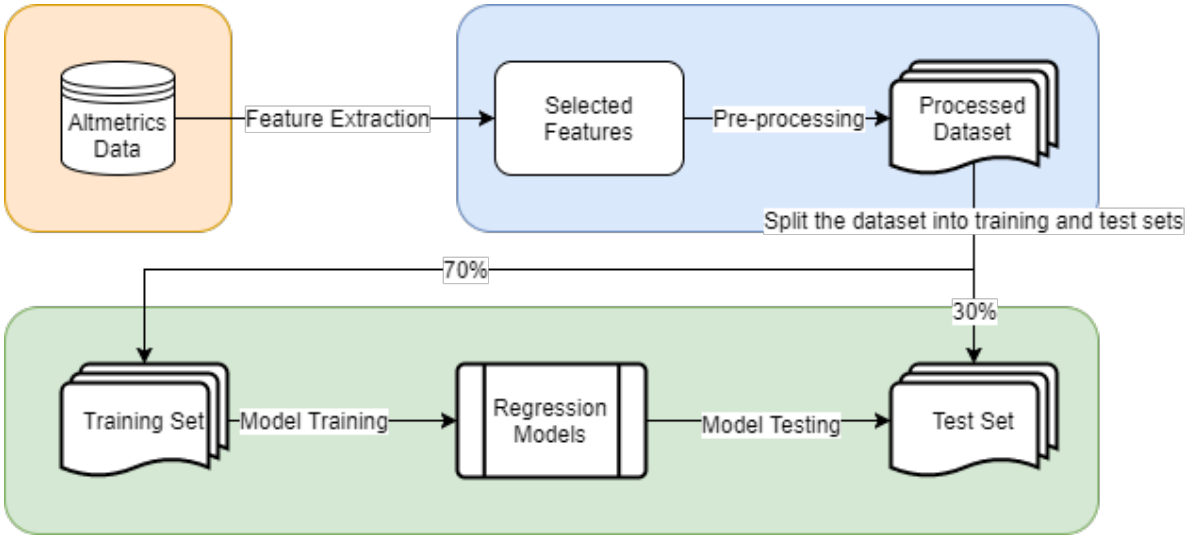

Figure 2.4: The Regression Process

2.4 Results

The result of this study is an ensemble of machine learning models that can be used to accurately predict if a scholarly article is likely to be cited in a public policy document or not. The results of the classification are shown in Table 2.1. The Random Forest performed best in terms of accuracy and precision, but the Multinomial Naive Bayes model performed better in terms of recall.

Table 2.1: Evaluation of classifiers used to predict policy citations

Model	Accuracy	Precision	Recall	F1-Measure
Multinomial Naive Bayes	0.842	0.802	0.905	0.850
Random Forest	0.870	0.826	0.870	0.844
Support Vector Machine	0.868	0.820	0.868	0.824

The Area Under the Curve (AUC) was also calculated by plotting the Receiver Operating Characteristic (ROC) curve which plots the model's true positive rate versus the false positive rate. This was done to calculate the accuracy of the classifier based on how well it separated the test set into those with and without policy citations. The ROC curves and the AUC of the classifiers have been compared in Figure 2.5. The Random Forest model performed best, with an AUC value of 0.98.

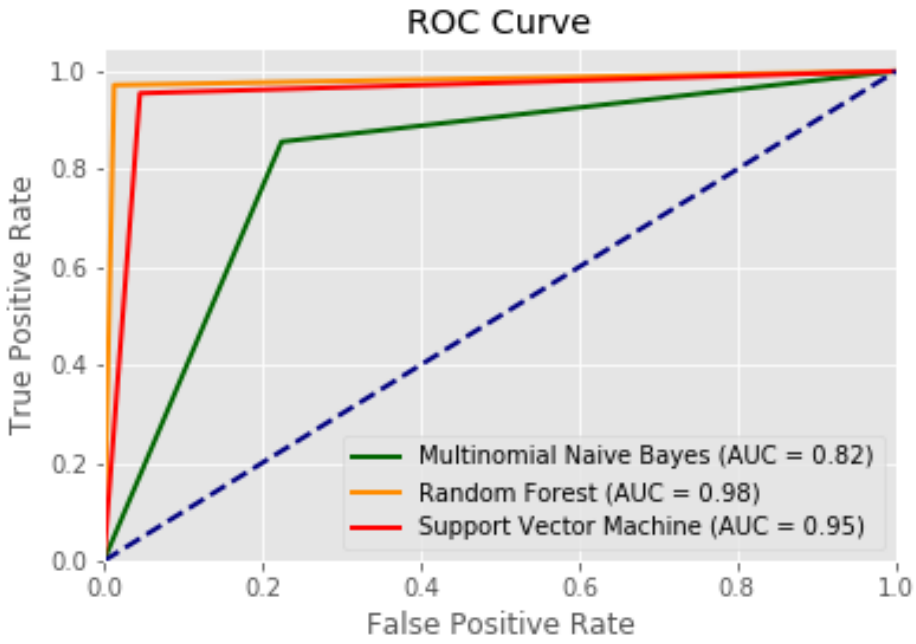

Figure 2.5: ROC for Public Policy Citations

We also plotted the precision-recall curve to study the tradeoff between precision and recall for different thresholds. The high Area Under the Curve in the case of Random Forest is an indicator of its high precision and recall further reinforcing its usefulness. The precision-recall curves for all 3 classifiers have been compared in Figure 2.6.

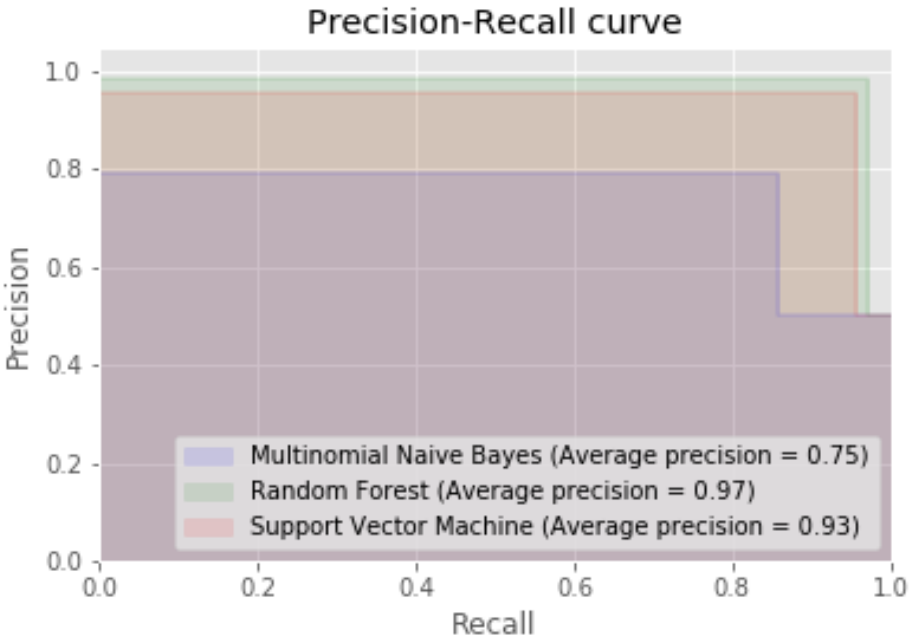

Figure 2.6: Precision - Recall Curve for Public Policy Citations

The Gini importance of each feature to the Random Forest classifier was also calculated to study which features played the most significant role in deciding whether a scholarly paper received a public policy citation or not. The values have been listed in Table 2.2. Peer reviews were the most important factor in the Random Forest model followed by the number of Google+ posts, Reddit threads, YouTube videos and tweets. This was slightly different from the Multinomial Naive Bayes model where the number of Wikipedia posts mattered most followed by YouTube videos, number of blogs, peer reviews and Reddit threads. The feature importances of each feature with respect to the Multinomial Naive Bayes model have been represented by the coefficients of each feature.

Table 2.2: Feature ranking for different classifiers used to predict policy citations

Platform	Random Forest (Gini Importance)	Multinomial Naive Bayes (Coefficients)
peer-review	0.273595	4.4267
Google+	0.197488	3.4210
Reddit	0.151016	4.4087
video	0.098035	4.9458
Twitter	0.068745	2.2421
Weibo	0.088242	3.7988
Mendeley	0.030116	0.3210
Wikipedia	0.026027	4.9668
blogs	0.018631	4.4571
Facebook	0.016189	3.2314
news	0.008926	3.7307

The relative importance of each feature to the model has been compared after normalization to a range of 0 to 1 as shown in Figure 2.7.

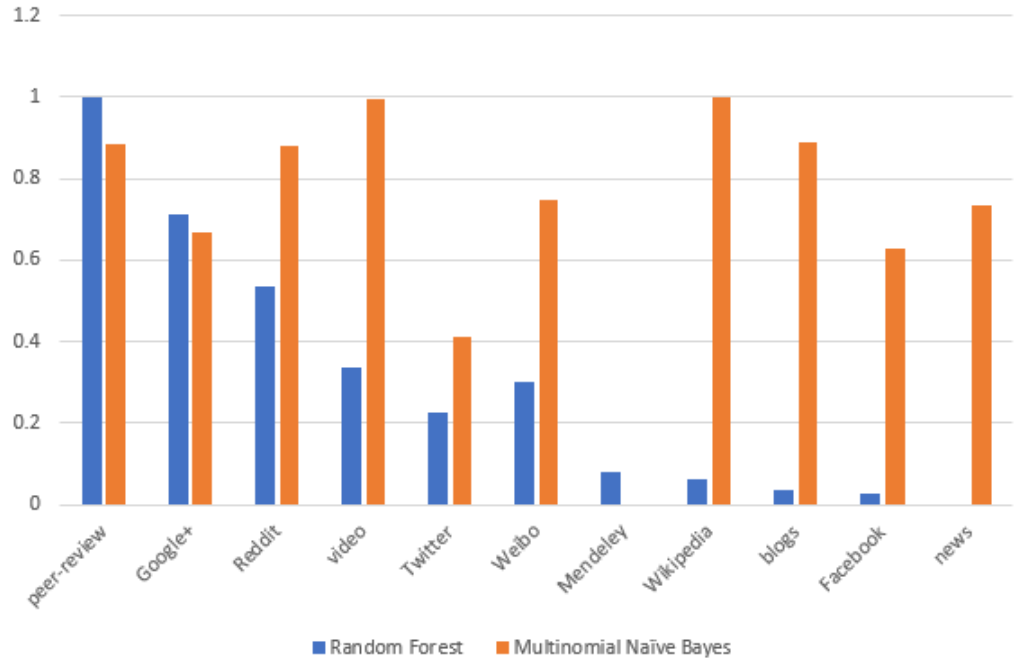

Figure 2.7: Comparison of relative importance of features to the models predicting policy citations

Also, if an article is predicted to receive citations from public policy documents, the number of such citations can also be predicted using the regression models which were evaluated based on their coefficients of determination and their Mean Squared Errors.

The coefficients of determination and the Mean Squared Errors of the regression models built in this chapter have been listed in Table 2.3.

Table 2.3: Evaluation of regression models used to predict policy citations

Model	R^2	Mean Squared Error
Linear Regression	0.6352	2.6063
Regression Tree	0.8741	0.5331
Support Vector Regression	0.7293	1.3126

The ability to make such predictions would help all the stakeholders involved in evaluation of research. It will assist funding agencies in their quest to identify research work that is likely to have significant impact on the society. It also makes it easier for policy makers engaged in evidence based policy making to find relevant research to build their policy upon.

CHAPTER 3

NEWSWORTHINESS

3.1 Introduction

The number of mentions research outputs receive from news outlets has been chosen as the second measure of societal impact of research. Being mentioned in stories published by news outlets is another approach to assessing the societal impact of research. Science news has become part of our daily lives and a crucial aspect of many general news outlets. Search engines have their own sections for science, medicine and/or health news, which include many references to the latest scientific findings. Research articles are increasingly being mentioned in online news stories and shared on other online platforms [53, 54, 55]. Researchers share links to those news stories on their websites as a sign of the societal impact of their work. According to Bornmann and Marx [56], research can be said to have societal impact when it is mentioned outside of scientific publications. The newsworthiness of research articles and social media metrics are good indicators of the societal impact of research [57, 58].

Research that has attracted the attention of the news media influences the perception of the relevance of research to society in general and becomes a topic of keen discussion among the public. With the rise of fact-checking journalism [59], reporters generally take steps to confirm findings by looking at other sources before reporting them as news. Any research that has the potential to significantly impact society would therefore be closely analysed and the inferences drawn by the public would not differ significantly from what the authors are trying to communicate.

Parts of this chapter were previously published [60].

3.2 Data

We randomly sampled 150,000 articles that had been mentioned in news articles with replacement. Due to problems such as missing data in other features, the actual number came down to 105,276. To create a balanced dataset for further analysis, along with the 105,276 articles that had been cited in a news article, we randomly chose another 105,276 articles that had not been mentioned in a news article. The result was a balanced dataset with more than 200,000 records, half of which had received attention from online news outlets in the form of mentions in news articles. The attention received by each set of articles from different sources is shown in Figures 3.1 and 3.2.

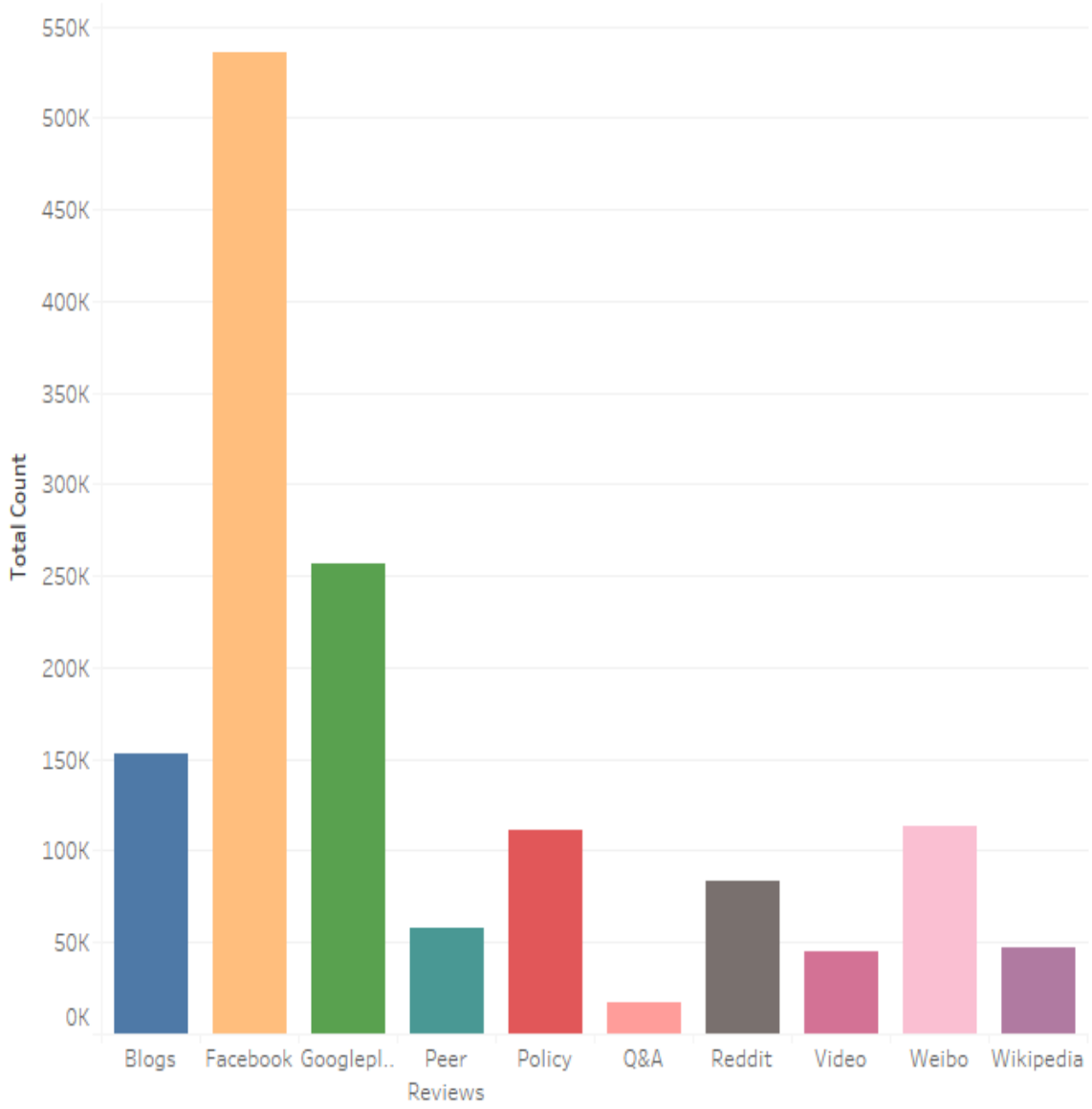

Figure 3.1: Online attention received by scholarly articles that have been mentioned in news articles

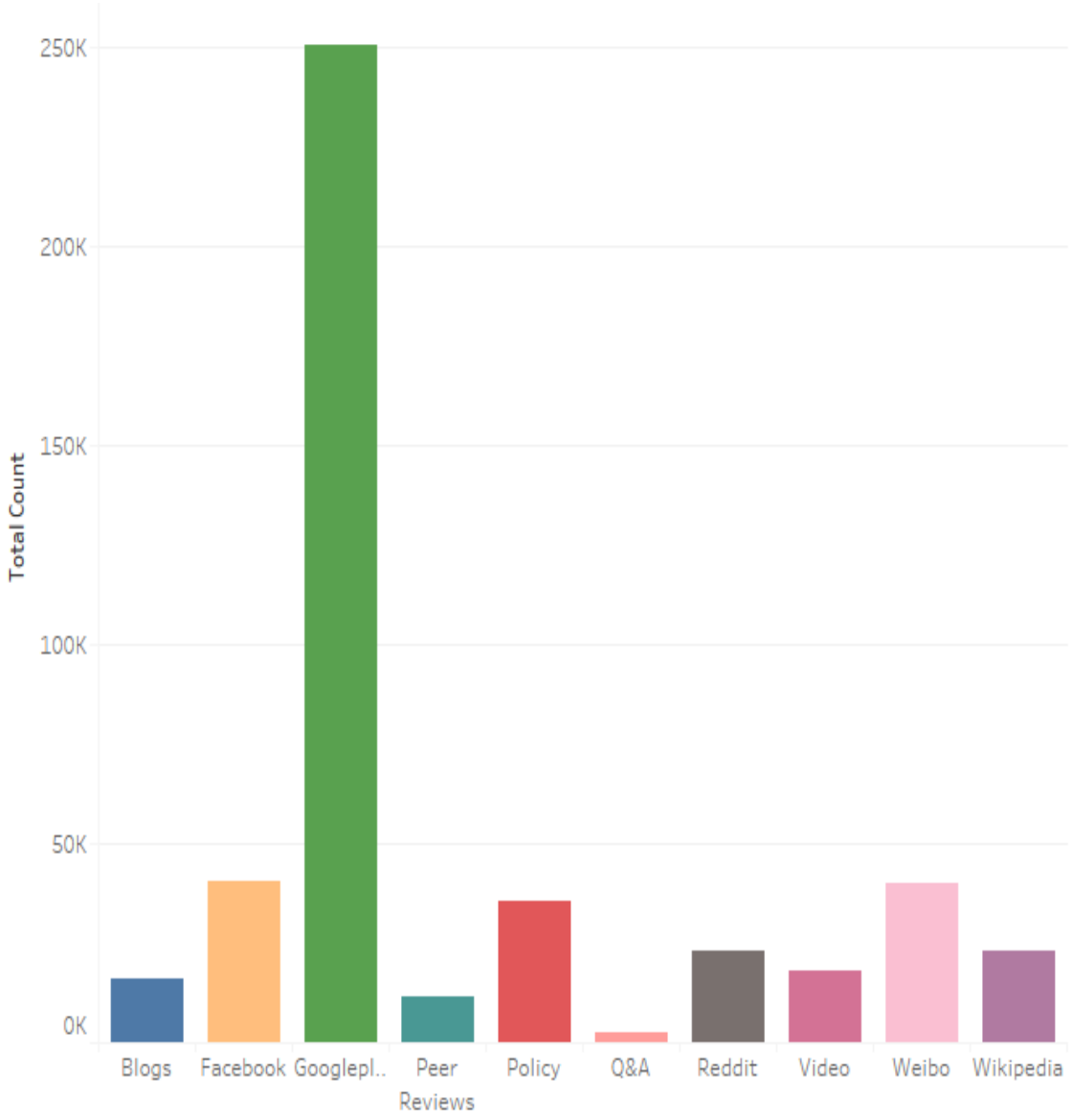

Figure 3.2: Online attention received by scholarly articles that have not been mentioned in news articles

3.2.1 Features

The following features were used as predictors in the classification and regression models in this chapter.

1. Peer Review - Number of peer reviews the article has received.
2. Google+ - Number of posts on the social media platform Google+ about the article.
3. Reddit - Number of reddit threads that talk about the article.
4. Video - Number of YouTube videos on the article.
5. Twitter - Number of tweets that mention the article.
6. Weibo - Number of posts on Weibo about the article.
7. Mendeley - Number of readers on Mendeley who read the scholarly article.
8. Wikipedia - Number of Wikipedia pages that mention the article.
9. Blogs - Number of blogs that discuss the article.
10. Facebook - Number of posts on Facebook that mention the article.
11. Policy - Number of public policy citations that the article has received.
12. QnA - Number of questions on StackOverflow relating to the article.

3.3 Methods

3.3.1 Classification

To predict the likelihood of a research article being mentioned in news, we implemented 2 classifiers: the Random Forest classifier with the number of trees set at 100, and a C-Support Vector Machine with the Radial Basis Function (RBF) kernel. we then divided the entire dataset into training and test sets comprising 70% and 30% of the entire dataset, respectively. we trained the models using 10-fold cross-validation technique and evaluated them based on accuracy, precision, recall, F1-measure, and Area Under the Curve (AUC) in the Receiver Operating Characteristic Curve (ROC). The entire process has been depicted in Figure 2.1.

With the classification models built, we also calculated the weight for each feature to determine the significance of each in making the final prediction. Given that feature weights in the case of a Support Vector Machine can be determined only for linear kernels, we ranked the features based on their relevance for only the Random Forest classifier. The importance of each feature with respect to the Random Forest model has been represented by its Gini index.

3.3.2 Regression

To predict the number of news mentions a scholarly article is likely to receive, we built regression models using the same features used for classification. The target variable used was the actual number of news mentions instead of the binary variable used for classification.

The models were evaluated based on their coefficients of determination (R^2) and their Mean Squared Errors. The entire process is depicted in Figure 2.2.

3.4 Results

The result of this experiment is a set of classification and regression models which can be used to accurately predict the extent of attention a scholarly article is likely to receive from news outlets. The classifiers help predict if a research work is likely to be found newsworthy. They were evaluated based on their accuracy, precision and recall values which have been listed in Table 3.1. The Random Forest model performed best overall with an accuracy of over 90% and higher recall and F1 scores.

Table 3.1: Evaluation of classifiers used to predict news mentions

Model	Accuracy	Precision	Recall	F1-Measure
Random Forest	0.924	0.796	0.658	0.720
Support Vector Machine	0.888	0.806	0.326	0.465

The models were also further evaluated based on the Area Under the Curve (AUC) by plotting the Receiver Operating Characteristic (ROC) curve which plots the model's true positive rate versus the false positive rate. The ROC curves and the AUC of each classifier have been shown in Figure 3.3. The Random Forest model performed best with an AUC value of 0.82.

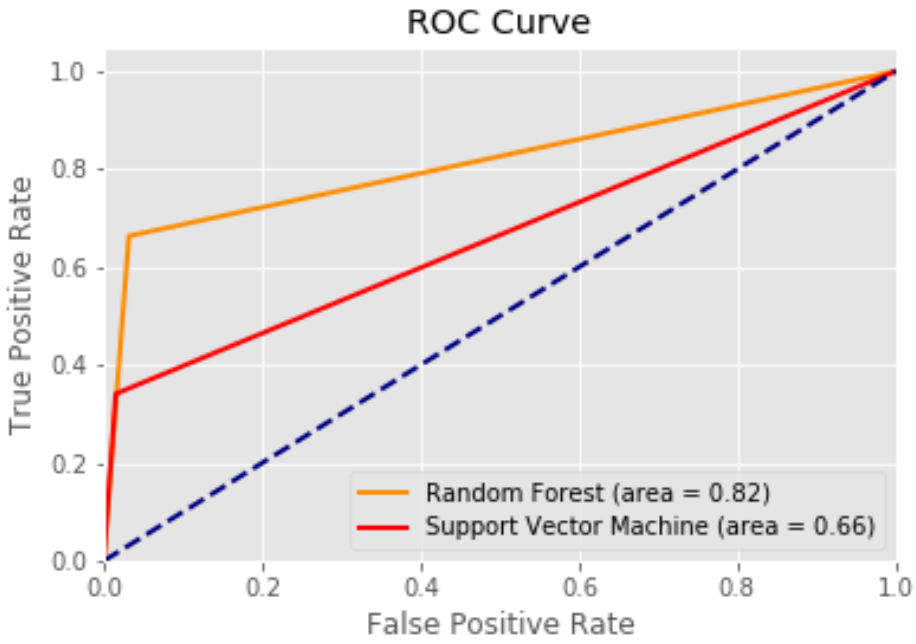

Figure 3.3: Comparison of ROC Curves for News Mentions

We also plotted the precision-recall curve to study the tradeoff between precision and recall for different threshold. The high Area Under the Curve in the case of Random Forest is an indicator of its higher precision and recall further reinforcing its usefulness. The precision-recall curves for both classifiers have been shown in Figure 3.4.

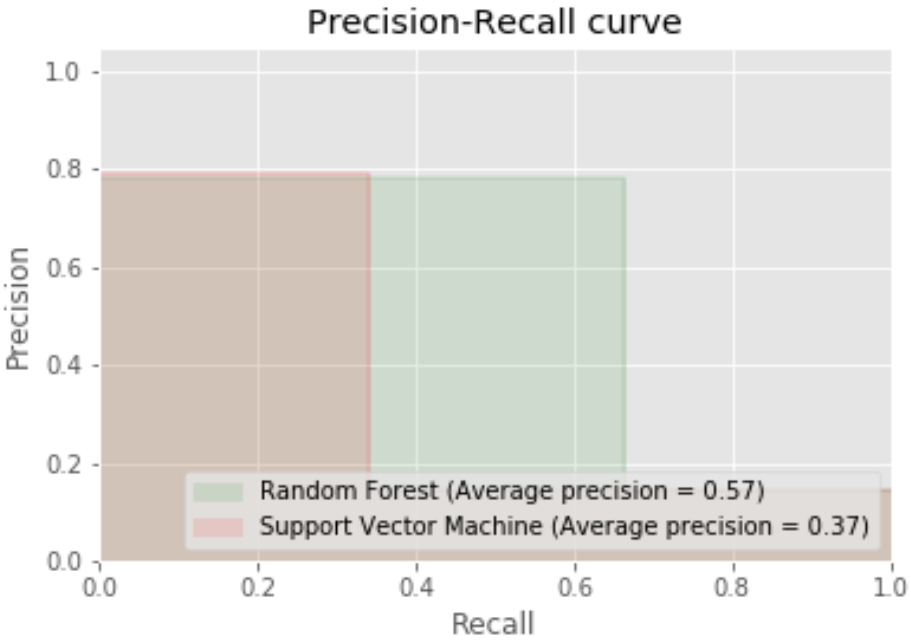

Figure 3.4: Precision – Recall Curves for News Mentions

Gini importance of each feature with respect to the Random Forest model was calculated to study which features were the most significant. It also helped study the differences in relevance of features in comparison to the importance of features with respect to policy citations. The values have been listed in Table 3.2. Mendeley was observed to be the most significant feature in the case of the Random Forest model followed by Facebook posts, tweets, blogs and Google+ posts.

Table 3.2: Feature ranking for classifiers used to predict news mentions

Platform	Random Forest (Gini Importance)
Mendeley	0.168083
Facebook	0.151553
Twitter	0.147885
Blogs	0.106562
Google+	0.093940
Wikipedia	0.060543
Reddit	0.048159
Peer Reviews	0.044373
Policy	0.042591
Weibo	0.035900
Video	0.031691
QnA	0.023858

Additionally, if an article is found newsworthy, the regression models can be used to predict the extent of attention it is likely to receive from news outlets represented by the number of mentions it receives in news articles. The models were evaluated based on their coefficients of determination and Mean Squared Errors which have been listed in Table 3.3.

Table 3.3: Evaluation of regression models used to predict news mentions

Model	R^2	Mean Squared Error
Linear Regression	0.4169	3.2655
Regression Tree	0.9102	0.8320
Support Vector Regression	0.6971	2.2192

The capability to make these predictions not only help estimate societal impact of research; they also help science journalists in their search for newsworthy research. With the deluge of scientific literature being published and the evolving nature of their collaborative relationships with their audiences, they need to find work that is likely to be of interest or relevance to their audience. These models can help narrow down the vast sea of literature they need to search for and make the job of finding newsworthy research easier.

CHAPTER 4

PUBLIC UNDERSTANDING OF SCIENCE

4.1 Introduction

The extent to which readers are likely to understand scientific text has been chosen as the third measure of its societal impact. Well understood and popular research are more likely to shape public opinion more than less popular research which could be academically brilliant. The Public Understanding of Science [61], a report published by the Royal Society, is widely considered to have given rise to the current interest in understanding scientific literacy. Promotion of public understanding of science has always been an important consideration associated with societal impact of research. Scientific literacy has been promoted as an important part of citizenship [62, 63]. According to McGinn and Roth [64], scientific literacy is an important quality in promoting good citizenship practices such as participation in scientific laboratories, activist movements, the judicial system, and other communities. Further, scientific literacy is a significant driver of economic growth, and for this reason virtually every modern society has shown a commitment to promoting scientific study and determining the publics understanding of scientific discoveries and advances.

Interest in this area is fueled by the widely held belief that science will be the ultimate beneficiary of any gains in the publics scientific literacy [65]. Studies have been conducted to investigate the relationship between text complexity and reading capability [66, 67]. The existence of a similar relationship between scientific texts and public understanding of science would allow identification of research that are likely to be well understood by the people.

Parts of this chapter will be published in [68].

4.2 Data

Initial analysis showed that of the 5.2 million articles, over 1.7 million articles had been shared and talked about on blogs. We further randomly sampled 1% of the dataset and extracted text from the abstract section and the blog posts to build a smaller dataset consisting of 17,736 data points for further analysis. We used regular expressions to filter out texts that contained only hyperlinks to the scholarly text. Also, we removed textual content that exactly matched the title or sentences from the abstract to avoid any bias caused by social media content in which only the scholarly output is noted without an accompanying discussion of it.

4.3 Methods

4.3.1 Feature Generation

We generated a set of five features - a target variable and four predictors which we later used to build the regression models. The entire process is demonstrated in Figure 4.1.

Target Variable: Since the objective is to predict public understanding of science, a target variable that is representative of the extent to which scientific text and textual content posted by the public about it mean the same is needed. Cosine similarity, though often used for similarity measurement between documents, does not suit the purpose of this study since it can not handle semantic similarity well [45]. The Wu and Palmer similarity is another popular method that owes its popularity to its computational speed [69], but it does not

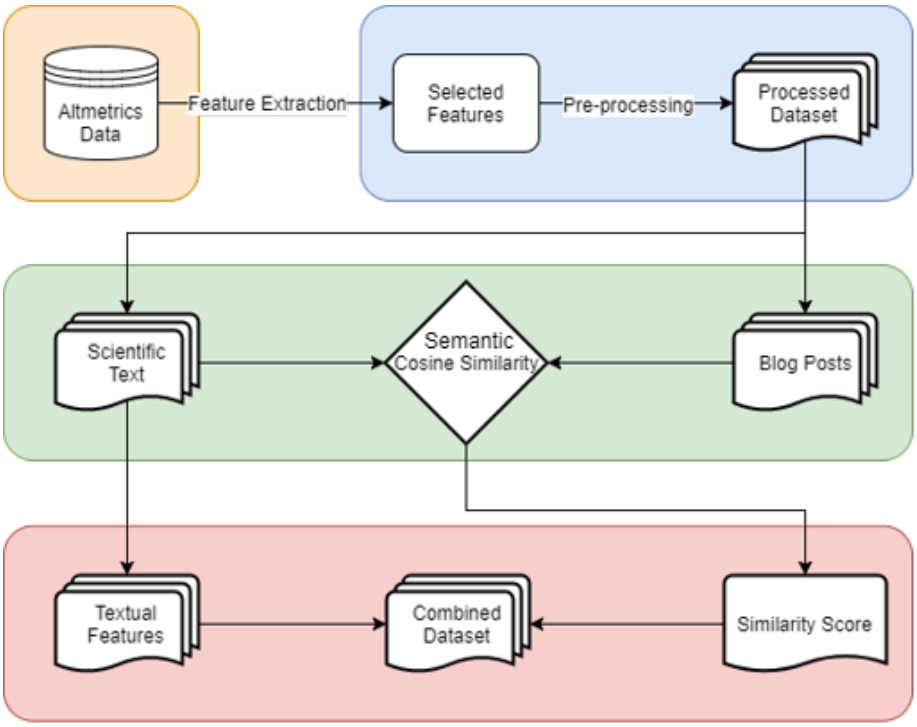

Figure 4.1: Feature Generation for Public Understanding of Science

consider how far apart the concepts are semantically [70]. The semantic cosine similarity [45] measure was finally chosen since it builds upon cosine similarity and enhances it by checking for synonym pairs using WordNet [71] and takes them into consideration when generating a similarity value. It considers semantic relation between the dimensions of two vectors which makes it suitable for this experiment. The performance of the semantic cosine similarity with a few test cases was also found reasonable based on human judgment.

Predictors: We also generated the following four features using the scientific text some of which are often considered when estimating the complexity of text [72] and others indicative of the writing style. They were used as predictors in the regression models.

1. Lexical diversity of the abstract - the ratio of unique word stems to the total words computed. It is an effective measure of the richness of vocabulary or verbal creativity of a text. We used Yule's measure [73] instead of a simple frequency-based measure,

since it yields an unbiased result by also taking the length of the text into consideration [74].

2. Average word length - the mean number of characters in each word in the abstract.

3. Average sentence length - the mean number of words in each sentence in the abstract.

4. Frequency of words longer than the average word length - a measure of the number of long words that have more characters than the average word in the abstract.

4.3.2 Regression

Using the processed data, we built five regression models: Decision Tree Regressor, Random Forest Regressor with 100 estimators, Support Vector Regressor, KNN Regressor, and a Gradient Boost Regressor. The models used the predictors generated in 4.3.1. to predict the comprehension score. The process has is depicted in Figure 4.2.

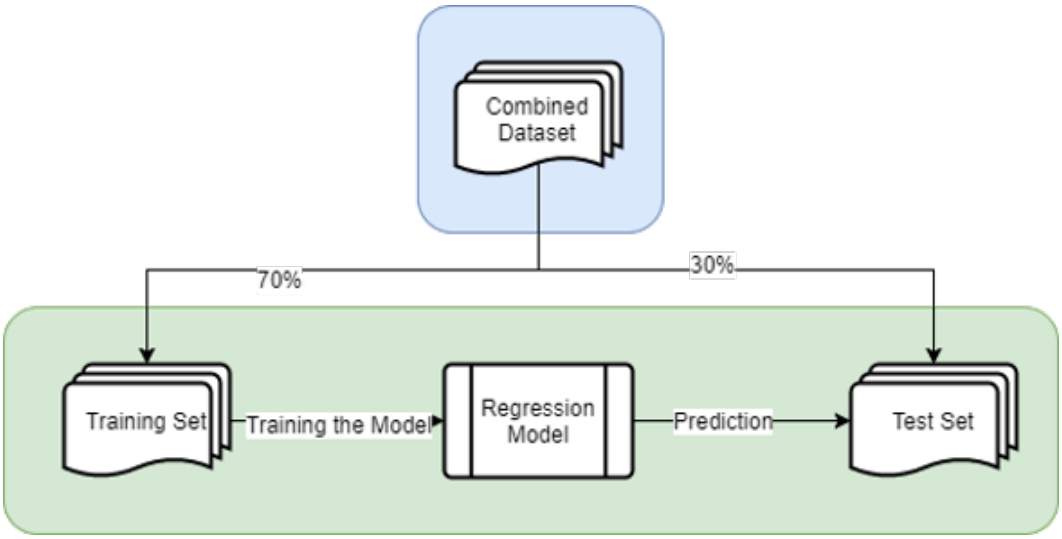

Figure 4.2: Regression model for Public Understanding of Science

4.4 Results

Table 4.1: Evaluation of regression models used to predict comprehension score

	R^2	Mean Squared Error
Decision Tree Regression	0.7356	0.0114
Random Forest Regression	0.6486	0.0074
Support Vector Regression	0.1202	0.0089
KNN Regression	0.0824	0.0074
Gradient Boost Regression	0.0609	0.0071

The Decision Tree Regressor and the Random Forest Regressor were observed to perform best compared to the other models. We calculated the Gini importance of each feature for both models to determine the relative significance of each feature to the public understanding. The results are shown in Table 4.2. The results show that using the Decision Tree Regressor,

Table 4.2: Importance of each feature to the regression models

	Decision Tree	Random Forest
Lexical Diversity	0.2588	0.2657
Average Word Length	0.3122	0.2799
Average Sentence Length	0.2476	0.2622
Frequency of words longer than average word length	0.1815	0.1922

the text complexity features used in this chapter can explain over 70% of the variance in how well readers understand a scientific article. These findings can be used to predict how well the public is likely to understand a given scientific text.

CHAPTER 5

CONCLUSION

5.1 Summary

The result of this entire study is a host of machine learning models that can be used by various stakeholders to assess the long term societal impact of research. This could be achieved by predicting factors that are indicative of societal impact using data from social media and other online platforms and the scientific text. This allows evaluation of research in a more broader sense than just the contribution it makes to the world of academia.

The idea to make societal impact a major part of the research evaluation process has been pursued by many governments and agencies. This thesis is a step in that direction. The machine learning models built as part of this work help predict three indicators of societal impact of research - use in public policy, attention received from news outlets, and the public's understanding of the work without the need of academic citations which often take years to accumulate.

5.2 Contribution

In chapter 2, we built models that predict if a scholarly article is likely to be cited by public policy documents and the number of policy citations it is likely to receive. In addition to helping funding agencies and other stakeholders identify research that is likely to have long term societal impact, these models also help policy makers identify relevant research

they can study and inculcate in their process of policymaking helping promote the practice of evidence based policy making.

In chapter 3, we built classifiers and regression models that predict if a scientific article is likely to be found newsworthy and the number of mentions it is likely to receive from news outlets. The models help identify research that is very likely to be of interest and relevance to the public. Science journalists are also likely to benefit from this study. With the increase in the amount of scholarly literature being published and increasing pressure to meet challenging deadlines, they can certainly make use of assistance in picking the right stories to be published.

In chapter 4, we built regression models that can estimate how well readers are likely to understand a given scientific text. This assists in the identification of research that is more likely to be understood easily by the public and improving their understanding of science. Governments and educators can make use of these models to identify research that will help improve public understanding of science and thus have significant societal impact.

5.3 Future Work

In the future, we plan to extend the work done in this thesis by studying additional indicators that can be used to assess the societal impact of research. Full texts of scholarly articles and additional text complexity features will also be used to better predict public understanding of science. We will also work on optimizing the models built in this thesis and improve their performance. The models will be deployed online available for public use.

REFERENCES

[1] E. Adie and W. Roe, "Altmetric: enriching scholarly content with article-level discussion and metrics," *Learned Publishing*, vol. 26, no. 1, pp. 11–17, 2013.

[2] H. Alhoori and R. Furuta, "Do altmetrics follow the crowd or does the crowd follow altmetrics?," in *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 375–378, IEEE Press, 2014.

[3] K. J. Holmberg, *Altmetrics for information professionals: Past, present and future*. 2015.

[4] S. Harnad, T. Brody, F. Vallieres, L. Carr, S. Hitchcock, Y. Gingras, C. Oppenheim, C. Hajjem, and E. R. Hilf, "The access/impact problem and the green and gold roads to open access: An update," *Serials review*, vol. 34, no. 1, pp. 36–40, 2008.

[5] M. Khabsa and C. L. Giles, "The number of scholarly documents on the public web," *PloS one*, vol. 9, no. 5, p. e93949, 2014.

[6] B. R. Martin, "The research excellence framework and the impact agenda: are we creating a Frankenstein monster?," *Research Evaluation*, vol. 20, no. 3, pp. 247–254, 2011.

[7] L. Bornmann, "Measuring the societal impact of research: Research is less and less assessed on scientific impact alone—we should aim to quantify the increasingly important contributions of science to society," *EMBO reports*, vol. 13, no. 8, pp. 673–676, 2012.

[8] S. Winnik, D. Raptis, J. H Walker, M. Hasun, T. Speer, P.-A. Clavien, M. Komajda, J. J Bax, M. Tendera, K. Fox, F. Van de werf, C. Mundow, T. F Lüscher, F. Ruschitzka,

and C. Matter, "From abstract to impact in cardiovascular research: factors predicting publication and citation," *European Heart Journal*, vol. 33, no. 24, pp. 3034–3045, 2012.

[9] A. Ibáñez, P. Larrañaga, and C. Bielza, "Predicting citation count of bioinformatics papers within four years of publication," *Bioinformatics*, vol. 25, no. 24, pp. 3303–3309, 2009.

[10] M. Callaham, R. L. Wears, and E. Weber, "Journal prestige, publication bias, and other characteristics associated with citation of published studies in peer-reviewed journals," *JAMA*, vol. 287, no. 21, pp. 2847–2850, 2002.

[11] M. H. MacRoberts and B. R. MacRoberts, "Problems of citation analysis: A critical review," *Journal of the American Society for Information Science*, vol. 40, no. 5, p. 342, 1989.

[12] P. O. Seglen, "Why the impact factor of journals should not be used for evaluating research," *BMJ: British Medical Journal*, vol. 314, no. 7079, p. 498, 1997.

[13] P. O. Seglen, "The skewness of science," *Journal of the American Society for Information Science*, vol. 43, no. 9, p. 628, 1992.

[14] R. Smith, "Measuring the social impact of research: Difficult but necessary," *BMJ: British Medical Journal*, vol. 323, no. 7312, p. 528, 2001.

[15] L. Bornmann, R. Haunschild, and W. Marx, "Policy documents as sources for measuring societal impact: How often is climate change research mentioned in policy-related documents?," *Scientometrics*, vol. 109, no. 3, pp. 1477–1495, 2016.

[16] L. Bornmann, "What is societal impact of research and how can it be assessed? a literature survey," *Journal of the Association for Information Science and Technology*, vol. 64, no. 2, pp. 217–233, 2013.

[17] H. Piwowar, "Altmetrics: Value all research products," *Nature*, vol. 493, no. 7431, p. 159, 2013.

[18] I. Viney, "Altmetrics: Research council responds," *Nature*, vol. 494, no. 7436, p. 176, 2013.

[19] J. B. Holbrook and R. Frodeman, "Peer review and the *ex ante* assessment of societal impacts," *Research Evaluation*, vol. 20, no. 3, pp. 239–246, 2011.

[20] C. S. Wagner and L. Leydesdorff, "An integrated impact indicator: A new definition of impactwith policy relevance," *Research evaluation*, vol. 21, no. 3, pp. 183–188, 2012.

[21] A.-W. Harzing and R. Van Der Wal, "A Google Scholar h-index for journals: An alternative metric to measure journal impact in economics and business," *Journal of the Association for Information Science and Technology*, vol. 60, no. 1, pp. 41–46, 2009.

[22] C.-T. Zhang, "The e-index, complementing the h-index for excess citations," *PLoS One*, vol. 4, no. 5, p. e5429, 2009.

[23] M. Thelwall, S. Haustein, V. Larivière, and C. R. Sugimoto, "Do altmetrics work? Twitter and ten other social web services," *PloS One*, vol. 8, no. 5, p. e64841, 2013.

[24] A. J. Salter and B. R. Martin, "The economic benefits of publicly funded basic research: a critical review," *Research Policy*, vol. 30, no. 3, pp. 509–532, 2001.

[25] C. Donovan, "The australian research quality framework: A live experiment in capturing the social, economic, environmental, and cultural returns of publicly funded research," *New Directions for Evaluation*, vol. 2008, no. 118, pp. 47–60, 2008.

[26] B. Van der Meulen and A. Rip, "Evaluation of societal quality of public sector research in the Netherlands," *Research Evaluation*, vol. 9, no. 1, pp. 11–25, 2000.

[27] S. P. Mostert, S. P. Ellenbroek, I. Meijer, G. Van Ark, and E. C. Klasen, "Societal output and use of research performed by health research groups," *Health Research Policy and Systems*, vol. 8, no. 1, p. 30, 2010.

[28] N. VSNU, "en knaw (2014)," *Standard evaluation protocol 2015*, vol. 2021, 2015.

[29] L. van Drooge, P. van den Besselaar, G. Elsen, M. de Haas, J. van den Heuvel, H. Maassen van den Brink, B. van der Meulen, J. Spaapen, and R. Westenbrink, "Evaluating the societal relevance of academic research: A guide," *ERiC*, 2010.

[30] J. Spaapen, H. Dijstelbloem, and F. Wamelink, "Evaluating research in context: A method for comprehensive assessment. the Hague, the Netherlands: Consultative committee of sector councils for research and development," 2007.

[31] REF Assessment Framework, *Decisions on assessing research impact Higher Education Funding Council for England.* Scottish Funding Council, Higher Education Funding Council for Wales, Department for Employment and Learning, Northern Ireland, 2011.

[32] J. Grant, P.-B. Brutscher, S. E. Kirk, L. Butler, and S. Wooding, "Capturing Research Impacts: A Review of International Practice," *Rand Corporation*, 2010.

[33] Australian Research Council, *Excellence in Research for Australia.* 2011.

[34] D. Council, "A tool for assessing research quality and relevance," *Copenhagen, Denmark: Danish Council for Research Policy*, 2006.

[35] K. Lahteenmaki-Smith, K. Hyytinen, P. Kutinlahti, and J. Konttinen, "Research with an impact. Evaluation practises in public research organisations," *VTT Tiedotteita*, vol. 2336, 2006.

[36] J. B. Holbrook, "Re-assessing the science–society relation: The case of the US National Science Foundations broader impacts merit review criterion (1997—2011)," *Peer Review, Research Integrity, and the Governance of Science–Practice, Theory, and Current Discussion*, pp. 328–362, 2012.

[37] J. Mervis, "Beyond the data," *Science*, vol. 334, no. 6053, pp. 169–171, 2011.

[38] J. B. Holbrook, "The use of societal impacts considerations in grant proposal peer review: A comparison of five models," *Technology & Innovation*, vol. 12, no. 3, pp. 213–224, 2010.

[39] H. Alhoori, R. Furuta, M. Tabet, M. Samaka, and E. A. Fox, "Altmetrics for country-level research assessment," in *International Conference on Asian Digital Libraries*, pp. 59–64, Springer, 2014.

[40] H. Alhoori, S. Ray Choudhury, T. Kanan, E. Fox, R. Furuta, and C. L. Giles, "On the relationship between open access and altmetrics," in *Proceedings of the iConference*, 2015.

[41] H. Alhoori, "How to identify specialized research communities related to a researcher's changing interests," in *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, pp. 239–240, ACM, 2016.

[42] H. Alhoori and R. Furuta, "Recommendation of scholarly venues based on dynamic user interests," *Journal of Informetrics*, vol. 11, no. 2, pp. 553–563, 2017.

[43] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[44] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3133–3181, 2014.

[45] F. Rahutomo, T. Kitasuka, and M. Aritsugi, "Semantic cosine similarity," in *The 7th International Student Conference on Advanced Science and Technology ICAST*, 2012.

[46] J. Edler and L. Georghiou, "Public procurement and innovation–Resurrecting the demand side," *Research Policy*, vol. 36, no. 7, pp. 949–963, 2007.

[47] R. Freeman and J. Maybin, "Documents, practices and policy," *Evidence & Policy: A Journal of Research, Debate and Practice*, vol. 7, no. 2, pp. 155–170, 2011.

[48] N. Black, "Evidence based policy: proceed with care," *BMJ: British Medical Journal*, vol. 323, no. 7307, p. 275, 2001.

[49] D. von Winterfeldt, "Bridging the gap between science and decision making," *Proceedings of the National Academy of Sciences*, vol. 110, no. 3, pp. 14055–14061, 2013.

[50] C. Bailey, B. Kale, J. Walker, H. V. Siravuri, H. Alhoori, and M. E. Papka, "Exploring features for predicting policy citations," in *Proceedings of the 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pp. 1–2, IEEE, 2017.

[51] B. Kale, H. V. Siravuri, H. Alhoori, and M. E. Papka, "Predicting research that will be cited in policy documents," in *Proceedings of the 2017 ACM on Web Science Conference*, pp. 389–390, ACM, 2017.

[52] L. Breimanm, J. Friedman, R. Olshen, and C. Stone, "Classification and Regression Trees Wadsworth," *Belmonm Inc., California*, 1998.

[53] A. A. Anderson, D. Brossard, and D. A. Scheufele, "The changing information environment for nanotechnology: Online audiences and content," *Journal of Nanoparticle Research*, vol. 12, no. 4, pp. 1083–1094, 2010.

[54] Y. Ding, E. K. Jacob, Z. Zhang, S. Foo, E. Yan, N. L. George, and L. Guo, "Perspectives on social tagging," *Journal of the Association for Information Science and Technology*, vol. 60, no. 12, pp. 2388–2401, 2009.

[55] S. Fausto, F. A. Machado, L. F. J. Bento, A. Iamarino, T. R. Nahas, and D. S. Munger, "Research blogging: Indexing and registering the change in Science 2.0," *PloS One*, vol. 7, no. 12, p. e50109, 2012.

[56] L. Bornmann and W. Marx, "How should the societal impact of research be generated and measured? A proposal for a simple and practicable approach to allow interdisciplinary comparisons," *Scientometrics*, vol. 98, no. 1, pp. 211–219, 2014.

[57] L. Bornmann, "Validity of altmetrics data for measuring societal impact: A study using data from altmetric and f1000prime," *Journal of Infometrics*, vol. 8, no. 4, pp. 935–950, 2014.

[58] P. Cress, "Using altmetrics and social media to supplement impact factor: Maximizing your article's academic and societal impact," *Aesthetic surgery journal*, vol. 34, pp. 1123–1126, 2014.

[59] L. Graves and T. Glaisyer, "The fact-checking universe in spring 2012: An overview (New America Foundation Media Policy Initiative Research Paper)," *New America*, 2012.

[60] H. V. Siravuri and H. Alhoori, "What makes a research article newsworthy?," *Proceedings of the Association for Information Science and Technology*, vol. 54, no. 1, pp. 802–803, 2017.

[61] The Royal Society, *The Public Understanding of Science*. 1985.

[62] S. Lee and W.-M. Roth, "Science and the good citizen: Community-based scientific literacy," *Science, Technology, & Human Values*, vol. 28, no. 3, pp. 403–424, 2003.

[63] J. D. Miller, "Scientific literacy: A conceptual and empirical review," *Daedalus*, pp. 29–48, 1983.

[64] M. K. McGinn and W.-M. Roth, "Preparing students for competent scientific practice: Implications of recent research in science and technology studies," *Educational Researcher*, vol. 28, no. 3, pp. 14–24, 1999.

[65] S. Miller, "Public understanding of science at the crossroads," *Public Understanding of Science*, vol. 10, no. 1, pp. 115–120, 2001.

[66] S. Štajner, R. Evans, C. Orasan, and R. Mitkov, "What can readability measures really tell us about text complexity," in *Proceedings of workshop on natural language processing for improving textual accessibility*, pp. 14–22, 2012.

[67] R. G. Benjamin and P. J. Schwanenflugel, "Text complexity and oral reading prosody in young readers," *Reading Research Quarterly*, vol. 45, no. 4, pp. 388–404, 2010.

[68] H. V. Siravuri, A. P. Akella, C. Bailey, and H. Alhoori, "Using social media and scholarly text to predict public understanding of science," in *Proceedings of the 2018 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, IEEE, 2018 (in press).

[69] A. Madylova and S. G. Oguducu, "A taxonomy based semantic similarity of documents using the cosine measure," *24th International Symposium on Computer and Information Sciences, ISCIS 2009.*, pp. 129–134, 2009.

[70] K. C. Shet, U. D. Acharya, and S. K. Manjula, "A new similarity measure for taxonomy based on edge counting," *CoRR*, vol. abs/1211.4709, 2012.

[71] G. A. Miller, "Wordnet: A lexical database for English," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[72] R. Flesch, "A new readability yardstick.," *Journal of Applied psychology*, vol. 32, no. 3, p. 221, 1948.

[73] C. U. Yule, *The statistical study of literary vocabulary.* Cambridge University Press, 2014.

[74] A. Miranda-García and J. Calle-Martín, "Yule's characteristic k revisited," *Language Resources and Evaluation*, vol. 39, no. 4, pp. 287–294, 2005.