X Modal
X Cultural
X Lingual
X Domain
X Site
Global OER Network

Grant Agreement Number: 761758
Project Acronym: X5GON
Project title: X5gon: Cross Modal, Cross Cultural, Cross Lingual, Cross Domain, and Cross Site Global OER Network
Project Date: 2017-09-01 to 2020-08-31
Project Duration: 36 months
Document Title: Quality assurance models
Author(s): Sahan Bulathwela, Emine Yilmaz, John Shawe-Taylor
Contributing partners: UCL, JSI, UPV, K4A
Date: 31/08/2018
Approved by:
Type: Report
Status: Final
Contact: m.bulathwela@ucl.ac.uk, emine.yilmaz@ucl.ac.uk, j.shawe-taylor@ucl.ac.uk

Dissemination Level		
PU	Public	✓
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Revision

Date	Lead Author(s)	Comments
08/08/2018	Sahan Bulathwela	Initial Draft
10/08/2018	Sahan Bulathwela, Erik Novak	Additional content about JSI contributions
20/08/2018	Sahan Bulathwela, Davor Orlic	Additional content about K4A contributions
20/08/2018	Sahan Bulathwela, Emine Yilmaz	Updates from the feedback from Emine Yilmaz
28/08/2018	Sahan Bulathwela, Erik Novak	Updates from the feedback from Erik Novak
29/08/2018	Sahan Bulathwela	Adding the list of figures, list of tables, abstract and appendix
31/08/208	Sahan Bulathwela	Final Version

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

Abbreviation	Meaning
OER	*Open Educational Resource*
VLN	*videolectures.net*
SVM	*Support Vector Machine*
MTL	*Multi-task Learning*
NLP	*Natural Language Processing*
IJS	*Institut "Jožef Stefan"*
JSI	*Institut "Jožef Stefan"*
API	*Application Programming Interface*
RMTL	*An R Library for Multi-task Learning*
RR	*Ridge Regression*
RMSE	*Root Mean Squared Error*

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

X5GON (Cross Modal, Cross Cultural, Cross Lingual, Cross Domain, and Cross Site Global OER Network) is a Horizon 2020 collaborative project aiming at providing the next generation network for learning with Open Educational Resources. The goal of this project is to leverage Information Technology and Artificial Intelligence to develop a learning resource network that can provide high quality, personalised learning pathways to learners by recommending open educational resources from multiple repositories.

This is the first deliverable of the work package relating to learning rich content representations (WP1) reporting on the development of quality models to assess the content quality of OERs automatically at scale. The study started by doing a thorough literature survey to identify the factors indicating quality in educational content. Based on the findings, a collection of features and labels were created from the data at hand. A series of regression and classification-based models were developed to automatically assess quality of educational content.

Based on the final comparison of results, the SVM model that uses pairwise comparison technique performed best on the test data with 71% classification accuracy.

1. INTRODUCTION

X5gon aims to build a global open education resources network that can automatically index digital educational material from multiple open resource repositories and deliver this content to informal learners in the form of recommendations to use these resources and significantly enhance the learning experience and effectiveness of the learner.

In the context of this project, work package 1 (WP1) led by University College London (UCL) involves developing quality assessment models for educational material and deriving high quality content and user representations to recommend effective learning trajectories to learners.

In the earlier stages of the project, quality assurance models are developed to automatically identify high quality content from low quality content. Quality assurance models help in three main ways.

1. Allows automatic identification of high quality vs. low quality education resources when X5GON integrates with new repositories
2. Allows ranking lectures and comparing between them to automatically rank them based on content quality
3. In the long term, leverage personalization by capturing patterns about quality features different users prefer when consuming OERs.

Automatic assessment of content quality at scale is essential for deriving accurate representations enabling personalization of learning trajectories. This is the focus of WP1 during the first 12 months of the project. In this report, we will explore the current state, solution directions and future potential of assessing online/digital educational content for quality assessment of open education resources.

1.1. DOCUMENT OVERVIEW

In ***section 2 and 3*** we outline the background context to the report by attributing the contributions from the partners that helped the project and certain challenges and setbacks we had to overcome during the reporting period.

Section 4 summarises the findings from the literature survey and setting the stage for the choices made in deriving the potential features, labels and approaches that would be useful for developing quality assurance models.

In ***section 5***, we discuss the proposed solution outlining different pre-processing steps and model training options that are available to us and promising.

Section 6 then proceeds to describing the actual raw data and developmental tools available to use and then proceeding to carefully explaining how exactly new tools, the final features and labels were derived. ***Section 7*** explains how the final dataset is used to train the proposed models.

Section 8 summarises the key results from the model evaluation stage.

Finally, the key observations and conclusions are summarised in ***section 9***.

2. CONTRIBUTION FROM PARTNERS

In the initial phase of deriving quality models, several partners in the consortium contributed significantly towards the data and tools devised during the study. This section outlines the contributions from different partners.

Major contributions to quality model derivation during month 1-12 came from three main partners.

1. Josef Stefan Institute (JSI)
2. Knowledge 4 All foundation (K4A)
3. Polytechnic University of Valencia (UPV)

2.1. JOSEF STEFAN INSTITUTE (JSI)

JSI provides support by providing the tools and infrastructure for X5GON. The data and infrastructure provided by JSI was essential to running developing quality assurance models for X5GON. The following sections explain different tools and platforms developed and maintained by JSI that helped us to leverage data and develop quality assurance models.

2.1.1. Platform Overview

The role of the X5GON platform is to connect different OER repositories by collecting their content, enrich it using different tools and services, store the enriched metadata, analyse it and provide valuable information to the users. The platform source code is available open source[1].

What follows are descriptions of components used in the platform. This is an extension of component descriptions found in deliverable D2.1 - Requirements & Architecture Report.

2.1.2. Technology

The data enrichment process uses different tools and services to extract information that is then used in different Work Packages. What follows are brief descriptions of these services and how we use them in the platform.

2.1.2.1. Wikifier

Wikifer *[Brank et al (2017)]* is a web service which takes a text document as input and annotates it with links to relevant Wikipedia concepts. The service supports cross and multi-linguality enabling extraction and annotations in different languages. This forms as the basis for comparing and analysing OER materials written in different languages. The tool was developed by IJS.

2.1.2.2. Enrycher

Enrycher[2] is a web service which automatically enriched a provided text document with topics, keywords, named entities and other natural language enrichments. Because data extracted using Wikifier already covers most of the Enrycher's output, the only component used is called DMOZ classification which extracts topics using the DMOZ ontology[3].

[1] JozefStefanIstitute/x5gon: https://github.com/JozefStefanInstitute/x5gon
[2] Enrycher, http://enrycher.ijs.si/
[3] DMOZ - The Directory of the Web, http://dmoz-odp.org/

2.1.2.3. qtopology

To build the pre-processing pipeline we decided to use qtopology[4] which is a distributed stream processing layer written in Node.js. It enables developers to create single processing components called Bolts, data retrieval components called Spouts and organize them using schemas called Topologies. The terminology has been adopted from the Storm project[5]. This enables us to easily create new processing components and include them into the existing pre-processing pipelines.

2.1.2.4. Apache Kafka

To communicate between different components of the platform we decided to use Apache Kafka[6], a distributed streaming platform. It is used to build real-time data pipelines and streaming apps.

Within the X5GON platform, apache kafka will be used as a messaging system which will redirect messages between different components - services developed within the project, the platform and the pre-processing pipeline.

2.1.3. User Activity Tracker

To retrieve user activity data, we developed a library[7] which enables sending user activity data to the platform. The user activity tracker has been presented in D2.1 - Requirements & Architecture report as well as in D4.1 - Initial Prototype of User Modelling Architecture.

2.1.4. Infrastructure

The platform infrastructure is hosted on the Posta Slovenije cloud named PosiTa[8]. The platform runs on a machine with 150GB of space, 32GB of RAM and 8 CPUs. The operating system installed on the machine is Linux Debian 8.6 (jessie). The machine can be dynamically scaled on request.

Additionally, we can request for additional machines to run services developed within the project.

2.2. KNOWLEDGE 4 ALL FOUNDATION (K4A)

K4A contributes to WP1 through disseminating the work done to the global OER community. This was done by presenting WP1 work at the "Course in Open Education Design" organised with JSI and introducing the results of WP1 into the "Open Education for a Better World" on-line mentoring program in which students from different backgrounds and different parts of the world developed 14 OER projects aligned on the UN SDG agenda. Finally, K4A and PS have identified quality as a main market driver in their work in WP8 and D8.1 Market Analysis.

K4A also helps WP1 through leveraging data from Videolectures.Net (VLN) website jointly powered by JSI and K4A. JSI and its case study in the project at the VLN website, has a collection of some 26152 videos summing-up to about 21,259 lectures and 15609 authors. The main reason why recurrent visiting users are coming back is to watch (peer-to-peer) validated high quality courses, research and conference talks with the average age group 26-30, and 23-25, being 56,7% university students for the USA, China, India, Germany, etc. The value of the work done in WP1 is highly relevant

[4] qtopology | Distributed stream processing layer, https://qminer.github.io/qtopology/
[5] Apache Storm, http://storm.apache.org/
[6] Apache Kafka, https://kafka.apache.org/
[7] JozefStefanInstitute/x5gon, https://github.com/JozefStefanInstitute/x5gon/tree/master/src/server/platform/snippet
[8] PosiTa | Digitalne storitve Pošte Slovenije, https://www.posita.si/

VLN as automatic understanding of quality levels of the content displayed to VLN customer base generates traffic and optimises the service.

2.3. POLYTECHNIC UNIVERSITY OF VALENCIA (UPV)

Polytechnic University of Valencia (UPV) contributes towards the quality assurance models by providing video transcripts and translations into different languages. UPV develops and maintains the MLLP Transcription and Translation Platform[9] (MLLP-TTP) that ingests educational materials (video content) from X5GON and generate English transcription/ translation files.

For detailed understanding about the application of MLLP-TTP service, performance and other system related details, we refer you to Deliverable D5.1: First report on Piloting.

[9] https://ttp.mllp.upv.es.

3. DATA ACQUISITION

Since the inception of the project, a lot of things had to be setup to pave way to the quality models to come. The most important task in the initial stage is to leverage relevant data to train the machine learning models on. Most efforts in the first phase of the project was put towards granting access to relevant data from www.videolectures.net that was available to the project through the project partners. Once the required authorizations were obtained, a significant amount of time was spent on downloading raw data, understanding the data, cleaning the data and compiling it into a usable dataset to be used to develop machine learning models for quality assessment. Data will be described in detail in section 6.1.

As this is a brand-new dataset that has not been used by the academic community before, rigorous measures were taken to sanity check every step of data processing with no prior assumptions about the correctness of data.

It is also fair to draw attention to the fact that www.videolectures.net was the primary and only source of data available in this stage. Careful evaluations should be done once new data sources are available to ensure the generalizability of the developed models to any educational resource.

4. LITERATURE REVIEW ON QUALITY ASSURANCE

From the initial analysis we could see that there is very little work done on automatic quality assurance in education domain. The main observation during the initial literature survey is that no one has done significant work in formally defining what quality would mean in the context of educational material. Due to this reason, multiple different sectors were studied to understand the definition of quality of content. There are multiple segments in research community looking at content quality assurance for different domains.

4.1. CHALLENGES

There are several challenges in quality assurance in education sector. Quality of educational material itself is far from being identified as a simple and straightforward concept. It is quite subjective and covers multiple dimensions. Due to the uncertainty around the definition of better-quality educational material, there has been very little research currently available and the datasets are very hard to come by.

One way to go about assuring quality of online content is to get a panel of experts to annotate and approve each educational material. There are instances in healthcare forums where medical practitioners would manually annotate and certify the quality of healthcare information in health forums *[Boyer (2017)]*. Doing this for educational material is time consuming. It also carries a large opportunity cost as teachers and domain experts should be doing these annotations. These skilled persons can utilize that time to do far more effective things. Also, this is not scalable as there the volume of available open educational resources is quite large.

4.2. QUALITY ASSURANCE IN DIFFERENT WEB DOMAINS

From the initial literature survey, we came across three main domains where research into quality assessment has been carried out.

Education: Not a lot of work on quality. But some work has been done around modelling knowledge learning and information retrieval for learning.

Healthcare: Significant work in applying machine learning to ensure trust-ability of healthcare information.

Information retrieval: Assessing quality of information searched as it is a strong factor affecting user satisfaction.

4.2.1. Education

Computer aided learning systems have shown to improve learning experience outside formal classroom learning environment leveraging personalised learning instructions, modern and up-to-date material and self-paced learning *[Pirolli & Kairam (2013)]*. Majority of recent work done in this space relates to improving quality of search results when doing an information search in a learning environment.

Collins-Thompson, Chica and Sontag has shown that incorporating reading level related features in user and document vectors can help towards improving the relevance of documents retrieved for a user [*Collins-Thompson et. al (2011)*]. There have also been several efforts in literature to identify a sensible metric to represent language level. There are several metrics such as Fletch Kuncaid Score, FOG, SMOG and they are widely adapted *[Si L., and Callan, J. (2001)].*

Amongst Intelligent Teaching Systems, it is common to use the Bayesian Tracing Model *[Yudelson, et al (2013)*] to represent the learning progress of a student. In this model, the knowledge a learner acquires is represented using a multinomial distribution of learning aspects about a field. Predicting learner's mastery of a field is done by observing how a learner applies learned skills to solve a problem (test). This approach looks like a promising way to represent both the contents of documents and the knowledge state of a learner.

Using this idea, Syed and Collins-Thompson *[Syed & Collins-Thompson (2017)]* has been successful towards improving information search results for learning tasks. In a vocabulary scenario, they frame mastery of a word by assigning the minimum number of times a learner needs to read a word, therefore each document contributing differently towards mastering different words.

We believe that it is possible to generalise this idea by deriving a set of concepts and representing educational resources in terms of these concepts. Coming up with a global taxonomy of concepts that cover all aspects of knowledge could be difficult. Wikification is an emerging method that tries to annotate free text with concept pages found in Wikipedia. Recent research in Wikification has led to numerous methods being invented in enriching natural text with global concepts *[Hoffart et al. (2011)*]. Josef Stefan Institute has developed and published www.wikifier.org *[Brank et al (2017)]* API which is a great tool to enrich natural text with Wikipedia concepts they belong to. It uses the substrings in a given input document to build a graph of different Wikipedia concepts mentioned in the document. Then it uses the PageRank algorithm *[Brin & Page (1998)]* to come up with the most influential Wikipedia concepts connected to the input document.

4.2.2. Healthcare

The main research interest in healthcare and information quality assurance revolves around the trust ability of health-related information posted in online forums. When people consume healthcare advice and opinions in the Internet, it is vital that this information is reliable and accurate. During our initial survey we came across studies that attempted to use machine learning to automatically detect quality in healthcare related content *[Sondhi et al (2012)]*.

Due to importance of trust ability, ideas such as HONcode Principals *[Boyer et al (2017)]* introduced by Health on Net (HON) Foundation has emerged. HONcode criteria is an acknowledged indicator of health information accuracy and reliability of information sources (websites) and are widely accepted by experts in medical community worldwide *[Gaudinat et al (2007)]*. There are also other organizations such as Quackwatch[10] who are also interested in controlling quality of healthcare related content in the Internet.

There are 7 main attributes considered in healthcare sector:

4.2.2.1. Authority

The reputation and quality of the authors is very important in healthcare forums. The qualifications of the authors and their affiliations give a huge weight towards the

[10] http://www.quackwatch.com/

reliability of information. Having the author and affiliation information explicitly mentioned in content is a strong indicator of quality.

4.2.2.2. Complementarity

In a health forum, information should be provided in a way that it won't compromise or damage the patients' relationship with their personal doctor. Information that supports and complements the patient's current knowledge is encouraged instead of replacing their personal doctor-patient relationship. Already, there is work where a Naïve Bayes Classifier is used to detect text demonstrating complementarity *[Boyer & Dolamic (2014)]*. This classifier is trained with unigram term frequency features based on human expert labelled text extracts.

4.2.2.3. Attribution to References

A very strong trait of an argument is to clearly cite references and evidence that supports the claims. Citing the sources that supports a piece of information usually indicate good quality.

4.2.2.4. Freshness of information

Validity of information may decay over time. This is true with healthcare information and educational content as well. As new knowledge is produced, older knowledge becomes outdated. In the context of healthcare forums, having the publication date mentioned is considered a good feature of quality content. We also found machine learning being used to automate detecting publication dates in healthcare forums *[Boyer et al (2017)]*. They used a set of expert labelled publication date extracts from websites to train the Stanford Named Entity Recognition (NER) model *[Finkel et al (2005)]*.

4.2.2.5. Policy Influence

In healthcare forums, attributes such as their privacy policy, advertising policy and financial disclosures are quite important. The privacy policy outlines how transparent an entity is when data is collected and managed. The Advertising policy on health forums give indications of how financially motivated a content website is. Commercial focus of a health website can heavily bias the type of information disseminated through it. Financial disclosures relate to how the studies are funded. This degree of transparency improves the trust ability of information.

4.2.2.6. Link Structure

Websites inherently link to other webpages in the Internet. Specific patterns in the link structure can also indicate commercial focus of healthcare websites. For instance, a more reliable healthcare website will have a lot of internal links and external links that point to neutral, reputed information sources rather than to pharmaceutical merchants *[Brin & Page (1998)]*. Raw features such as

- Number of internal links
- Number of external links
- Total number of links
- Presence of links to contacts, privacy policy & etc...

Can be used to capture the link structure of a document.

4.2.2.7. Presentation Features

The nature of presenting information can also be an important feature when deciding on the quality of an information source. Features such as percentage of coherent text indicates if text is scattered around the webpage with advertising spaces in between this text. Sondhi et al *[Sondhi et al (2012)]* uses tools such as ELinks[11] to extract the textual representation of an html page to extract presentation related features.

4.2.3. Information retrieval

Education in certain aspects can be viewed as a unique case of information retrieval. In contrast to information search, which addresses facilitating relevant information to a user based on different types of information requirements (Eg: Navigational, Informational and Resource) *[Rose & Levison (2004)]*, education focusses on delivering relevant information that is useful to a person in long-term growth. Therefore, it is sensible to consider that learnings from information search domain is applicable and relevant to delivering educational content as well.

In terms of assessing quality of content, it is evident that some factors mentioned in the above sections such as difficulty level of language *[Yilmaz et al (2014), Collins-Thompson et al (2011)]*, link structure *[Brin & Page (1998)]* etc. But when investigating quality-based information retrieval specifically, we observed that textual quality features that go beyond readability are used *[Bendersky et al (2011)].*

4.2.3.1. Linguistic style

Apart from the level of language, the style of language used can also affect the information delivery. We came across studies that uses features such as the intersection between English stop words and vocabulary in web pages to detect spam web pages *[Ntoulas et al (2006)]*. These features can be used to represent the style of language used in different educational resources.

4.2.3.2. Document Entropy

Document entropy can be used to quantify the "focus" of a document. *[Bendersky et al (2011)]* uses document entropy as a feature in modelling quality biased information search. Lower the entropy value in document, the more focussed that document is.

4.3. QUALITY LABELS

Training machine learning models to improve engagement, user satisfaction, user retention in web services has been quite popular in the recent past. Due to this reason, there are numerous publications that discuss about this emphasised in the earlier sections, there is no formal definition for quality of content in the current research landscape. We came across several studies that used different user satisfaction indicators as target variables for high quality content. These signals fall under two main categories:

1. **Explicit feedback:** data captured through the system explicitly to understand user satisfaction.
2. **Implicit Feedback:** data captured as part of fulfilling a different function, but also strongly indicative of user satisfaction/intent.

[11] http://elinks.or.cz

Some examples are:

1. Explicit feedback:
    a. Star ratings
    b. Likes / comments
2. Implicit feedback
    a. Click through data
    b. User dwell time / watch time

4.3.1. Explicit Feedback

As seen from examples, explicit user feedback is very fine-grained and effective for training supervised machine learning models. However, users do not come to websites to spend significant amount of time rating content. Asking for excessive explicit feedback on content hinders user experience. Due to this reason, explicit feedback is usually rare in terms of data points.

From the industry, we can recall the Netflix price competition *[Amatriain (2009, 2012)]* where the recommendation problem was solely based on star ratings by users. A lot of personalised recommendations system use a for or rating as training labels. But the main drawback of using explicit ratings remains, they are very scarce.

Likes and comments are also very resourceful in mining user opinion, Like, comment data also are extremely scarce. Another disadvantage of this type of data is that the users can be motivated by numerous reasons to like or comment on a piece of web content. Often these actions are motivated by reasons different from the factors we want to measure (such as personal biases and satisfaction / dissatisfaction of irrelevant features)

4.3.2. Implicit Feedback

Clickthrough data is also a very useful source of user engagement and preference. There have been numerous studies that has used click data as a representation of user preference. In information search domain, click through is used frequently to measure effectiveness of search results *[Serdyukov et al (2014)]*. Additionally, there has been extensive studies showing that click through signals captures relative relevance of search engines although they carry a small bias when representing absolute relevance of search results *[Joachims et al (2017)]*.

Most recent work has shown that engagement related signals tend to be useful as target variables in machine learning models. This is mainly because, such features capture user satisfaction and retention. Even business organizations such as Youtube use user engagement signals such as view time to train machine learning models that recommend their users with new content [*Covington et al (2016), Meyerson (2012)*]. *Meyerson* explains how using watch time can improve predicting engagement of users as opposed to metrics such as number of views that can be easily contaminated by attractive titles, thumbnails etc... (click baits).

4.4. Discussion

One of the main patterns that we have observed through this literature survey is that no one wants to explicitly specify neither a family of features nor target variables that represent quality. Different applications of content assessment tend to use subset of content, author and user related attributes to capture quality. Based on the findings above, we think overall quality of a document can be categorised into 5 main verticals.

- **Coverage of a topic:** To what extent the topics in knowledge area are covered. The knowledge areas, the focus or generic nature of documents fall under this vertical.
- **Authority:** This aspect represents the reputation and credibility of the authors of content. Any information relating to author or the reputation of their affiliations.
- **Understandability (easiness to understand):** The readability of the content including the level of language used.
- **Presentation quality:** attributes related to the presentation of materials such as whitespace, pauses, disconnect of knowledge etc...
- **Freshness:** how recent/ up-to-date the resource is

We believe that training a model using above features is the most effective way to develop automatic, scalable quality assurance models.

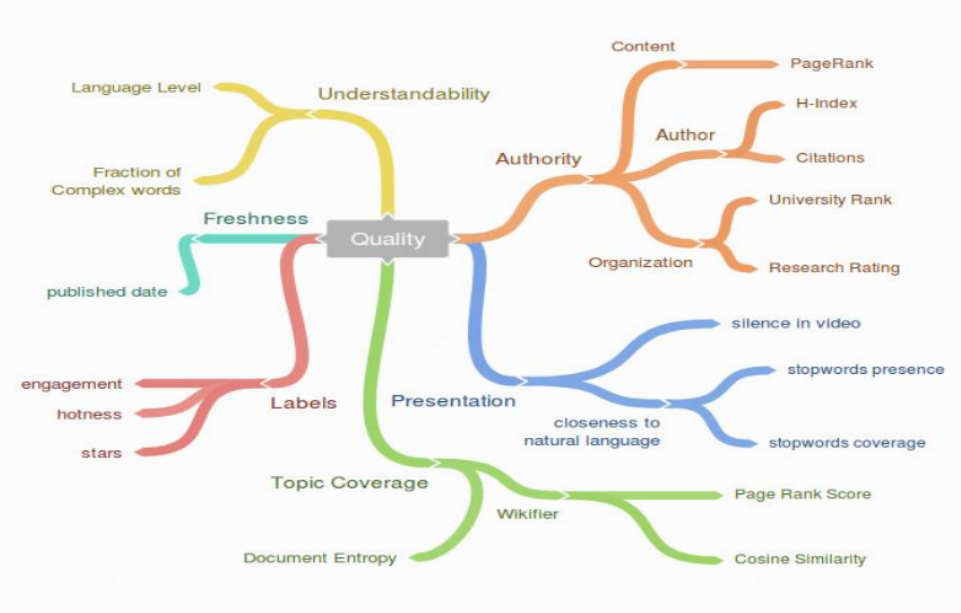

Figure 1: *Summary of potential features and labels indicative of quality*

From the literature survey, we realised that measuring quality is quite the difficult task. Research community uses different observable variables such as star ratings, engagement, number of views etc. to capture quality.

When comparing numerous observable signals, we can see from Figure 1 that there are numerous signals that are used to measure user satisfaction and acceptance. Explicit feedback signals such as user ratings, comments and likes can be considered as few of the most straightforward type of feedback available. However, explicit feedback is quite rare in datasets although they carry stronger signals. In contrast, user activity related data such as click through rate, engagement rate are more widely and densely available. But these implicit feedback mechanisms tend to be weaker in terms of signal compared to explicit feedback. However, there is plenty of evidence in literature that suggest its usefulness as discussed in the earlier section.

5. PROPOSED METHOD

5.1. DATA

The main dataset available to us at this point is from www.videolectures.net. The educational materials served through this domain are mainly videos of conference talks and lectures. This dataset provides information about the lecture such as the subject, authors and their affiliations. It also provides access to the English transcription of the video. In addition to this, access to anonymous user sessions are available for us to extract engagement signals from lecture views. A detailed explanation of the raw data and the features will be provided in "Data and tools" section.

5.1.1. Potential features

Based on Figure 1 in the background section, we can conclude that there are 5 main drivers of quality of information. We use these quality verticals to extract features that are indicative of these verticals. In the "Data and tools" section, we explain in detail what exact features are extracted and used in the models.

In a nutshell, we can use textual representation of educational resources to extract features such as level of language, topic coverage, style of language, length of content, entropy of the document etc...

5.1.2. Potential Labels

Three potential label variables are found in the videolectures.net dataset. Namely, they are:

1. "Hotness" score per lecture
2. Average Star rating per lecture
3. Lecture view related data

5.2. METHODOLOGY

Based on the background literature survey to the field, it is observable that most of the studies treat the problem as a supervised learning problem. From the previous section, we can see that there have been several attempts to use Naïve Bayes Classifier, Support Vector Machines, Feedforward Neural Networks and various other supervised learning algorithms to solve the problem at hand. We think that is a great starting point for this work. Based on the data at hand, the main objective is to build a supervised machine learning model that understand superior quality content based on features extracted from the educational resources.

As our application deals with ill-defined, yet sensitive, topic of quality assessment of educational content, it is ideal to derive a machine learning model that is accurate but, also, highly interpretable at the same time. Due to this reason, it is suitable to initially build models that are easily interpretable. This also gives us the opportunity to observe and understand more about the problem at hand based on what the story that the data tells.

5.2.1. Quality by Subject

Quality of educational material also change amongst different subject areas. The expectation of linguistic style, level of language and many other aspects change significantly between subjects. For example, the composition of a Computer Science

material is very different from one of Chemistry or Philosophy. This variation mainly roots to pedagogical techniques these different disciplines have developed to transfer knowledge effectively. We need to account for this variability when deciding on the experimental setup for training the models.

Multitask Learning is a popular machine learning technique used in settings like the subject level variability *[Evgeniou and Pontil (2004)]* . The main idea of multi task learning is learn a set of models to fulfil multiple tasks simultaneously. By doing this, individual task learners can learn models that fit to their own task while sharing some information with other task models. There are several ways information can be shared depending different regularisations such as joint feature selection *[Lui, Ji and Ye (2009)]*, trace norm learning *[Fang et al (2017)]*, etc. Multi task learning allows different models to share information about patterns that help all the models while having freedom to learn task specific patterns. This setting fits very well for learning quality assessment for different subject areas (e.g. Computer Science, Biology, etc.). If we treat different subject areas as different tasks, multitask learning allows the models to learn subject specific quality predictors. But this setting allows them to share information about patterns that govern universal quality of educational content in general.

5.2.2. Pairwise preference for quality

The main objective of quality assessment of educational content is to be able to distinguish bad educational content from good ones. This problem can also be treated as a ranking problem. The idea was first coined by Thurstone *[Thurstone (1927, 1929, 1959)]* in his work around measuring intangible variables such as preference, attitude, emotion in psychology. This approach has been used by many scientists to interpret variables such as importance of a decision *[Saaty (2008)]*, personal skill level *[Elo (2008), Herbrich et al (2006)].*

Pairwise preference *[Herbrich et al (1998)]* has emerged recently being used in "Learning to Rank" problems. In this approach, we model the ranking problem by comparing pairs of observations than ranking the whole set of examples into a global order. The focus shifts to teaching the model to identify what factors lead to superiority/ inferiority of items. There are several studies that has used pairwise preference to rank items in information retrieval domain with boosting algorithms *[Freund et al (2003)]* and neural networks *[Burges et al (2005)]*.

As quality of educational material is highly intangible, pairwise preference approach is very suitable for this problem. Instead of trying to have an exact measurement that summarises overall quality, relative preference of users can be used to rank lectures based on absolute quality.

5.3. MODELS

We consider three main machine learning models that are ideal for this problem setting.

5.3.1. Ridge Regression

Ridge regression is one of the most popular techniques used in Machine Learning. As shown in (1) this model is a supervised learning technique that takes **p** number of features **(x_i … x_p)** in Real space ($X \in \mathbb{R}_p$) as input to predict a target variable **($\hat{y}$)** which

is in Real value space ($Y \in \mathbb{R}$) as well. It does this by learning a linear vector **w** where $\mathbf{w_0}$ is the intercept and $\mathbf{w_1} \dots \mathbf{w_p}$ are the weight coefficients for the features.

$$\hat{y}(w, x) = w_0 + w_1 x_1 + ... + w_p x_p \tag{1}$$

Deferring from Ordinary Least Square regression, it is known as "Ridge" regression as it enforces l2-regularization [*Ng (2004)*] to control the complexity of the learned linear function w.

$$\min_w ||Xw - y||_2^2 + \alpha ||w||_2^2 \tag{2}$$

As shown by (2), Ridge regression employees a hyperparameter **α** that controls model complexity. The second term controlled by **α** penalises the model when the weight coefficients in vector **w** gets bigger.

5.3.1.1. Advantages of Ridge Regression

This model highly suitable for ranking problems when the target variable is a real value ($Y \in \mathbb{R}$). Ordinary least square regression also has a closed form solution that allows finding the most suitable weight coefficients efficiently. The l2-regularization also helps us to have control over the generalization error. As the model determines a linear set of weight coefficients for each feature in the dataset, the model is also highly interpretable.

5.3.1.2. Disadvantages of Ridge Regression

The model is extremely simple. This will limit the range of models that we can fit to this data. Therefore, there is a chance that the model is not complex enough to capture non-linear patterns in the data. The regularization also limits the solution space the algorithm can search in.

5.3.2. Pairwise Preference Classification

Classification case is quite different from the regression setting that we explained in the section above. Instead of trying to predict a real value as target variable, we attempt to predict a target value in a discrete state space. In other words, it is a supervised learning technique that takes **p** number of features $\mathbf{(x_i \dots x_p)}$ in Real space ($X \in \mathbb{R}_p$) as input to predict a target variable **(ŷ)** which is in discrete state space ($Y \in \{-1, +1\}$) in binary class case, and $Y \in \{1, 2, \dots, k-1, k\}$ in multiclass classification with **k** classes).

We use Support Vector Machines (SVM) *[Cristianini & Shawe-Taylor (2000)]* to solve the problem at hand. SVM algorithm solves the classification problem by learning **b**, the intercept, and the linear vector **w** where $\mathbf{w_1} \dots \mathbf{w_p}$ are the weight coefficients for the features. SVM algorithm tries to maximize the distance between the decision boundary and the training examples of the classes using the optimization outlined in (3).

$$\begin{aligned} \min_{w,b,\zeta} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \zeta_i \\ \text{subject to } y_i (w^T \phi(x_i) + b) \geq 1 - \zeta_i, \\ \zeta_i \geq 0, i = 1, ..., n \end{aligned} \tag{3}$$

Conventionally, classification setting is not designed to solve ranking problems, although it has been adapted for ranking by several researchers using methods such as Ranking SVM *[Joachims (2002)]* and Ordinal Regression *[Herbrich et al (1999)]*. We use a similar method as Ranking SVM *[Joachim, (2002)]* to model rank lectures in terms of quality. We convert our dataset into a pairwise preference dataset to solve (4) where **f** represents the differences between Lecture 1 and 2 while **y** represents the superior lecture between lecture 1 and 2. This way, we can solve for **b**, the intercept, and the linear vector **w** where $\mathbf{w_1}$ **…** $\mathbf{w_p}$ are the weight coefficients for the features to derive a highly interpretable model that .

$$y = w^T f(x_{lecture1} - x_{lecture2}) + b$$

$$where \begin{cases} y = +1 \text{ if } lecture\ 1 > lecture\ 2 \\ y = -1\ otherwise \end{cases} \quad (4)$$

We will discuss in detail about the exact features and labels used in the "Data and Tools" section.

5.3.2.1. Advantages of Pairwise Classification

As the model tries to capture how pairwise preference between lectures occur, the dataset will capture pairwise comparisons between lectures rather than a global ranking score. Therefore, we create more examples where individual examples are more informative.

The l2-regularization also helps us to have control over the generalization error. As the model determines a linear set of weight coefficients for each feature in the dataset, the model is also highly interpretable.

One might think that creating a global order of lectures also becomes non-trivial in the pairwise setting. This is because the pairwise preference ranks may not necessarily propose a unique ranking in an unequivocal way. In the context of pairwise classification, several methods have been suggested and empirically evaluated for this task *[Fürnkranz & Hüllermeier (2010)]*. Therefore, there are multiple approaches available to solve the global ranking problem.

5.3.2.2. Disadvantages of Pairwise Classification

We can run to similar limitations on model complexity as discussed in Ridge Regression.

But the main disadvantages of this model lie at the computational complexity of training the model. As pairwise comparisons exponentially increase the size of the training data, the number of training examples increase from N to N^2. As the computational complexity of Primal SVM training lies around $O(num\_examples^2)$. The overall training complexity is around $O(N^4)$ for our case.

5.3.3. Multitask Learning

"**Multi-task learning** (MTL) is a subfield of machine learning in which multiple learning tasks are solved at the same time, while exploiting commonalities and differences across tasks."

[Wikipedia, (May 2018)]

MTL is usually enforced with regularization induced by requiring the weights to comply with a pattern that represent relationships and differences between tasks is more superior than regularization that prevents overfitting. This works well when tasks share a significant number of commonalities.

There are multiple algorithms that use multi-task learning ideas. By regularizing the weight matrix **w** in a way that it either selects a subset of features for all tasks *[Liu et al (2009)]* or learn a combination of highly correlated task parameters *[Maurer & Pontil (2013)]*, information can be shared between tasks.

5.3.3.1. Advantages of Multi-task learning

Quality of educational resources is governed by certain patterns that indicate general quality. However, some aspects of quality change between subject fields (Science vs. Arts vs. Business). For example, having pauses in a lecture contributes differently towards overall quality of a philosophy lecture in comparison to a computer science lecture. Multi-task learning is a great way to exploit this structure.

There are also techniques to incorporate information about Network/Graph *[Widmer et al (2012)]* or Clustering *[Zhou et al (2011)]* structure to the multi-task learning setting. Given that some subjects tend to be related to other subjects, being able to represent the clustering/ graphic structure of the data can be useful.

5.3.3.2. Disadvantages of Multi-task learning

Heavily relies on regularization techniques that can limit the hypothesis space. This can lead to underperforming model as there is a risk of too much information being suppressed. Also, the model unnecessarily complex if the assumptions about the task structure are wrong.

5.4. Discussion

From the method exploration in the above section, we can see that there are various advantages and disadvantages of the proposed methods. Linear Regression has the advantage of training a computationally efficient model while preserving interpretability. However, the simplicity of the model may mean that it wouldn't be the most effective algorithm to capture the patterns expressed by data. On the other hand, Pairwise Preference Classification generates a far more expressive dataset where each combination of items is compared with each other. But this method has the disadvantage of computational complexity due to the exponential increase of training examples. However, our case is quite like RankNet, where the number of examples would not increase to N^2 scale as the pairwise comparisons are only done between documents belonging to a certain query *[Burges et al (2005)]*. Our case will only compare lectures within a subject area (Biology, Computer Science, Arts etc...). Within pairwise preference setting, both linear classification and /or multitask learning may yield promising results.

Due to these reasons, it is sensible to apply both ridge regression and pairwise classification to our data and investigate what model will perform better. We conclude that the best approach would be to try linear regression, pairwise classification and multitask classification on our dataset and evaluate which model yields best results.

6. DATA AND TOOLS

In this section we discuss about the tools and data that were available to us for deriving the final dataset used for the analyses. We first describe the raw data that was available to use from our partners. Then we proceed to tools that were already available from the scientific community and our partners. Then we follow-up with the new tools that were developed to enrich the raw data that was already available to us.

6.1. VIDEOLECTURES.NET DATA

The main source of data available to us during the initial period of the project is from www.videolectures.net. Videolectures (VLN) is a website run by Josef Stefan Institute (JSI) and Knowledge 4 All (K4A) foundation. According to its Wikipedia page, it is one of the largest online academic video repositories in the world *[Wikipedia, (July 2018)]*. Videolectures.net also releases most of its content in non-restrictive Creative Commons licence (Creative Commons Attribution-Non-commercial-No Derivative Works 3.0) which makes it very simple to use their data. VLN repository is also the main data source for "Translectures" project *[Knowledge 4 All (2018)]* which focusses on building technologies to do cross-lingual machine translation at scale.

6.1.1. Variety of data

As mentioned above, Videolectures data revolves around videos of lectures, presentations, conference talks given by research community at various research events around the world. This repository is concentrating on hosting Ph.D. level research talks and presentations from peer reviewed conferences around the globe. Most of the contents hosted in this repository come from Computer Science related knowledge areas such as Data Science, Semantic Web, Big data etc. although there is a reasonably big collection of content available from other fields such as Biology, Physics, Arts etc... All the field categories present in the dataset are outlined in the section below. In terms of different features available, data pertaining to the lecture such as its authors, author affiliations, research event and venue related information is available with supplementary information such as the slides from the presentations from the respective lectures. In addition to this, anonymised user session data relating to what parts of the lecture learners watched and skipped is also available giving a detailed viewed into user engagement and interaction with these lectures.

The text transcriptions and multiple translations of the video content is also available with the dataset. Translectures project transcribes and translates video content in Videolectures repository to text for subtitling. This enables learners who may be fluent in different languages to have closed-captioning on video lectures to improve their learning experience.

6.1.1.1. Field Categories (Subjects)

The lectures in videolectures.net repository is divided into field categories. The full taxonomy of field categories is a tree structure with 629 leaf categories. However, We categorise lectures only up to the top-most-level field categories for this analysis. *Table 1* outlines the set of **21** top-most categories used.

Category
Philosophy
Science
Computers
Astronomy
Military
Humanities
Chemistry
Earth Sciences
Arts
Architecture
Medicine
Mathematics
Technology
Business
Environment
Physics
Social Sciences
Computer Science
Data Science
Regional
Biology

Table 1: *List of field categories of lectures*

6.1.1.2. Potential Labels

This dataset carries three main attributes that potentially represents quality as a target variable.

Star Ratings: Users can rate the lecture using a star rating. A user can assign 1 to 5 stars for a lecture and the average star rating across all the ratings per lecture is used displayed with the lecture to every user who views a lecture.

Hotness Score: Hotness score is an internal metric used by videolectures.net to rank their lectures. (5) defines how the hotness score for any lecture is calculated.

$$Hotness = \frac{Number\ of\ views}{Number\ of\ days\ since\ publication^2} \quad (5)$$

Where

Number of days since publication = current date – date of publication

Unfortunately, we haven't been able to find any work that has used (5) as an indicator of quality. But we decide to not to rule it out until we analyse its suitability.

Engagement: Engagement can be calculated using the user sessions for the lectures. (6) defines how the engagement rate for each lecture can be calculated.

$$Engagement\ Rate = \frac{Total\ Duration\ of\ lectures\ watched\ by\ learner}{Length\ of\ lecture} \quad (6)$$

Total duration of lecture watch time can be calculated using the user session data available. Length of lectures is available as a field in the lecture data provided by videolectures.net.

YouTube[12] uses watch time as one of the main measurements of engagement with their videos *[Meyerson (2012)]*.

6.1.2. Volume

All the lecture related data and a subset of anonymised user engagement data was provided for deriving quality models. The volume of raw data initially considered for the experiments is summarised below.

- 7, 040 organizations/ universities that authors are affiliated to
- 16, 438 authors
- 25, 697 individual lectures
- 26, 042 raw videos that belong to the 25, 697 lectures mentioned above
- 155, 850 anonymised user sessions (how people navigate through the videos)

As mentioned in the previous section, caption transcriptions are also available for these videos.

- 76, 472 transcriptions (in original language) and translations

One of the most important statistics to consider is how much data is usable out of this full dataset. This is the number of lectures where labels can be derived.

- 3, 014 lectures with at least a single star rating
- 25, 230 lectures with hotness score
- 14, 877 lectures with at least one engagement datapoint
  - 6, 270 lectures with 5 or more user sessions
  - 3, 223 lectures with 10 or more user sessions

6.2. AVAILABLE TOOLS

In this section, we will outline numerous tools used in transforming data and training the machine learning models. Initially, we will discuss Apache Spark, Wikifier and PyCaption, tools that provide data processing capabilities relating to this dataset. Then we will proceed to Scikit-Learn

6.2.1. Apache Spark (PySpark)

Apache Spark[13] is an open-source cluster computing framework initially developed by the AMPLab at University of California, Berkeley, USA and later donated to Apache

[12] https://www.youtube.com
[13] https://spark.apache.org/

foundation. Apache Spark implements the ecosystem of software and driver tools needed to process data massively parallelly using MapReduce *[Dean & Ghemawat (2004)]* computing paradigm. While providing super-fast computation capabilities using in-memory computation *[Zaharia et al (2012)]* in comparison to its main contender Apache Hadoop[14], Spark also provides addional features such as being able to work with distributed data using sql or R like syntax *[Xin et al (2013)]*.

Having a python programming interface (PySpark), Apache Spark enables seamlessly working with other data science libraries such as Sciki-Learn. Although the programming style is a bit different from conventional single core programmes, Spark programs leverage taking full use of all the cores in a small computer when run in local mode yet can scale into 100s or 1000s of parallel computing cores when data gets bigger with no additional programming efforts.

The only disadvantages of Apache spark are the slightly different programming style and the few additional dependences that must be installed to the development environment.

Given that X5gon plans to ingest data from numerous repositories of different sizes, using parallel data processing capabilities of Apache Spark greatly reduces the risks that may arise with ingesting data from repositories at scale.

6.2.2. NLTK

Natural Language ToolKit (NLTK)[15] *[Bird et al (2009)]* is an open-source Python library for working with natural language data. It comprises of a range of text processing algorithms, NLP models and easy to use interfaces to corpora and linguistic resources such as English stop words, WordNet etc... NLTK is quite useful to us when extracting content related textual features from lecture transcripts.

6.2.3. Wikifier

Wikifer *[Brank et al (2017)]* is a web service which takes a text document as input and annotates it with links to relevant Wikipedia concepts. This service is hosted[16] and maintained by Josef Stefan Institute in Slovenia who is one of the project partners. The service supports cross and multi-linguality enabling extraction and annotations in different languages. This forms as the basis for comparing and analysing OER materials written in different languages. The tool was developed by Institut "Jožef Stefan" (IJS).

6.2.4. PyCaption

PyCaption[17] is the standard python programming library for working with subtitle files. This library allows reading/ converting and wring differently formatted subtitle files. Pycaption can deal with popular file formats such as SRT, DFXP, SAMI, WebVTT and Transcript. As the transcript files and translation files of the lectures are main stored in DFXP format, pycaption is a great candidate for working with the transcription files.

[14] www.hadoop.apache.org
[15] https://www.nltk.org/
[16] http://wikifier.org/
[17] https://pycaption.readthedocs.io/en/stable/

6.2.5. Scikit-learn

Scikit-Learn[18] is a very popular open source machine learning library that provides a user intuitive API to multiple families of machine learning algorithms such as Regression, Classification, Clustering etc... *[Pedregosa et al (2011)]*. Scikit-learn is a powerful candidate for developing machine learning models as it seamlessly integrates with the rest of pythonic tools that are used in the pipeline (pySpark etc...). A lot of business organizations use Scikit-learn in their production systems and it has also been proven to be stable in production settings.

The only disadvantage about this library is that it focusses on very common algorithms and therefore do not focus on niche areas in machine learning such as multi-task learning algorithms.

6.2.6. RMTL

RMTL: An R Library for Multi-task Learning[19] is an R library that implements a few machine learning algorithms that allow multi-task learning using approaches such as joint features selection with L_{21} norm *[Lui et al (2009)]* and trace-norm regularization *[Fang et al (2017)]*.

6.3. TOOLS DEVELOPED

In this section we discuss the tools we had to develop to enrich the data at hand. The main tool developed as part of data processing endeavour is the transcription conversion tool.

6.2.1. DFXP to Text converter

As mentioned in section above, videolecture.net repository also provides us with the transcriptions of the lecture videos. These transcriptions are provided in Distribution Format Exchange Profile (DFXP). Pycaption library is a great tool to convert .dfxp files to text files. The text version of the video is vital to quality analysis as those files are the only data source that will enable extracting features from the text about lecture content. Unfortunately, the library was unable to parse the .dfxp files from videolectures repository. In addition, we also needed a representation that can measure timing related to words and silence tags.

Due to these reasons, we decided to develop our own tool to read the .dfxp file and extract features such as the content of the lecture, the duration of silence tags in the lecture etc... We will discuss the exact features extracted from the files in the forthcoming section when we describe the final dataset.

6.3. FINAL DATASET

In this section, we describe the final dataset extracted from the raw data and was subjected to analysis and model training. As the main interest at this phase of the project is to identify absolute quality of content, the features extracted are features that wouldn't have any personal biases.

For the descriptive statistics of the different feature variables, please refer to *Appendix A1.*

[18] http://scikit-learn.org/stable/
[19] https://github.com/transbioZI/RMTL

6.3.1. Features

The features that we extracted for the study are mainly content based features that represent the quality related features. The included features are described below.

Document Entropy: Document entropy represents the degree of focus (cohesiveness). As found in found in *[Bendersky et al (2011)],* Document Entropy is defined as (7)

$$Entropy_D = -\sum_{w \in D} p_D(w)\ log\ p_D(w)$$

Where:

$$p_D(w_i) = \frac{term\ \ frequency\ _{w_i,D}}{\sum_{w_j \in D} term\ \ frequency\ _{w_j,D}} \quad (7)$$

According to (7), if the document entropy is low, this means that the lecture is focussed into a small number of topics as a few numbers of unique words are used. When the document entropy is larger, this means that the lecture includes a lot of topics and hence less focussed.

Easiness: Easiness measures the level of language that is being used to present the lecture. The frequently used Flesch-Kincaid reading ease test *[Flesch (1979)]* is used to measure the level of language of text. This test uses formula (8) to derive a score that corresponds to the reading levels in Table 1:

$$F{-}K\ reading\ ease = 206.835{-}1.015\left(\frac{total\ \ words}{total\ \ sentences}\right){-}84.6\left(\frac{total\ \ syllables}{total\ \ words}\right) \quad (8)$$

Score	School level	Notes
100.0-90.0	5th grade	Very easy to read. Easily understood by an average 11-year-old student.
90.0–80.0	6th grade	Easy to read. Conversational English for consumers.
80.0–70.0	7th grade	Fairly easy to read.
70.0–60.0	8th & 9th grade	Plain English. Easily understood by 13- to 15-year-old students.
60.0–50.0	10th to 12th grade	Fairly difficult to read.
50.0–30.0	College	Difficult to read.
30.0–0.0	College graduate	Very difficult to read. Best understood by university graduates.

Table 2: *Interpretation of F-K reading ease test*

Fraction of Complex words: This feature also captures the complexity of words used in the lecture. A Complex word is a word with more than three or more syllables[20]. This scalar is a fraction between 0 and 1 computed according using (9).

$$Fraction\ Complex\ Words_{\ D} = \frac{\sum_{w_i \in\ D} S_i}{total\ words\ in\ D} \begin{cases} S_i = 1 & \text{if } w_i \text{ is a complex word} \\ S_i = 0 & otherwise \end{cases} \quad (9)$$

Fraction of silent words: This is the faction of silence in the video. In the subtitle files, there are tags that indicate durations where no words are spoken. This value is a scalar between 0 and 1 computed according to (10)

$$Fraction\ Silence_{\ D} = \frac{\sum_{s_i \in\ D} S_i}{total\ duration\ of\ D}$$

Where:

S_i is the duration of a silent phrase in lecture D (10)

Fraction stopword coverage: This is the proportion of stopwords that are covered in a document. Where there are words **w_i** in document **D**, and there exists the global stopword set **S**, this value is calculated using (11) by dividing the number of unique stopwords in document **D** by the number of stopwords in the stopword set **S.**

$$Fraction\ Stopword\ Coverage_{\ D} = \frac{|\{w_i \in\ D, w_i \in S\}|}{|S|} \quad (11)$$

Fraction stopword presence: This is the proportion of stopwords that are in the document. Where there are words **w_i** in document **D**, and there exists the global stopword set **S**, this value is calculated using (12) by dividing the number of stopword occurrences in document **D** by the number of words in the document**.**

$$Fraction\ Stopword\ Presence\ D = \frac{|[w_i \in\ D, w_i \in S]|}{|D|} \quad (12)$$

Fraction stopword coverage and *Fraction stopword presence* together represent the style of language in the lectures. These two features will help us understand if there are stylistic preferences that attribute to better quality. Both fraction of stopwords coverage and presence can be used to represent the divergence between the document and language models which have been used as quality predictors before *[Zhou & Croft (2005)]*

Published date epoch days: This is the published date using epoch days. In other words, this feature is the time difference (in days) between the lecture publish date and January 01, 1970. The bigger this number is, fresher the lecture is. Smaller this value is, older the lecture is.

[20] https://github.com/nltk/nltk_contrib/blob/master/nltk_contrib/readability/textanalyzer.py#L94

Title word count: Number of words in the lecture title. *Ntoulas et al [2006]* has found the number of words in the title to be a useful feature for spam webpage detection.

Word count: Number of words in the lecture. Represents the duration/ effort a learner must commit to complete the educational resource.

6.3.2. Labels

After interpreting Hotness score, it is evident from (5) that this value suffers a heavy discount everyday as the denominator of (5) is **days**^{2}. This means that the number of views have to increase exponentially over time to keep the hotness score constant. This suggests that hotness, as its name represents popularity more than it represents the quality of content.

Based on the numbers in data volume section, it is evident that Star ratings are two scarce as target variable.

6.3.2.1. Median Engagement Rate

Engagement on lectures is the best candidate for a target variable that represent quality. Engagement with a lecture indicates that the user is motivated to stay with the educational resource for longer. We use equation (6) to compute *Engagement Rate* per lecture per session. As there are multiple sessions per lecture, we compute a summary statistic that represents all the sessions per lecture. There are two types of main outliers that deviate the centre in our case.

1. Users who immediately leave the lecture as soon as they view the page without giving any time to assess the quality of material. This occurs mainly when learners arrive in the page mistakenly. The summary engagement rate should not be sensitive to such cases.
2. Users who watch the videos repeatedly in the same session leading to engagement rates far greater than 1. The shorter the lecture is the engagement rate becomes larger due to equation (6).

Median is the most robust centre statistic amid outlier scenarios outlined above. We use **median engagement rate** of lecture as the target label.

6.3.3. Pairwise Preference Setting

In the pairwise preference scenario, we make pairwise preferences between lectures in the same field category (Computer Science, Biology, Philosophy etc...). The pairing and data preparation procedure is outlined by (13). The distance between the features of the pair of lectures compared ($\mathbf{x}_{l1}$ and $\mathbf{x}_{l2}$) becomes the feature set ($\mathbf{x}_{l1,l2}$). The label is a Boolean variable which turns True if the Median Engagement Rate of lecture l1 is greater than that of lecture l2 where the engagement rate of sessions is calculated using equation (6).

$$x_{l1,l2} = x_{l1} - x_{l2}$$

$$y_{l1,l2} = \begin{cases} \textbf{True if } MER_{l1} > MER_{l2} \\ \textbf{False otherwise} \end{cases}$$

$$\begin{aligned} &\textbf{where :} \\ &x_{l1,l2},\ x_{l1},\ x_{l2} \in \mathbb{R}^d \\ &l_1 \neq l_2 \\ &category(l_1) = category(l_2) \\ &MER_l \Rightarrow Median\ Engagement\ Rate\ of\ lecture\ l \end{aligned} \tag{13}$$

7. MODEL TRAINING

Model training is done in the conventional setting. Initially, the lecture data is partitioned into 70%:30% Train: Test split using Stratified Sampling. The stratification is done on category fields outlined in Table 1. This way, 70% of lectures from every category fall into the training set while 30% falls into the test set.

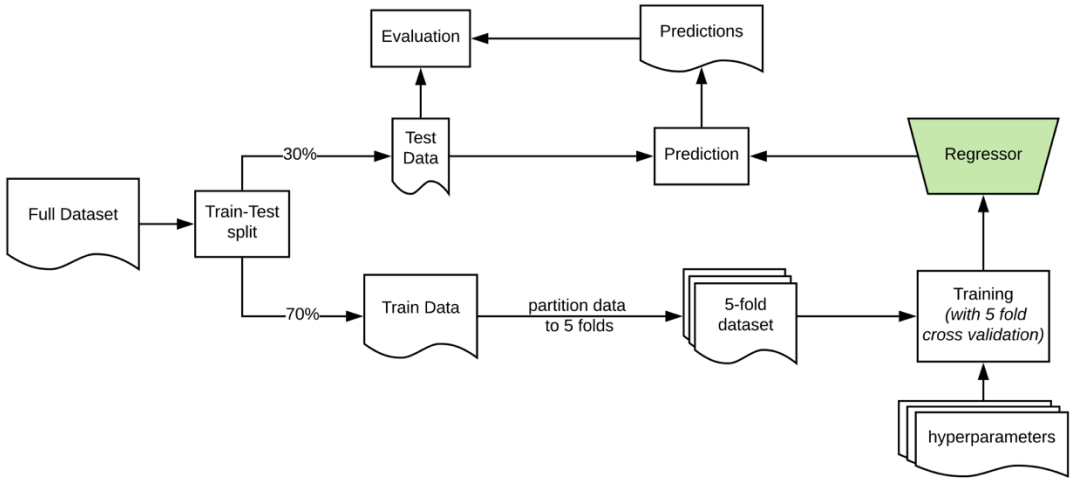

Figure 2: *Training process of Ridge Regression model*

As shown in Figure 2, this split dataset is used for training the regression model. 5-fold cross validation is used to find the optimal regularization parameter for ridge regression.

In the pairwise preference case (outlined in Figure 3), the partitioned data is then processed to generate pairwise observations within the train and test sets complying to the conditions in (13). The pairing is done after the train test split has been created as shown in Figure 3. This assures that any pair of observations ($l_1 > l_2$ and $l_2 < l_1$) do not fall to the train and test sets respectively hence guaranteeing there is no label leakage.

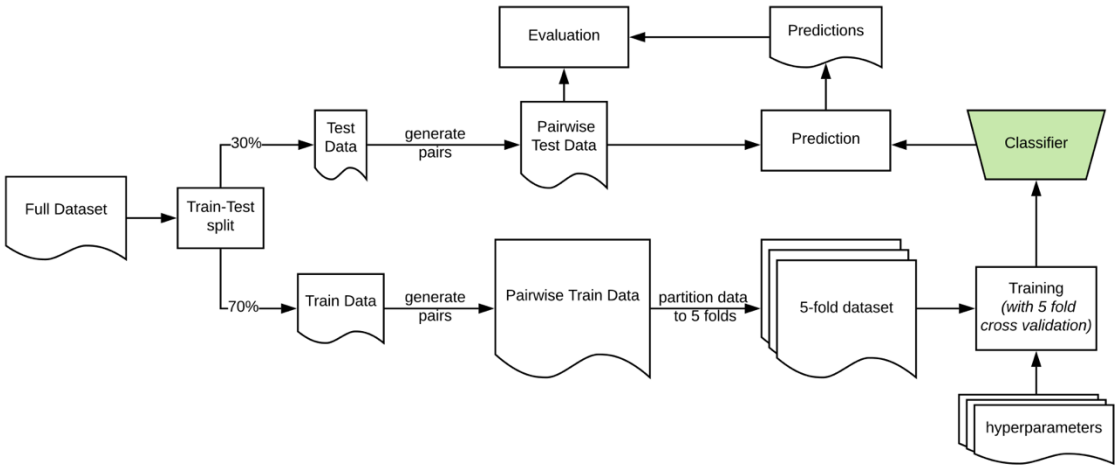

Figure 3: *Training process of Classification models*

As you have noticed from figure 2 and 3, there are regularization related hyper parameters relating to all the models proposed. During training, hyperparameter turning for the models is done using 5-fold cross validation.

8. RESULTS

This section summarises the main results obtained from the three main models fitted to the dataset.

1. Ridge Regression (RR)
2. Linear Pairwise Classification using SVM (SVM)
3. Pairwise Multitask Classification using Trace norm regularization (MTL)

The following section only presents a concise summary of the model evaluation process and the selection criteria. Please refer to *deliverable D1.2* for a more detailed description of the evaluation and selection process.

8.1. EVALUATION METRICS

Multiple evaluation metrics had to be used when evaluating models as both Regression and Classification models have been used to solve this ranking problem.

8.1.1. Regression

In the context of ridge regression, Root Mean Squared Error (RMSE) is one of the widely used evaluation metrics for regression problems. RMSE indicates model's predictive power with the mean deviation between the prediction and the true value on the test set.

Our objective is more to predict the relative rank of lectures in terms of quality rather than to predict the exact median engagement rate. Although RMSE is a highly suitable metric for regression problems, the problem we face is a "ranking" problem that is framed in the form of a regression problem. Due to this reason, a rank-correlation metric such as Spearman rank correlation coefficient is more suitable to evaluate the predictive power of the model than RMSE.

8.1.2. Classification

Accuracy Score can be considered as one of the main evaluation metrics used to evaluate classification models. However, as we use classification to solve a pairwise preference problem here, the ranking would be more accurate when the number of actual comparisons agree with the predicted outcomes ($l_1 > l_2$ vs. $l_2 > l_1$). In other words, the true ranking and the prediction ranking is more correlated when more pairwise comparisons are predicted correctly. Therefore, larger classification accuracy means more alignment between the true ranking and the predicted ranks of lectures.

8.2. Results Overview

RMSE and Spearman correlation coefficient (*Spearman R*) has been used to evaluate the regression results obtained by Ridge Regression. Table 3 summarises the evaluation results from the ridge regression model trained with the data.

Evaluation Metric	Training Data	Test Data
RMSE Training data	0.1907	0.1838
Spearman R (p-value[21])	0.5638 (7.63e-303)	0.5814 (6.01e-142)

Table 3: *Model evaluation results from Ridge Regression*

In the classification setting, classification accuracy is the metric used. Table 4 below summarises the classification accuracy score obtained for both SVM (Ranking SVM) and Multitask Classification using Trace norm (Trace-Norm MTL).

Classification Accuracy	Training Data	Test Data
Ranking SVM	0.7191	0.7121
Trace-Norm MTL	0.7210	0.7105

Table 4: *Classification Accuracy of Ranking SVM and Trace Norm MTL models*

8.3. COMPARING MODELS

To compare the different models developed to resolve this ranking problem, we need to convert the results from different to the same result space where we can fairly compare them. We use classification accuracy as the metric that is generalizable to all the models developed. In the regression case, we use the global rank to generate a pairwise preference dataset where classification accuracy is comparable. Please refer to deliverable D1.2 for further details about the process. Table 5 summarises the results from the classification accuracy results from the three models under investigation.

Classification Accuracy	Training Data	Test Data
Ridge Regression	0.7120	0.7115
Ranking SVM	**0.7191**	**0.7121**
Trace-Norm MTL	0.7210	0.7105

Table 5: *Final Comparison of all three models (i) Ridge Regression, (ii) Ranking SVM, and (iii) Trace-Norm MTL*

[21] within brackets is the p-value of Spearman correlation. Smaller p-values indicate that the correlation is highly significant.

9. DISCUSSION AND CONCLUSION

From our work so far, we can observe that quality of educational material is an ill-defined topic where very little work has been carried out. By investigating work done in Education sector and several other domains such as Information search and healthcare, it was possible to identify a series of potential features that are indicative of quality.

It is also evident that Star ratings are quite a popular target variable used to evaluate quality. However, star ratings are quite scarce in real world datasets and usually one has to resort to an alternate implicit feedback signal that is available in larger scale.

From the results, it is evident that both regression and classification techniques perform well on understanding the features that determine superior quality (71% accuracy on average). It is also observable that the models perform "equally well" on the task. Table 5 also gives strong evidence that the models are quite robust in terms of fighting overfitting as train accuracy and test accuracy of all the models align well.

9.1. CONCLUSION

From the current study, we can conclude the following:

- Quality of content is determined by attributes that fall under five main verticals.
  - Coverage of a topic
  - Authority
  - Understandability (easiness to understand)
  - Presentation quality
  - Freshness
- Quality is best represented by Explicit feedback such as star ratings
- However, at the scarcity of such data points, implicit feedback such as video watch time, clickthrough rate are suitable alternatives for measuring quality of content.
- All three approaches used perform equally well on the prediction task while giving reasonably good results in general (71% classification accuracy)
- Based on results obtained in Table 5, it is fair to say that all models are robust against overfitting,
- Based on the held-out data performance summarised in Table 5, we can conclude that using an SVM for pairwise preference classification is the most suitable model to go forward.
- SVM is suitable due to the following reasons:
  - Superior held-out set (test data) performance
  - No evidence of overfitting
  - Simple highly interpretable model

9.2. FUTURE WORK

There are numerous other algorithms that have shown to outperform RankSVM in pairwise classification. RankNet, LambdaNet and LambdaMART are few of those algorithms *[Burges (2010)]*. We can make immediate performance improvements by using these more sophisticated neural ranking models.

We are expanding our work to understand how to improve our evaluation metrics and training process by accounting for the quality difference between lectures.

A detailed explanation of our future research avenue with pointers of diagnostic evidence is found in section 7.3 in deliverable D1.2.

REFERENCES

Janez Brank, Gregor Leban, Marko Grobelnik. Annotating Documents with Relevant Wikipedia Concepts. Proceedings of the Slovenian Conference on Data Mining and Data Warehouses (SiKDD 2017), Ljubljana, Slovenia, 9 October 2017

Celia Boyer, Cedric Frossard, Arnaud Gaudinat, Allan Hanbury, Gilles Falquetd. 2017. How to sort trustworthy health online information? Improvements of the automated detection of HONcode criteria, Procedia Computer Science, Volume 121, Pages 940-949, DOI= https://doi.org/10.1016/j.procs.2017.11.122.

Peter Pirolli & Sanjay Kairam. 2013. User Model User-Adap Inter (2013) 23: 139. https://doi.org/10.1007/s11257-012-9132-1

Kevyn Collins-Thompson, Paul N. Bennett, Ryen W. White, Sebastian de la Chica, and David Sontag. 2011. Personalizing web search results by reading level. In Proceedings of the 20th ACM international conference on Information and knowledge management (CIKM '11), Bettina Berendt, Arjen de Vries, Wenfei Fan, Craig Macdonald, Iadh Ounis, and Ian Ruthven (Eds.). ACM, New York, NY, USA, 403-412. DOI: https://doi.org/10.1145/2063576.2063639

Luo Si and Jamie Callan. 2001. A statistical model for scientific readability. In Proceedings of the tenth international conference on Information and knowledge management (CIKM '01), Henrique Paques, Ling Liu, and David Grossman (Eds.). ACM, New York, NY, USA, 574-576. DOI= http://dx.doi.org/10.1145/502585.502695

Micheal V. Yudelson, Kenneth R. Koedinger, Geofferey J. Gordon. 2013. Individualized Bayesian Knowledge Tracing Models. In: Lane H.C., Yacef K., Mostow J., Pavlik P. (eds) Artificial Intelligence in Education. AIED 2013. Lecture Notes in Computer Science, vol 7926. Springer, Berlin, Heidelberg

Rohail Syed and Kevyn Collins-Thompson. 2017. Retrieval Algorithms Optimized for Human Learning. In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17). ACM, New York, NY, USA, 555-564. DOI= https://doi.org/10.1145/3077136.3080835

Sergey Brin, Larry Page. 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*. 30: 107–117. DOI: 10.1016/S0169-7552(98)00110-X. ISSN 0169-7552

Parikshit Sondhi, Vinod Vydiswaran, and Cheng Xiang Zhai. 2012. Reliability prediction of webpages in the medical domain. In *Proceedings of the 34th European conference on Advances in Information Retrieval (ECIR'12)*, Ricardo Baeza-Yates, Arjen P. Vries, Hugo Zaragoza, B. Barla Cambazoglu, and Vanessa Murdock (Eds.). Springer-Verlag, Berlin, Heidelberg, 219-231. DOI= https://doi.org/10.1007/978-3-642-28997-2_19

Arnaud. Gaudinat, Natalia Grabar, and Celia Boyer. 2007. Machine Learning Approach for Automatic Quality Criteria Detection of Health Web Pages. In Proc. of the World Congress on Health (Medical) Informatics – Building Sustainable Health Systems, volume 129, pages 705–709

Celia Boyer and Ljiljana Dolamic. 2014. Feasibility of automated detection of HONcode conformity for health-related websites. International Journal of Advanced Computer Science and Applications (IJACSA), 5(3). DOI= http://dx.doi.org/10.14569/IJACSA.2014.050309

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363-370.

Daniel E. Rose and Danny Levinson. 2004. Understanding user goals in web search. In Proceedings of the 13th international conference on World Wide Web (WWW '04). ACM, New York, NY, USA, 13-19. DOI= http://dx.doi.org/10.1145/988672.988675

Emine Yilmaz, Manisha Verma, Nick Craswell, Filip Radlinski, and Peter Bailey. 2014. Relevance and Effort: An Analysis of Document Utility. In Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM '14). ACM, New York, NY, USA, 91-100. DOI= http://dx.doi.org/10.1145/2661829.2661953

Michael Bendersky, W. Bruce Croft, and Yanlei Diao. 2011. Quality-biased ranking of web documents. In Proceedings of the fourth ACM international conference on Web search and data mining (WSDM '11). ACM, New York, NY, USA, 95-104. DOI= https://doi.org/10.1145/1935826.1935849

Alexandros Ntoulas, Marc Najork, Mark Manasse, and Dennis Fetterly. 2006. Detecting spam web pages through content analysis. In Proceedings of the 15th international conference on World Wide Web (WWW '06). ACM, New York, NY, USA, 83-92. DOI= https://doi.org/10.1145/1135777.1135794

Xavier Amatriain. 2009. The Netflix Prize: lessons learned. TechnoCalifornia (29 September 2009). Retrieved July 23, 2018 from http://technocalifornia.blogspot.com/2009/09/netflix-prize-lessons-learned.html

Xavier Amatriain, 2012. Building industrial-scale real-world recommender systems. In Proceedings of the Sixth ACM Conference on Recommender Systems, RecSys '12, pages 7–8, New York, NY, USA, 2012. ACM

Pavel Serdyukov, Georges Dupret, and Nick Craswell. 2014. Log-based personalization: the 4th web search click data (WSCD) workshop. In Proceedings of the 7th ACM international conference on Web search and data mining (WSDM '14). ACM, New York, NY, USA, 685-686. DOI= http://dx.doi.org/10.1145/2556195.2556207

Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2017. Accurately Interpreting Clickthrough Data as Implicit Feedback. SIGIR Forum 51, 1 (August 2017), 4-11. DOI= https://doi.org/10.1145/3130332.3130334

Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep Neural Networks for YouTube Recommendations. In Proceedings of the 10th ACM Conference on Recommender Systems (RecSys '16). ACM, New York, NY, USA, 191-198. DOI: https://doi.org/10.1145/2959100.2959190

Eric Meyerson. 2012. Youtube now: Why we focus on watch time. (August 2012). Retrieved July 30, 2018 from http://youtubecreator.blogspot.com/2012/08/youtube-now-why-we-focus-on-watch-time.html

Theodoros Evgeniou and Massimiliano Pontil. 2004. Regularized multi-task learning. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '04). ACM, New York, NY, USA, 109-117. DOI= http://dx.doi.org/10.1145/1014052.1014067

Jun Liu, Shuiwang Ji, and Jieping Ye. 2009. Multi-task Feature Learning via Efficient l2,1-norm Minimization. In Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI '09), 339–348. Retrieved August 30, 2018 from http://dl.acm.org/citation.cfm?id=1795114.1795154

Meng Fang, Jie Yin, Lawrence O. Hall, and Dacheng Tao. 2017. Active Multitask Learning With Trace Norm Regularization Based on Excess Risk. IEEE Trans Cybern 47, 11 (November 2017), 3906–3915. DOI:https://doi.org/10.1109/TCYB.2016.2590023

Rudolf Flesch. 1979. HOW TO WRITE PLAIN ENGLISH, Chapter 2: Let's Start with the Formula, Retrieved July 25 2018 from http://pages.stern.nyu.edu/~wstarbuc/Writing/Flesch.htm

Yun Zhou and W. Bruce Croft. 2005. Document Quality Models for Web Ad Hoc Retrieval. In Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM '05), 331–332. DOI:https://doi.org/10.1145/1099554.1099652

Louis Leon Thurstone. 1927. A law of comparative judgement. Psychological Review, 34, 278–286.

Louis Leon Thurstone. 1929. The measurement of psychological value. Essays in Philosophy by Seventeen Doctors of Philosophy of the University of Chicago. Chicago: Open Court (1929), 157–174.

Louis Leon Thurstone. 1959. The Measurement of Values. Chicago: The University of Chicago Press.

Thomas L. Saaty. 2008. Relative measurement and its generalization in decision making why pairwise comparisons are central in mathematics for the measurement of intangible factors the analytic hierarchy/network process. Rev. R. Acad. Cien. Serie A. Mat. 102, 2 (September 2008), 251–318. DOI:https://doi.org/10.1007/BF03191825

Arpad E. Elo. 2008. 8.4 Logistic Probability as a Rating Basis. The Rating of Chessplayers. Past & Present. NY: Press International (2008).

Ralf Herbrich, Tom Minka, and Thore Graepel. 2006. TrueSkillTM: A Bayesian Skill Rating System. In Proceedings of the 19th International Conference on Neural Information Processing Systems (NIPS'06), 569–576.

Ralf Herbrich, Thore Graepel, Peter Bollmann-Sdorra, and Klaus Obermayer. 1998. Learning preference relations for information retrieval. In ICML-98 Workshop: text categorization and machine learning, 80–84.

Yoav Freund, Raj Iyer, Robert E. Schapire, and Yoram Singer. 2003. An Efficient Boosting Algorithm for Combining Preferences. J. Mach. Learn. Res. 4, (December 2003), 933–969.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. 2005. Learning to Rank Using Gradient Descent. In Proceedings of the 22Nd International Conference on Machine Learning (ICML '05), 89–96. DOI= https://doi.org/10.1145/1102351.1102363

Nello Cristianini and John Shawe-Taylor. 2000. An Introduction to Support Vector Machines: And Other Kernel-based Learning Methods. Cambridge University Press, New York, NY, USA.

Thorsten Joachims. 2002. Optimizing Search Engines Using Clickthrough Data. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '02), 133–142. DOI:https://doi.org/10.1145/775047.775067

Ralf Herbrich, Thore Graepel and Klaus Obermayer. 1999. Support vector learning for ordinal regression. In 1999 Ninth International Conference on Artificial Neural Networks ICANN 99. (Conf. Publ. No. 470), 97–102 vol.1. DOI:https://doi.org/10.1049/cp:19991091

Andrew Y. Ng. 2004. Feature Selection, L1 vs. L2 Regularization, and Rotational Invariance. In Proceedings of the Twenty-first International Conference on Machine Learning (ICML '04), 78–. DOI:https://doi.org/10.1145/1015330.1015435

Johannes Fürnkranz and Eyke Hüllermeier. 2011. Preference Learning and Ranking by Pairwise Comparison. In Preference Learning, Johannes Fürnkranz and Eyke Hüllermeier (eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 65–82. DOI:https://doi.org/10.1007/978-3-642-14125-6_4

Wikipedia. 2018. Multi-task learning. Retrieved August 30, 2018 from https://en.wikipedia.org/w/index.php?title=Multi-task_learning&oldid=840806852

Jun Liu, Shuiwang Ji, and Jieping Ye. 2009. Multi-task Feature Learning via Efficient L2, 1-norm Minimization. In Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI '09), 339–348.

Massimiliano Pontil and Andreas Maurer. 2013. Excess risk bounds for multitask learning with trace norm regularization. In Conference on Learning Theory, 55–76.

Christian Widmer, Marius Kloft, Nico Görnitz, and Gunnar Rätsch. 2012. Efficient Training of Graph-Regularized Multitask SVMs. In Machine Learning and Knowledge Discovery in Databases (Lecture Notes in Computer Science), 633–647.

Jiayu Zhou, Jianhui Chen, and Jieping Ye. 2011. Clustered Multi-Task Learning Via Alternating Structure Optimization. Adv Neural Inf Process Syst 2011, (2011), 702–710.

Wikipedia. 2018. VideoLectures.net. Retrieved August 30, 2018 from https://en.wikipedia.org/w/index.php?title=VideoLectures.net&oldid=849915975

TransLectures – Transcription and Translation of Video Lectures | Knowledge 4 All Foundation Ltd. Retrieved August 30, 2018 from http://www.k4all.org/project/translectures/

Jeffrey Dean and Sanjay Ghemawat. 2008. MapReduce: Simplified Data Processing on Large Clusters. Commun. ACM 51, 1 (January 2008), 107–113. DOI:https://doi.org/10.1145/1327452.1327492

Matei Zaharia, Mosharaf Chowdhury, Tathagata Das, Ankur Dave, Justin Ma, Murphy McCauley, Michael J. Franklin, Scott Shenker, and Ion Stoica. 2012. Resilient Distributed Datasets: A Fault-tolerant Abstraction for In-memory Cluster Computing. In Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation (NSDI'12), 2–2.

Reynold S. Xin, Josh Rosen, Matei Zaharia, Michael J. Franklin, Scott Shenker, and Ion Stoica. 2013. Shark: SQL and Rich Analytics at Scale. In Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data (SIGMOD '13), 13–24. DOI: https://doi.org/10.1145/2463676.2465288

Steven Bird, Ewan Klein, and Edward Loper. 2009. Natural Language Processing with Python (1st ed.). O'Reilly Media, Inc.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12, (November 2011), 2825–2830.

APPENDIX

A1: DESCRIPTIVE STATISTICS OF THE FINAL DATASET

In this section, we describe the full VLN repository dataset used for the analyses. In the first section, how the lectures are distributed over different field categories is presented. Then, the mean and standard deviation of different features variables are presented in the next section.

A1.1. Frequency of lectures for each field category in the dataset

Table A1.1 explains in detail how many lectures were available in the dataset based on the field category each lecture belongs to. If a lecture belongs to multiple categories, that lecture will be included in all the categories it belongs to.

Category	Observations
Philosophy	60
Science	122
Computers	138
Military	14
Humanities	110
Chemistry	76
Earth Sciences	11
Arts	81
Architecture	29
Medicine	103
Mathematics	280
Technology	249
Business	116
Environment	34
Physics	189
Social Sciences	367
Computer Science	2,763
Data Science	231

Regional	53
Biology	155
Total	***5,181***

Table A1.1: *Frequency of lectures belonging to different field categories*

A1.2. Mean and Standard Deviation of features and labels

In this section, we present the value distribution for each features and label in the final dataset. The mean and standard deviation of the values are presented in a plot where the statistics are computed for each field category. The value distribution is marked using a coloured vertical bar where the dot in the centre of the bar is the mean value of that feature. The range of the bar is 1 standard deviation above and below from the mean value.

A1.2.1. Features

In this section, we outline the descriptive statistics of each feature in the following order.

- Document Entropy (Figure A1.1)
- Easiness (Figure A1.2)
- Fraction of Complex Words (Figure A1.3)
- Fraction of Silent Words (Figure A1.4)
- Fraction Stopword Coverage (Figure A1.5)
- Fraction Stopword Presence (Figure A1.6)
- Published Date Epoch Days (Figure A1.7)
- Title Word Count (Figure A1.8)
- Word Count (Figure A1.9)

Please refer to subsection *6.3.1* for a detailed account of how exactly the features are computed. For a full list of field categories, please refer to Table 1.

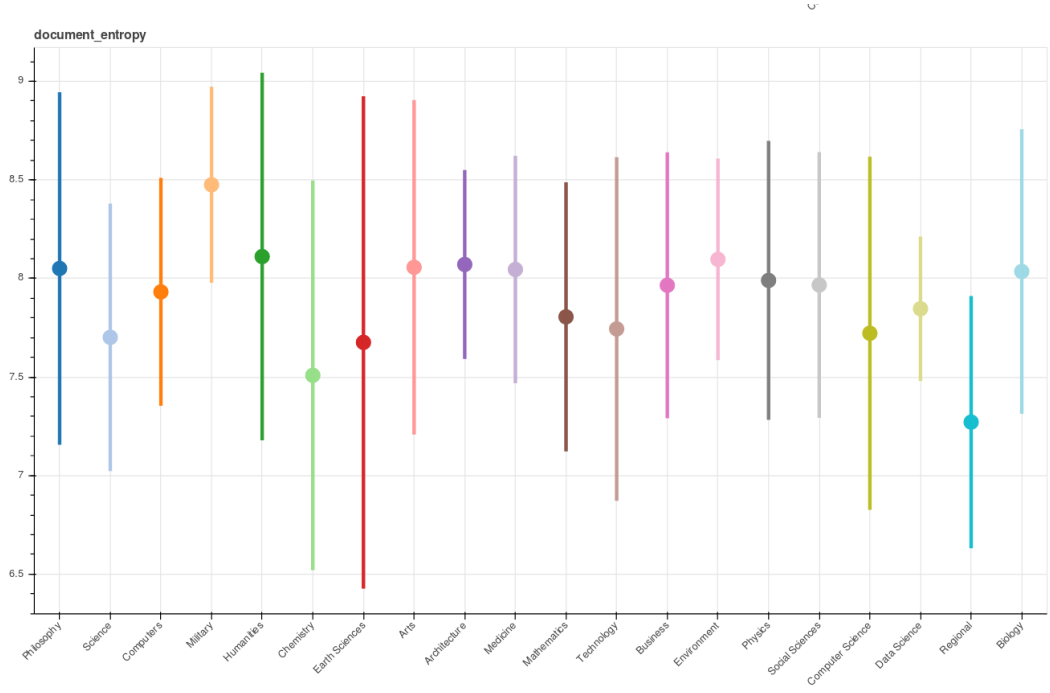

Figure A1.1: *Value distribution of document entropy*

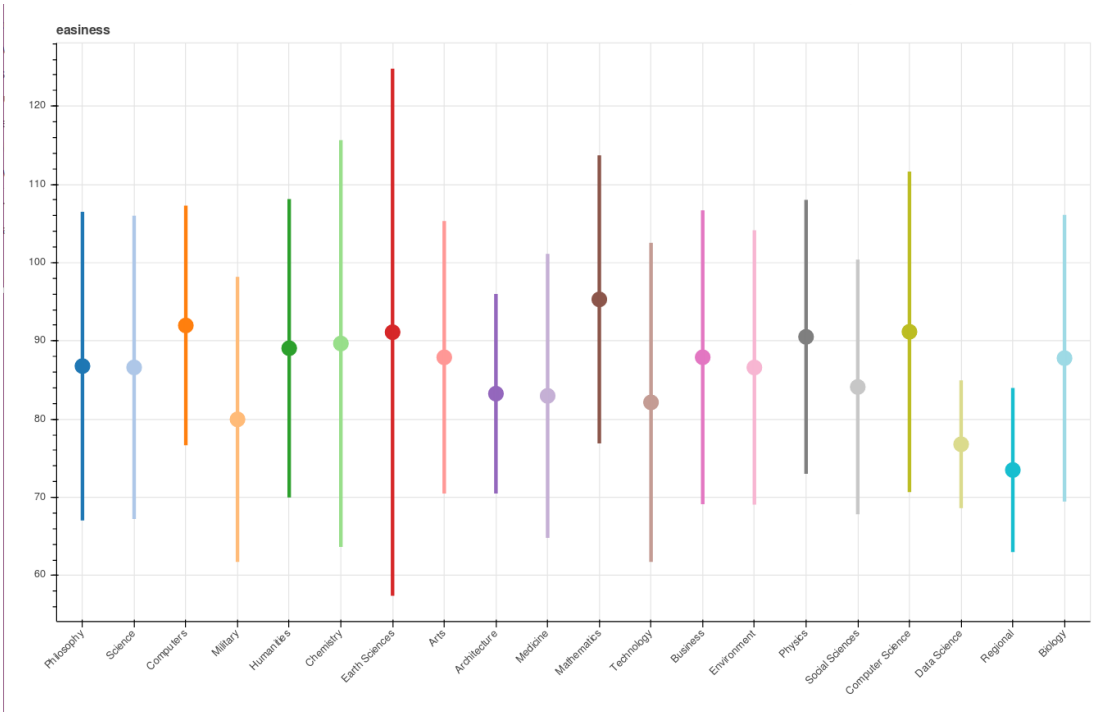

Figure A1.2: *Value distribution of easiness*

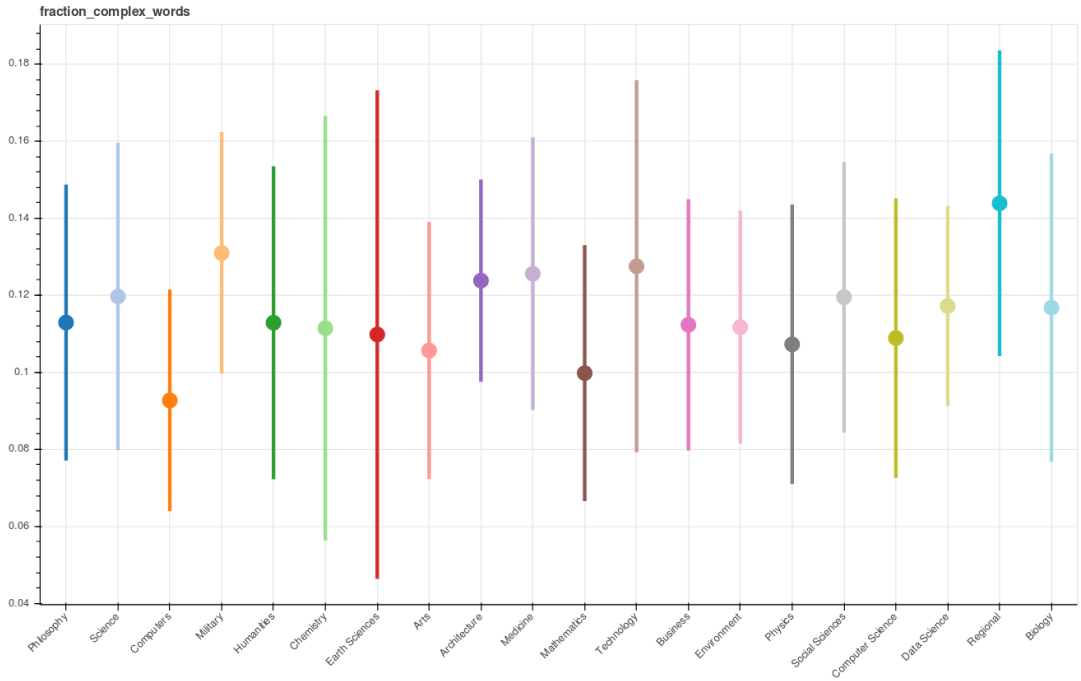

Figure A1.3: *Value distribution of fraction of complex words*

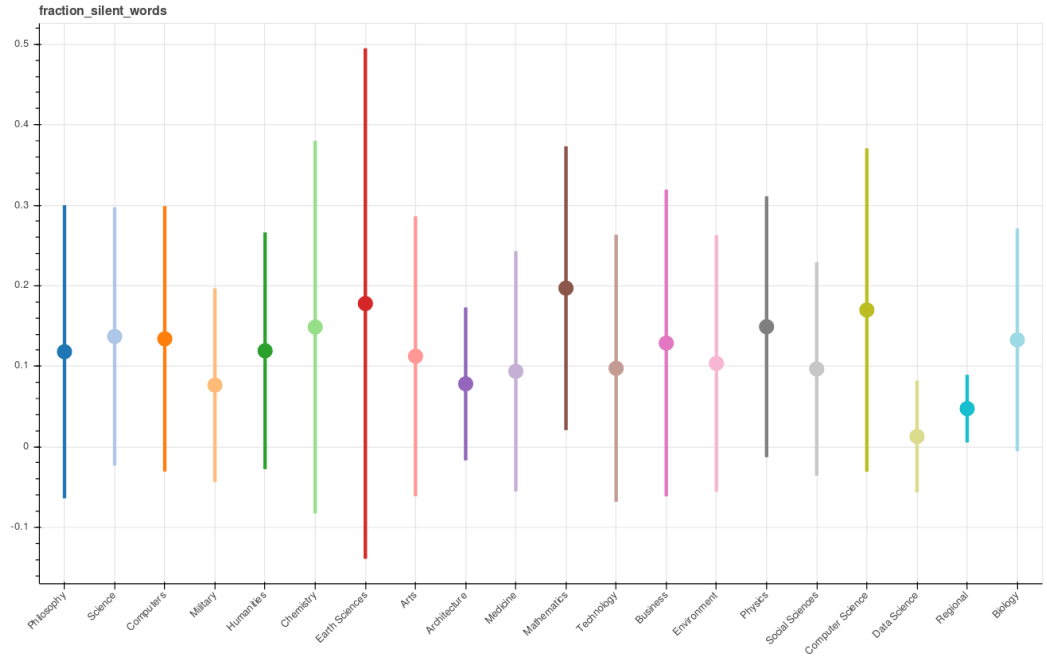

Figure A1.4: *Value distribution of fraction of silent words*

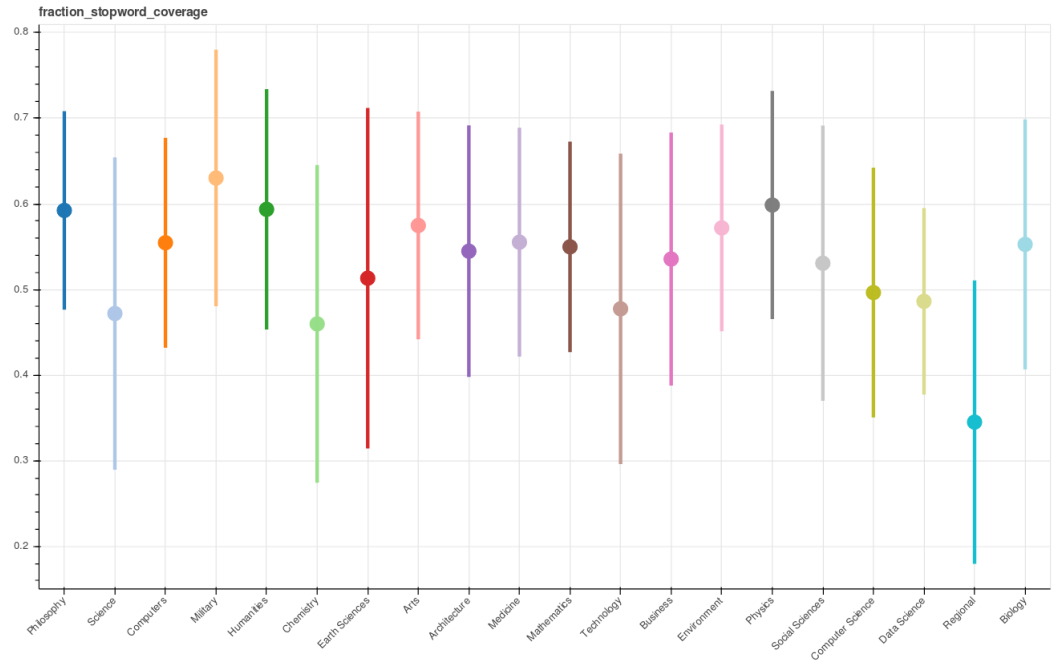

Figure A1.5: *Value distribution of fraction stopword coverage*

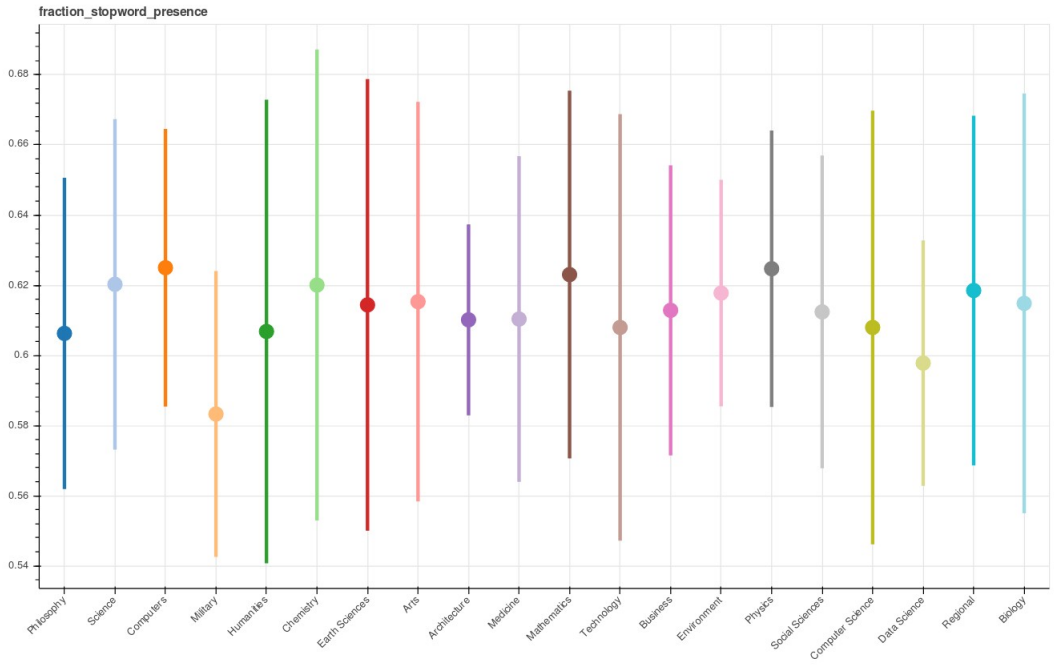

Figure A1.6: *Value distribution of fraction stopword presence*

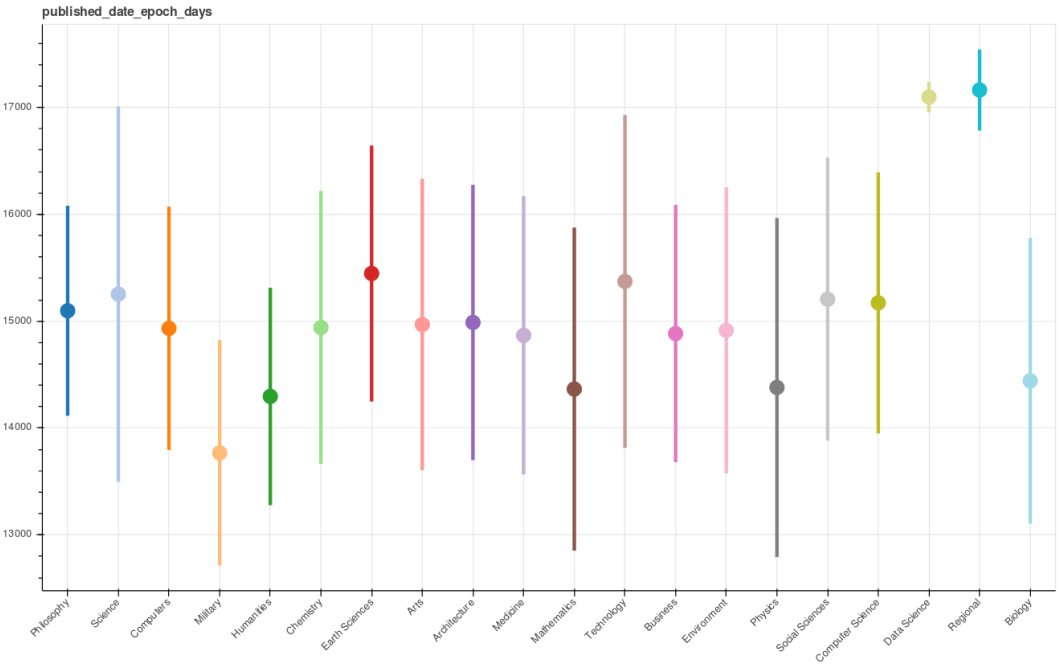

Figure A1.7: *Value distribution of published date epoch days*

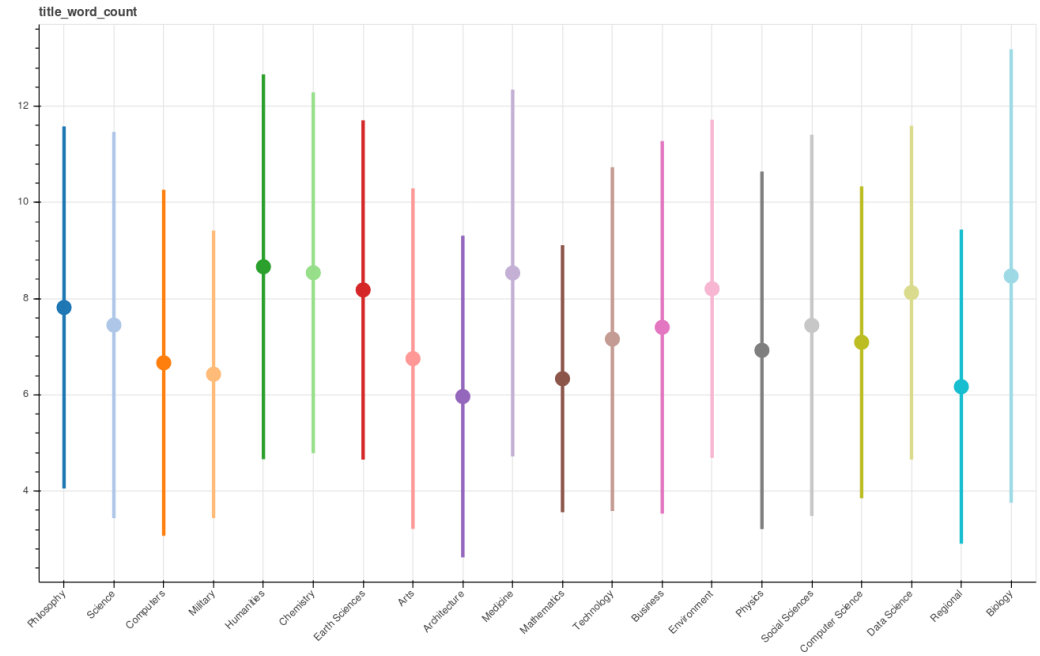

Figure A1.8: *Value distribution of title word count*

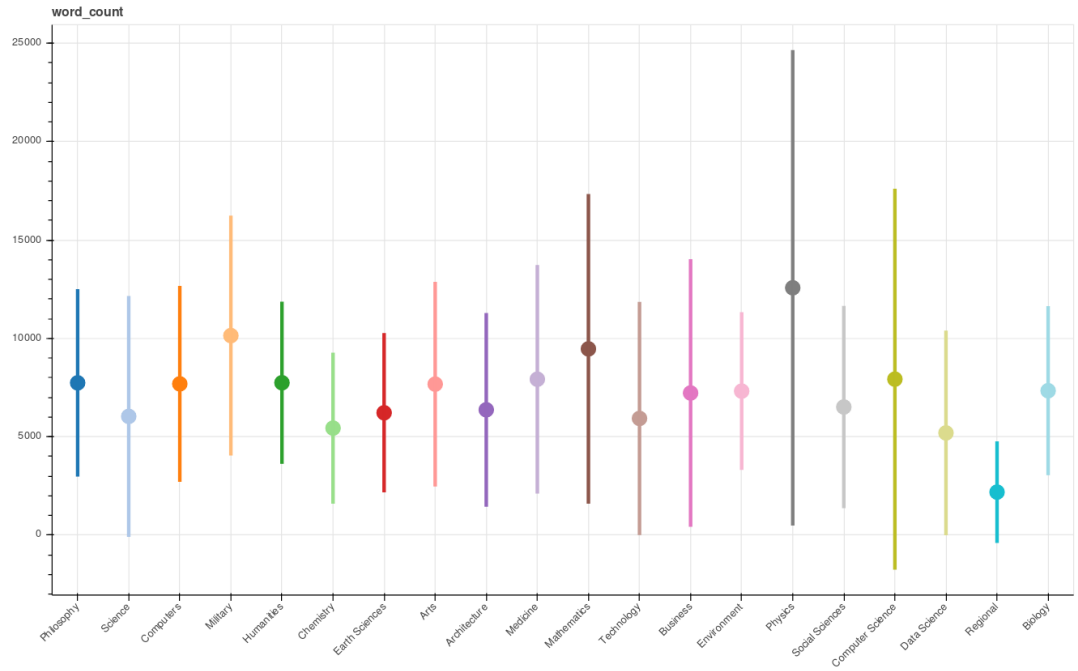

Figure A1.9: *Value distribution of word count*

A1.2.2. Labels

In this section, we show the descriptive statistics of median engagement rate. Figure A1.10 outlines the mean and standard deviation values for median engagement rate categorised by different field categories.

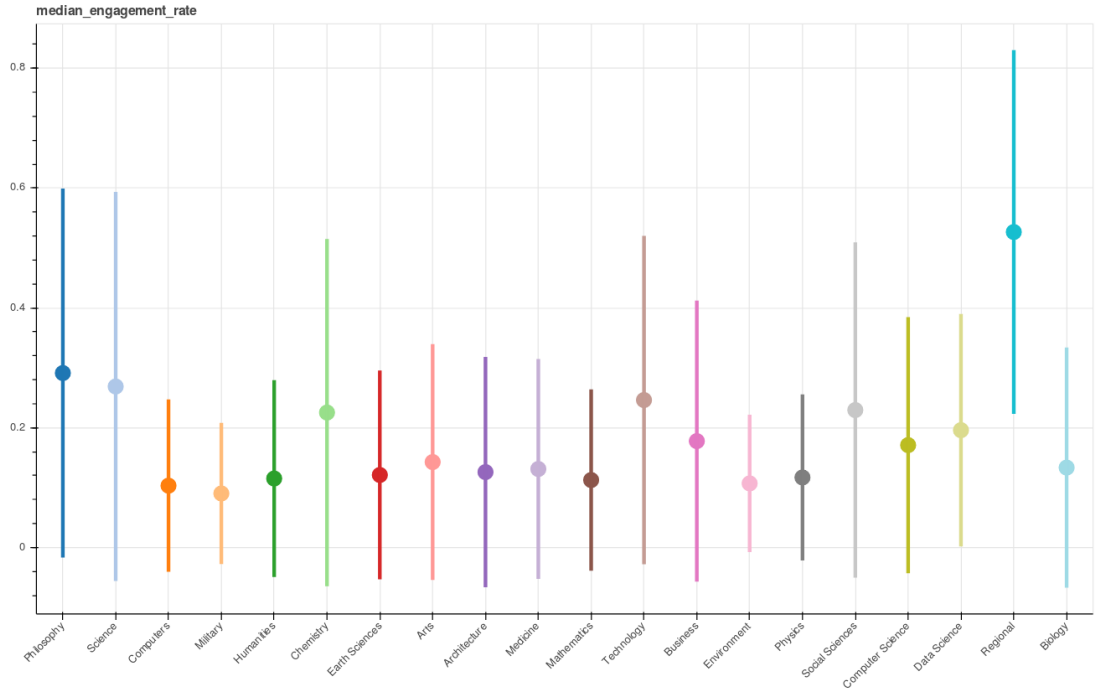

Figure A1.10: *Value distribution of median engagement rate*

Please refer to subsection *6.3.2* for a detailed account of how exactly the label variable is computed. For a full list of field categories, please refer to Table 1.