S	
E1	
E2	
For Office Use Only	

Masters Project Final Report

(MCS)

2019

Project Title	Automatic creation of e-learning content using blog posts
Student Name	K. A. S. N. Wijerathna
Registration No. & Index No.	2015MCS080 15440802
Supervisor's Name	Dr. T. A. Weerasinghe

For Office Use ONLY

Automatic creation of e-learning content using blog posts

A dissertation submitted for the Degree of Master of Computer Science

K. A. S. N. Wijerathna

University of Colombo School of Computing

2019

Declaration

The thesis is my original work and has not been submitted previously for a degree at this or any other university/institute.

To the best of my knowledge it does not contain any material published or written by another person, except as acknowledged in the text.

Student Name: K. A. S. N. Wijerathna

Registration Number: 2015MCS080

Index Number: 15440802

____

Signature: Date:

This is to certify that this thesis is based on the work of

Mr./~~Ms.~~ K. A. S. N. Wijerathna

under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Certified by:

Supervisor Name: Dr. T. A. Weerasinghe

____

Signature: Date:

Abstract

In the contemporary world, blogging is no longer used just for personal stuff and providing expert opinions, blogs often carry useful "learning" content, but most of the time in unstructured manner. Tech companies often use internal and external blogs to allow their employees to share the expert knowledge within the company instead of incorporating this knowledge with existing e-learning tools like virtual learning environments as it takes an extra effort and cost to integrate the knowledge in the blogs with the e-learning tools. Deriving e-learning course content from blogs helps those companies as they can still use their own blogs and adopt e-learning tools without much burden and cost.

This study presents a model to automate the process of integrating the new knowledge from the blogs with an e-learning platform. There are three main issues addressed in the proposed model: identifying the appropriate methods to extract knowledge content from blog posts, extract the content from blogs using identified methods and create learning content, integrate the created learning content with selected e-learning tool. Finally, the developed model was run against the selected blog posts and the created e-Learning content was evaluated using human judgement.

Normally a blog page contains boilerplate and a lot of clutter like pop-ups, extra links and unnecessary icons. Hence extracting the "meat" of a blog page and eliminating the clutter is an eminent task. Content extraction from webpages, blogs is heavily studied area as there are lot of applications like sentiment analysis, content summarization and text classification. Also, there are already built open source and proprietary tools to extract the core content from the web pages. Each of them has their own merits. This study evaluates python based three content extractor libraries: Readability, Boilerpipe and NewsPaper to find out which one is giving the best results.

This study revealed that among six of the selected extractors, both Readability and NewsPaper libraries outperform other four extractors in Boilerpipe library. When comparing Readability and NewsPaper libraries, Readability library has a slight edge over the NewsPaper library as it can extract code segments and emojis as well. When integrating the extracted content with the

Moodle, an already available Forum has been used. The content was transferred as a new discussion under the forum using Moodle REST API functions.

The proposed model can extract the content from a blog and then extracted content transferred to Moodle. Without manually creating the course content by copy and pasting, trainers and technical experts can easily integrate their blogs with an e-Learning system.

Keywords*: blogs, web content extraction, e-learning, lms integration*

Acknowledgement

First, I would like to thank University of Colombo School of Computing for the opportunity given me to study for the Master of Computer Science program.

Foremost I would like to extend my sincere thanks to Dr. Thushani Weerasinghe, my research project supervisor, for making the opportunity to explore my knowledge in the area of e-learning and for encouraging me technically and spiritually during my studies. Her guidance helped me in all the time of the research.

Further, I would like to thank the five Instructional Designers from the e-learning center of the UCSC who have participated in the survey which I have conducted to gather the requirements for usability aspect of the proposed model. Then I wish to thank the other staff members of the UCSC, for their support which directs us to achieve our goals in life.

Special thank should go to Gihan Ubayawardana and Oshada Samarasinghe for their voluntary participation in the blinded experiment conducted for assessing the outcome of the proposed model and for the continuous support and encouragement.

I want to express my most profound gratitude to my parents, wife, brother and other family members for their love, support, patience and all the encouragement. They all kept me going and this work would not have been a reality without them.

Last but not least, I would like to take this opportunity to express my gratitude to everyone who has supported me throughout the course of this MCS project.

Contents

List of Figures

List of Tables

List of Abbreviations

API	Application Programming Interface
CLT	Cognitive Load Theory
CMM	Conditional Markov Model
CMS	Course Management System
DOM	Document Object Model
FTP	File Transfer Protocol
ID	Instructional Design
IE	Information Extraction
IR	Information Retrieval
ISD	Instructional Systems Design
LCMS	Learning Content Management System
LMS	Learning Management System
LSS	Learning Support System
ML	Managed Learning Environment
MLE	Machine Learning
NLP	Natural Language Processing
OSS	Open Source Software
REST	Representational State Transfer
SCORM	Shareable Content Object Reference Model
SME	Subject Matter Expert
TM	Text Mining
UMELMS	Ubiquitous Multimedia Enhanced Learning Management System
VLE	Virtual Learning Environment
XHTML	EXtensible HyperText Markup Language

Chapter 1

Introduction

This study tries to identify an appropriate method to extract the knowledge content from blog posts and integrate those knowledge content with an existing Learning Management System (LMS). The topic is related to Instructional Design in e-learning as well as Information Extraction (IE) which is a subfield of Natural Language Processing (NLP) [1]. This chapter describes the background of the study and gives the introduction to the upcoming activities of the study.

1.1 Blogs

Blog (short form of "weblog") is a discussion or a website published for informational purposes [2] and blogs closely resemble diary or journal, but they are not necessarily private and most of the blogs have audience [3] and consist of text entries called "posts". Blogs have many uses in many areas and motivations for blogging are different form one another. According to Luján-Mora and Juana-Espinosa [4], five major motivations factors influence blogging. Those are: "documenting one's life; providing commentary and opinions; expressing deeply felt emotions; articulating ideas through writing; and forming and maintaining community forums".

Blogs can be viewed as an improved version of traditional learning materials for both teachers and students as they can use blogs as supplementary to traditional lectures or as e-learning materials [4]. There are two main types of educational blogs according to the role of the writer who can be either instructor or student. Those are:

- Instructor blogs: written by instructors, purpose is to use as an additional communication channel between the instructor and students, usually consist of course content and general instructions to students.
- Student blogs: written by students, this can be divided into two sub types as learning blog and project blogs [4].

Stakeholders in education field uses blogs to facilitate their various communication needs in favor of e-learning [4]. Hence the use of blogs in education field has drastically increased as pedagogical and technological innovations have caused in the next level of higher education [5, p. 20].

1.2 Content Extraction

An e-learning course can be developed using existing resources such as text documents, video, audio, examinations, assignments, surveys etc. [6]. Among those resources, blogs have very useful and up to date information as people used to share their experiences and new knowledge via blogs. Sharing those experiences and new knowledge is very helpful for students as well as co-workers in the workplace.

When developing an e-learning course using existing blog posts, extracting the knowledge content from the blog posts is vital due to the nature of blog posts. Those have no restrictions in the format. Also blog posts can be noisy. This leads to the need of special strategies for automatic content extraction from the blogs:

- “Information Extraction refers to the automatic extraction of structured information such as entities, relationships between entities, and attributes describing entities from unstructured sources” [7, p. 1].

1.2.1 General Information Extraction

Generally, information extraction discusses automatic methods for structuring the content extracted from Natural language text [8]. Extracting content from noisy and unstructured sources is challenging and researchers have engaged with this area for over two decades now [7]. As a result, information extraction systems are developed to identify the entities and relationships between those entities. The identified entities and relationships are stored in structured databases [8].

Information extraction is used in variety of applications of following categories [7]:

- Enterprise applications like news tracking, customer care, classified advertisements and data cleaning.
- Personal information management applications to organize documents, emails, projects and contacts.
- Scientific applications like as bioinformatics.

- Web applications such as citation databases, opinion databases, community websites, comparison shopping and structured web searches.

Information extraction systems fall into different categories such as [8], [9], [10];

- Manually Tuned Information Extraction Systems.
- Learning-Based Information Extraction Systems.
- Supervised Information Extraction.
- Unsupervised and Partially Supervised Information Extraction.

Text preprocessing is a vital task in information extraction systems. It is a set of activities applied to the input text because text data often comes with number formats, data formats, uncommon words, prepositions, articles, pro-nouns etc. Text preprocessing helps to eliminate those. Activities involved in Text preprocessing are [11]:

- Tokenization: which breaks text into words, phrases, symbols or other elements (tokens).
- Stop word removal: removes common words like ‘and’, ‘are’, ‘this’ etc.
- Stemming: extract the common representation from various forms of a word.

“Preprocessing is an important task and critical step in Text mining, Natural Language Processing (NLP) and information retrieval (IR). In the area of Text Mining, data preprocessing used for extracting interesting and non-trivial and knowledge from unstructured text data” [11].

1.2.2 Content Extraction from Blogs

When it comes to extracting content from blogs we need to consider that the whole web page might not be appropriate as web page usually contains clutter (or also called boilerplate) [12]. There could be various content included as clutter: navigation, contact info, navigation, pop-ups, unnecessary images, links. Those “extra” content might not be related to the topic of the web page [13]. Content extraction is widely used for supporting visually impaired and blind.

Information Retrieval (IR) and Natural Language Processing (NLP) algorithms will benefit from the content extraction as it helps to get only the extracted content without clutter and boilerplate coming from the actual web page [12]. In the web extraction domain, segmentation and text classification uses structural features in web pages. Mainly HTML tags (headers-h1..h6, paragraphs-p, divisions-div links, images and etc..) [14]. Study done in [15] discusses about different elements in a blog page. After main content extraction, there are three main sections: Header, Body and Footer. Along those 3 sections, the Body can be further broken down to: main content and sidebars. Following Figure 1.1 depicts how the blocks are arranged in a blog page.

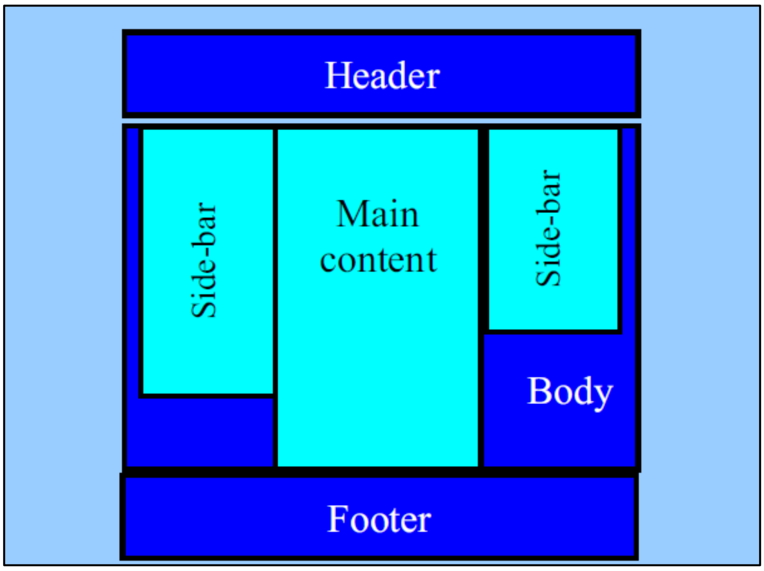


Figure 1.1: Macro-elements of a blog Web page [23]

1.3 e-Learning

The term “e-learning” is an umbrella term [16], hence there are different definitions for the term e-learning:

- Electronically mediated asynchronous and synchronous communication for the purpose of constructing and confirming knowledge [5, p. 21].
- The process of learning which is supported using ICT (e.g. the Internet, network, standalone computer, interactive whiteboard or portable device). Also used loosely to describe the actual content delivered on-screen, and the more general use of ICT to contribute to learning processes.

It has been coined in the mid 1990 with the betterment of World Wide Web. Primary goal of the e-learning is to create a group of individuals who can engage via internet and related technologies. Due to the interactive nature of the e-learning, it demonstrates very different characteristics from the traditional learning and teaching practices. “...e-learning is combination of technology and specially designed learning material…” [17] learning materials depend on the medium used, hence special design is needed. Based on the e-learning content involvement in learning process, e-learning is categorized into two main types: Complete online learning and blended learning.

Complete online learning solely depends on e-learning tools. All the activities like delivering course materials, discussions, assignments, examinations and all other evaluations are handled via e-learning tools. In contrary, Blended Learning use e-learning tools as supplementary, to make the conventional face-to-face methods more effective [17].

Gaur [17] describes the advantages of e-learning over traditional face-to-face learning as follows.

- Global Connectivity: there is no boundary such as university, school or class.
- Fast Access: endless e-resources are just a mouse-click away.
- Convergence of different mediums: use of text, graphics, audio, video and even virtual reality.
- Flexibility: in contrast with the conventional classroom-based learning, student can choose the time and place of learning.
- Quick Creation, upgradation and revision of course material: with e-learning, the learning material can be created, upgraded and revised faster than other conventional learning methods.
- Quick access to supporting material through hyperlink: student can view details about any term using links in the study material.
- Ability to serve large number of students at low cost: providing learning facilities to larger community costs more, e-learning can reduce the cost while facilitating large number of students.
- Distribution of quality material by virtual classes: using virtual classes, teachers can distribute same good quality learning materials even for the students in rural areas.

1.3.1 e-Learning theories

Learning content of an e-learning course is delivered via e-learning courses, often called as courseware. The word courseware is a combination of two words, course and software. When comparing the e-learning courseware, those bundled with all the materials that is required to complete a e-learning course where as traditional course only contains learning materials [18].

There is a whole philosophy behind designing e-learning courses and it is tightly connected with learning theories. Cognitive Load Theory (CLT) is a one of those theories connected to learning:

- Cognitive load theory is an instructional theory generated by this field of research. It describes learning structures in terms of an information processing system involving long term memory, which effectively stores all our knowledge and skills on a more-or-less permanent basis and working memory, which performs the intellectual tasks associated with consciousness. Information may only be stored in long term memory after first being attended to, and processed by, working memory. Working memory, however, is extremely limited in both capacity and duration [19].

The limitations described in Cognitive Load Theory (CLT) might impede the learning under some conditions. In such situations, the quality of instructional design is crucial. Since early 1980's, Cognitive Load Theory (CLT) has been used to develop successful instructional design strategies [20];

- The goal free effect.
- The worked example and problem completion effect.
- The split attention effects.
- The redundancy effects.
- The modality effects.

Instructional Design (ID) is the process of creating learning activities to transfer the knowledge efficiently and effectively. There are three main components in instructional design [21]:

- Understand how people learn.
- Construct learning activities based on how people learn.
- Measure the effectiveness of the learning activities.

As good instructional design can boost the learning efficiently by reducing the Cognitive Load. Following steps help to achieve a good instructional design [22];

- Keep it simple.
- Use different instructional techniques.
- Make learning “bite sized”.

There are many instructional systems design models, ADDIE (Analyze, Design, Develop, Implement and Evaluate), Rapid Prototyping, Dick and Carey, etc. Among those models, ADDIE model is the most commonly used one [23].

E-learning course design requires skills in different areas such as technology, media, academic and managerial. That leads to the existence of different roles in instructional design process [24, p. 22,23];

- Instructional designers (IDs) – responsible for designing specific e-learning materials and activities.
- Subject matter experts (SMEs) – contribute the knowledge and other related information for the course.
- Web developers and media editors – responsible for creating courseware and media.
- Course administrators, online facilitators and tutors – support learners.
- Technical support specialists – provide technical support for e-learning process.
- Human resources/Capacity development manager – responsible for all managerial activities.

Figure 1.2 shows the distribution of the responsibility among key roles in ADDIE process.

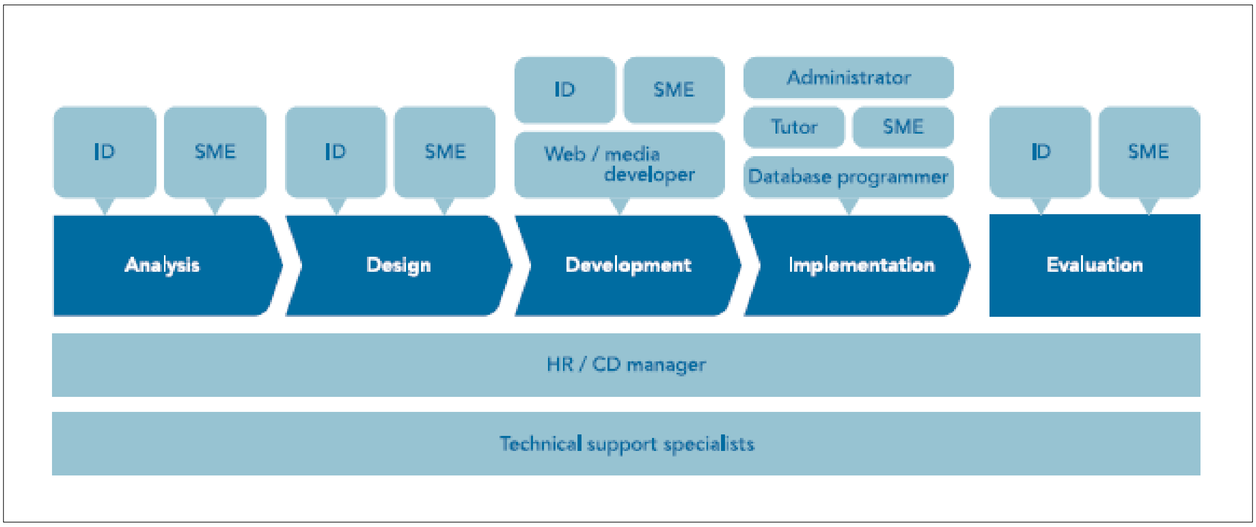


Figure 1.2: Areas of responsibility for key roles in the ADDIE process [13]

1.3.2 Learning Management Systems

A Learning Management System (LMS) can be seen as an "online learning hub" which facilitate efficient teaching and learning by providing indispensable set of features [25]. LMSs are called in different names [16];

- Learning Content Management System (LCMS)
- Learning Management System (LMS)
- Course Management System (CMS)
- Virtual Learning Environment (VLE)
- Managed Learning Environment (MLE)
- Learning Support System (LSS)

But the usage of is not consistent among those. Normally an LMS is used for managing online educational activities like: creating and delivering learning content, assessing the students and analyzing results, collaboration work in projects and tracking achievement of the students [25]. There are many commercial and open source software (OSS) available for providing LMS functionality. Moodle is such system that is worldwide popular [26].

When talking about the learning management industry, it's expected to expand the global LMS market from $5.22 billion to $17 billion by 2020 [25]. Studies reveal that in 2018 U.S. higher education market Blackboard [27] was leading with 31% of the market share, Canvas [23] was next with 30% of market share and Moodle [28] was with 18% of market share [29]. But when it comes to the global view, specially the Europe, Latin America, and Oceania regions the situation is totally different as Moodle is leading with more than 50% of the market share [30].

1.3.3 Moodle Integration

Moodle provides REST API to enable third party applications to access the Moodle database through web services. Learning content creation should be placed in Moodle section table according to the Moodle architecture. When it comes to a Moodle course, sections are the most important and essential parts as a section can contain forum, wiki, document, assignment, quiz or any SCORM component as well [31].

1.4 The Problem

Even though large number of researches show the contribution in both information extraction from text and e-learning content design, there are no any evident previous work carried out to generate e-learning content from extracted text from a blog automatically. This study tries to address that knowledge gap and attempts to find a solution for the problem of "Automatic creation of e-learning content using blog posts".

1.4.1 Knowledge Gap

There are no any automatic ways or tools to support extracting important learning content such as experiences and new knowledge from blogs and to provide those through an e-learning course. A large number of prior researches ([7], [32]–[34]) shows the popularity and the growth of knowledge in information extraction and text mining. Also, e-learning researches reveals the important factors to consider when designing e-earning content like instructional design. But there is no an evident connection between information extraction and e-learning content creation. This is identified as a gap in the knowledge and a problem to be solved.

1.4.2 Motivation

Normally, educational institutes which use e-learning, are equipped with defined processes, tools and human resources in their instructional design process [24]. When it comes to workplaces, in most cases they do not have such processes, tools or defined roles. Normally workplaces, especially small tech companies maintain blogs and even use public blogs (instead of having e-learning tools) to share new knowledge and experiences. If it is possible to automate the process of deriving e-learning course content from blogs, those companies also benefit form that as they can re-use their own blogs also, they can adopt e-learning tools without much burden and cost.

1.5 Aims and Objectives

The research reported in this thesis aimed to;

- Identify the best appropriate method to extract knowledge content from blog posts.
- Design and implement a model to pre-process and extract contents from blog posts.
- Design and implement a model to get the extracted knowledge content verified by human user.

- Select an appropriate technique to integrate knowledge content with existing LMSs.
- Integrate the verified knowledge content with existing LMS.

1.6 Assumptions and Delimitations

This section discusses assumptions and delimitations of this research.

1.6.1 Assumptions

- Extracting information from Blog Posts will be handled automatically. But those cannot be directly exported as knowledge content (as every information are not knowledge content). Hence there will be a filtering process to select “knowledge content” from “extracted information” with manual intervention. The present research assumes that the trainers will be the experts in organization, and they will use their own blogs to create learning content
- Content extraction and the e-learning course content generation will be done by the original author of the blog or Subject Matter Expert (SME), hence reviewing of the generated content is not required
- No copyright issues will be raised as this is intended to internal use of a company and proper acknowledgement will be given by adding the original source
- There are different types of the output formats available in some extractors, for the convenience of the evaluation and for the easiness of the e-content generation, single output format (HTML) is selected
- When generating e-learning content in LMS (as Course materials), the format of the course matters. In this research the only pre-defined formats will be considered (courses with Forums and Wikis)

1.6.2 Delimitations

- Blogs consist of audio, video, images and text. Among those, this study primarily focusses on the textual content
- Blog posts might be noisy and unstructured. Those can also range from a simple topic like “How to make a tea?” to a high-tech one. Due to that, selecting blog posts needs a considerable amount of manual intervention and filtering. Due to that, this research is based on only selected blog posts

- Selection of blog is restricted to a single discipline
- There are lot of LMSs available. Among those, “Moodle” is selected for implementing the integration part

1.7 Research Approach

In this study, the researcher tries to solve the problem in a workplace: how to extract e-learning content from blog and integrate them with an LMS. It is not evident that the exact same type of study conducted in available literature. But the similar studies conducted in [14], [35], [36], the researchers have employed both qualitative and quantitative methods in tandem as their research approach. Hence to solve the current problem, Pragmatic approach is used (mixed methods) and the research employs both qualitative and quantitative approaches such as surveys, experiments and pre-determined approaches (for usability) for data collection and analysis [37].

1.8 Chapter Outline

This report is organized into six chapters; Introduction, Literature Review, Methodology, Results and Evaluation, Discussion and Conclusion. This chapter discusses about the background of the research. It starts with an introduction to the area of study and then it explains the main subject areas: Information Extraction and e-leaning in detail. The study has been motivated by the current use of blogs in e-learning, the knowledge gap and the potential benefits for stakeholders. Aims and objectives of this research are in line with the problem being addressed. Assumptions and Delimitations describes the important consideration on research design and scoping the work to be carried out. At the end of the first chapter, research approach which employed is presented.

The second chapter reviews extant literature on Blogs and e-learning, content extraction from blogs, modeling e-learning content, Learning Management Systems (LMSs) and integrating external systems with LMSs.

The methodology adopted to conduct the study is discussed in the third chapter. The design overview, architecture of the proposed model, programing environment and testing and evaluation topics are included in this chapter.

Fourth chapter is dedicated to present and interpret the results of the data obtained through the experiments. This chapter discuses briefly about the test data used and the results obtained. Then those data will be analyzed.

Fifth chapter discusses the outcome from the analysis done in the chapter 5 and summarizes the key implications emerged.

Final chapter provides concluding remarks of the study. It includes a brief summary of the entire study, conclusion, real world implications of the findings and suggestions for further research.

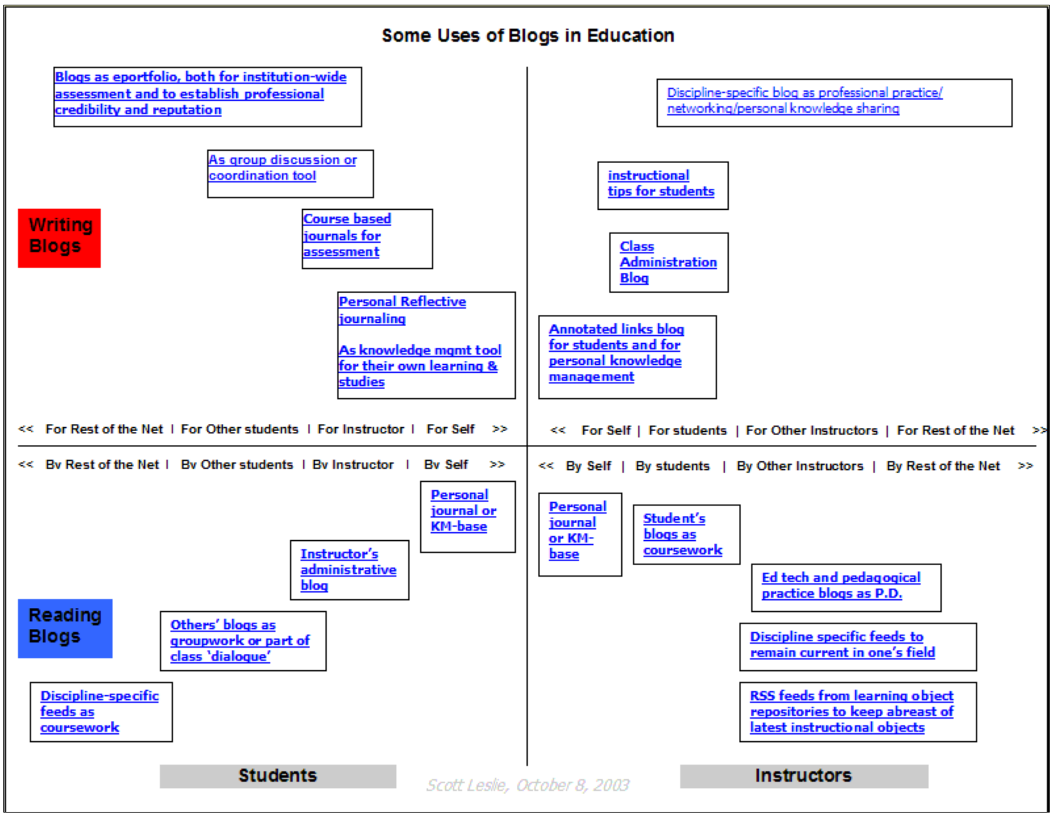
Chapter 2

Literature Review

This chapter reviews the literature in order to find an appropriate method to extract the knowledge content from blog posts and integrate those knowledge content with an existing Learning Management System (LMS). This chapter is organized as follows.

2.1 Blogs and e-learning

In e-learning blogs are being used to facilitate variety of communication needs [4]. Figure 2.1 shows matrix of some uses of blogs in education field.



Some Uses of Blogs in Education
Blogs as eportfolio, both for institution-wide assessment and to establish professional credibility and reputation
As group discussion or coordination tool
Course based journals for assessment
Personal Reflective journaling
As knowledge mgmt tool for their own learning & studies
Discipline-specific blog as professional practice/ networking/personal knowledge sharing
instructional tips for students
Class Administration Blog
Annotated links blog for students and for personal knowledge management
Writing Blogs
<< For Rest of the Net | For Other students | For Instructor | For Self >>
<< For Self | For students | For Other Instructors | For Rest of the Net >>
<< By Rest of the Net | By Other students | By Instructor | By Self >>
<< By Self | By students | By Other Instructors | By Rest of the Net >>
Personal journal or KM-base
Instructor's administrative blog
Others' blogs as groupwork or part of class 'dialogue'
Discipline-specific feeds as coursework
Reading Blogs
Personal journal or KM-base
Student's blogs as coursework
Ed tech and pedagogical practice blogs as P.D.
Discipline specific feeds to remain current in one's field
RSS feeds from learning object repositories to keep abreast of latest instructional objects
Students
Scott Leslie, October 8, 2003
Instructors

Figure 2.1: Some Uses of Blogs in Education [22]

Those uses are analyzed in two-dimensional space, users of blogs (instructors or students) vs intention of using blogs (writing or reading).

There are set of inherited advantages of blogs in e-learning due to the internet-based nature. Apart from those, blogs have their own advantages as follows [4] :

- Blogs are easy to setup and administrate in contrast to other technologies.
- Blogs makes easier to publish all types of resources (text, images, video, etc.) to the Web when compared to traditional web publishing.
- Blogs allow instant publishing with just one click: blogs are easy to create and maintain, as opposed to traditional web pages that are labor-intensive and require at least some web design knowledge (HTML, CSS, JavaScript).
- Blogs can be updated easily, from anywhere without having to worry about FTP connections, web authoring software, etc.
- Blogs could reach a large audience without losing information quality and allowing for different levels of detail. Blogs break the tradeoff between reach and richness of information.
- 24/7 (anytime, anywhere) access to information posted in blogs.
- No special blogging software is needed to create a weblog: some bloggers use plain HTML to create their blogs. Blogs can also be created with some scripts coded in Perl or using templates that makes blogging easier. However, blogging software allows a person to create and maintain a weblog without knowing HTML. Even more, bloggers may focus on content without the worries of periodically archiving, nor keeping accurate recording times. Still, complexity of blogs has greatly increased, therefore, blogging software becomes more necessary with time.
- Instructor does not need to periodically request the learning logs to the students.
- Other technologies can be applied jointly. For instance, using of Wikis as enablers for group writing and knowledge sharing. For example, building glossaries.

Blogging is no longer used just for sharing personal stuff and expert opinions [38], It has been recommended as a tool for supporting e-learning courses and already proven its success as a versatile tool in e-learning [39].

2.2 Content extraction from Blogs

An e-learning course can be developed using existing resources such as text documents, video, audio, examinations, assignments, surveys etc. [6]. Among those resources, blogs have very useful and up to date information as people used to share their experiences and new knowledge via blogs. Sharing those experiences and new knowledge is very helpful for students as well as co-workers in the workplace.

2.2.1 Generic text extraction

Extracting key terms in a text can be addressed using different approaches, the simplest approach is to use TFxIDF model (Term Frequency–Inverse Document Frequency) to select key words in a document [40]. But this method was generally led to poor results, hence new methods like supervised learning methods were discovered where a system is trained to recognize keywords in a text using lexical and syntactic features.

Text Mining technologies can be used to extract the information from unstructured textual data. The study conducted by [41] presents two examples of automatically extracting information using different Text Mining (TM) techniques: automated key-word association extraction, and prototypical document mining.

Study of Grineva et al. [42] depicts how to extract the key terms from noisy and multi-theme document. In their model, document is modeled as a graph of semantic relationships between terms of that document. Even though the state-of-the-art approaches in key terms extraction are based on statistical learning but those require hand-created training sets. In this study they present a new approach to key terms extraction using Wikipedia and that approach utilizes processed document rather than a hand-created training set. Their method consists of five steps as (1) candidate terms extraction; (2) word sense disambiguation; (3) building semantic graph; (4) discovering community structure of the semantic graph; and (5) selecting valuable communities. Further, they have performed an evaluation on web pages as it is important to automatically extract key terms from noisy documents.

Machine learning approaches have been proven their accuracy and usefulness when extracting content from text. But supervised training of extracting is costly. Study of Carlson et al. [10] present a method of achieving higher performance in information extraction using semi-supervised learning by coupling the simultaneous training of many extractions. The study focusses on a “bootstrapping” method for semi-supervised learning as such approaches have proven their performance in several areas like web page classification, named entity classification, parsing, and machine translation.

Study conducted in [9] addresses the problem of automatically populating forms and databases with unstructured electronic information. This study focusses on extracting contact information from various places such as signature of emails, on web pages, and on fax cover sheets using discriminatively trained context free grammar and conditional Markov chain models (CMM). According to the experimental results, a discriminatively trained context free grammar can more accurately extract contact information than a similar conditional Markov model.

Named Entity Extraction is a vital task in NLP, hence there are popular services for extracting, classifying and disambiguating named entities. In [43], Rizzo and Troncy present "NERD: A Framework for Unifying Named Entity Recognition and Disambiguation Extraction Tools", a framework with 10 popular named entity extractors available on the web.

"TagAssist: Automatic Tag Suggestion for Blog Posts" in [44] presents a system which is capable of suggesting tags for new blog posts by using existing tagged posts. The system takes a new post and finds other similar tagged posts and then recommends a tag set. In the stage of Tag Compression, system has two primary phase (1) Tag Normalization, in which token scrubbing is performed to trim white spaces and punctuations from each tag and (2) Compression Validation, which confirms the grouping from normalization phase has not grouped tags with different meanings under same normalized root. Tag Suggestion Engine (TSE) is used to suggest set of tags to the user based on normalized and compressed tag space.

In [8], Agichtein et al. present their "Snowball system" which is capable of extracting a relation from text by staring with only few initial "seed" tuples. Further, the system is also capable of discovering the rest of the tuples by analyzing the context where the "seed" tuples occur.

2.2.2 Text extraction from Blogs

Up to now we have discussed about generic text extraction, different algorithms and techniques used. When it comes to contents blogs (blogs are also web pages), those are semi-structured text documents presented in XML, HTML or XHTML. Hence it's lacking formatted structure of a document [45]. Before the actual extraction part the content of a web/blog page will be re-formatted: reproduce the entire web/blog page in a more convenient form by removing some HTML and removing some data components like images [12].

Extracting the core content from blog (or news sites) is very crucial for various types of applications like article narration, offline reading, generating article previews as well as generating content for visually impaired and blind [12]. Also this is a prerequisite for further processing the content for the tasks falls in the NLP methods such as sentiment analysis, content summarization and, text classification [46].

As also mentioned in the background of the study, boilerplate or the clutter removal from the original blog is a difficult problem. Main problem is to determine the which blocks of the text contribute to the core content of the blog as different blogs may use different mark-up (headers, div, link, iframe, images, etc.) and also the main content can be placed in in anywhere of the DOM tree. And most of the cases the DOM that contain the meat of the article (core content) is scattered and lot of boilerplate (banners, ads, comments, contact info, related articles, etc.) is added in between [46].

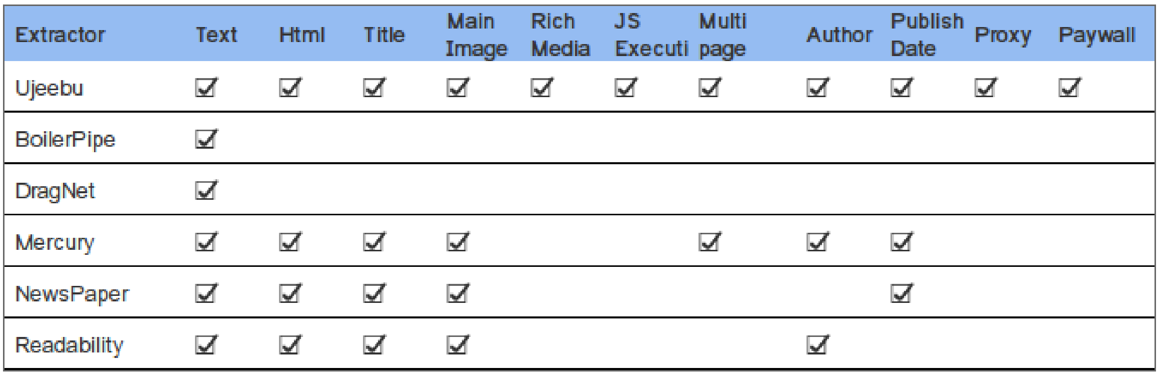
When it comes to extracting text and other data from a webpage there are two main sets of commonly used techniques: statistical and Machine Learning (ML). Most of the statistical methods use the methods based on computing heuristics (technique/approach to solve problems or self-discovery that a practical method is employed which might not be the optimal/perfect. In such cases, finding an optimal solution might be impossible or impractical. Using heuristic methods, the process can be speed up to finding a solution of a satisfactory level [47]) like, frequency of certain characters, distance from title, link density (number of words in a block that are linked versus the overall number of words in the block [48]), etc. [46].

Both techniques have their own pros and cons. Statistical method is less computationally intensive than the Machine Learning but most of the cases provides acceptable results whereas Machine Learning based methods have proven their success in complex cases like outliers. When it comes to Machine Learning, quality of the training data set matters. Some cases the two techniques were used in tandem as well [46].

Even though the above methods provide satisfactory results, those can also fail to provide results due to a totally different html structure that those extractors “have not seen earlier”. That means the available extractors (both statistical and ML based) might not work for all sorts of available sites [46]. One or the other extractor works better depending on the scenario. And most of the algorithms assume that the core content is textual [48].

There are lot of commercial and open-source extractors available and as in the different techniques used in content extraction (statistical and Machine Learning (ML)) there are different strong points in each extractor. If we look into some open source offerings there are libraries such as Readability[49], Mercury[50], Boilerpipe[51], Dragnet[52], NewsPaper[53] and Goose[54]. Among those libraries most of them are written in python or ported to python [46]. This study will be based on few selected open-source libraries and those will be discussed further in the Chapter 3.

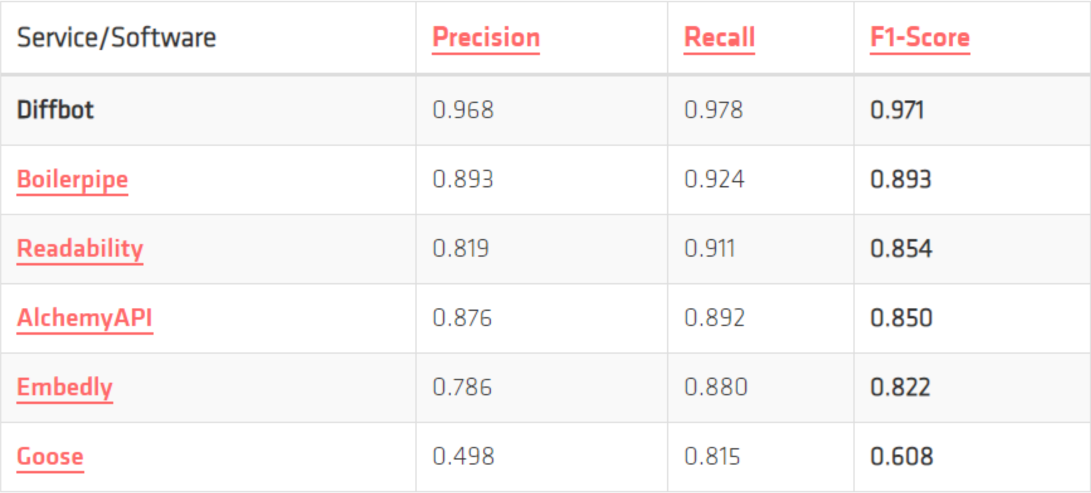
Here you can find two different analysis done for comparing different APIs. Table 2.1 shows a comparison done among 6 different APIs based on their capabilities.



Extractor	Text	Html	Title	Main Image	Rich Media	JS Executi	Multi page	Author	Publish Date	Proxy	Paywall
Ujeebu	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
BoilerPipe	☑										
DragNet	☑										
Mercury	☑	☑	☑	☑			☑	☑	☑		
NewsPaper	☑	☑	☑	☑					☑		
Readability	☑	☑	☑	☑				☑			

Table 2.1: Ujeebu extraction API vs. Open Source [43]

Table 2.2 depicts the results of another analysis done based on the Precision, Recall and the F1- Score [55].



Service/Software	Precision	Recall	F1-Score
Diffbot	0.968	0.978	**0.971**
Boilerpipe	0.893	0.924	**0.893**
Readability	0.819	0.911	**0.854**
AlchemyAPI	0.876	0.892	**0.850**
Embedly	0.786	0.880	**0.822**
Goose	0.498	0.815	**0.608**

Table 2.2: Comparing text Extraction Methods [52]

Precision: the fraction of retrieved text that was correct. For example, if a method returns only one sentence of a ten-paragraph article, it will have a precision of 1.0.

Recall: the fraction of all correct text that was returned. However, a method that returns all the content on a page (not only including the full article text, but the header/footer/advertising content/etc.) can achieve a recall of 1.0.

F1-score: a weighted average of the precision and recall, taking both into account. Best score is 1.0 and worst is 0.0.

2.3 Creating e-learning content

Web based course plays a major role in e-learning to carry out teaching activities, according to the study conducted in [56] presents difficulties that teachers from non-computer profession might come across. As consequences, the developed low-quality e-learning courses waste time, money and resources of developers nevertheless, they also lead to student to gain less and teachers lose their interest. With the presented approach, e-learning courses also include Instructional Design Philosophy and guides teachers to develop courses.

As e-learning becomes an essential part in every learning institute, authors and publishers face difficulties when guaranteeing accessibility of learning objects. Non-annotated objects are major problem since it hinders system's ability to recommend useful information, nevertheless, prevent learners form finding new information. The problem is commonly known as "cold-start" problem [57]. The study carried out in [58] addresses this problem using a state-of-the-art method for automatic tag annotation, "TaggingLDA": Latent Dirichlet Allocation probabilistic topic model based approach. Result shows automatically generated tags were preferred 35% more than the original tags by authors. Those generated tags were 17.7% more relevant for users as well. This study implies that the automatic tag suggestion can facilitate information access in e-learning hence solves the "cold-start" problem.

Leaning objects are small instructional components which can be defined as:

- Learning Objects are defined here as any entity, digital or non-digital, which can be used, re-used or referenced during technology-supported learning. Examples of technology-supported learning include computer-based training systems, interactive learning environments, intelligent computer-aided instruction systems, distance learning systems, and collaborative learning environments. Examples of Learning Objects include multimedia content, instructional content, learning objectives, instructional software and software tools, and persons, organizations, or events referenced during technology supported learning [59].

2.4 Integration with LMSs

Lecture recording systems are essential components in e-learning. But current lecture recording systems carry many problems. “Flash based lecture recording system” which has integrated with LMS as presented in [60] tries to overcome those problems due to its unique features such as (1) classroom recording using a simple equipment setup with minimum burden to lecturers, (b) editing the timeline of the lecture record and bookmarking using metadata, and (c) integration with the assets in LMS. When it comes to the implementation of the above-mentioned lecture recording system, Anh et al. present the prototype with UMELMS (Ubiquitous Multimedia Enhanced Learning Management System).

Figure 2.2 shows the complete implementation of the system.

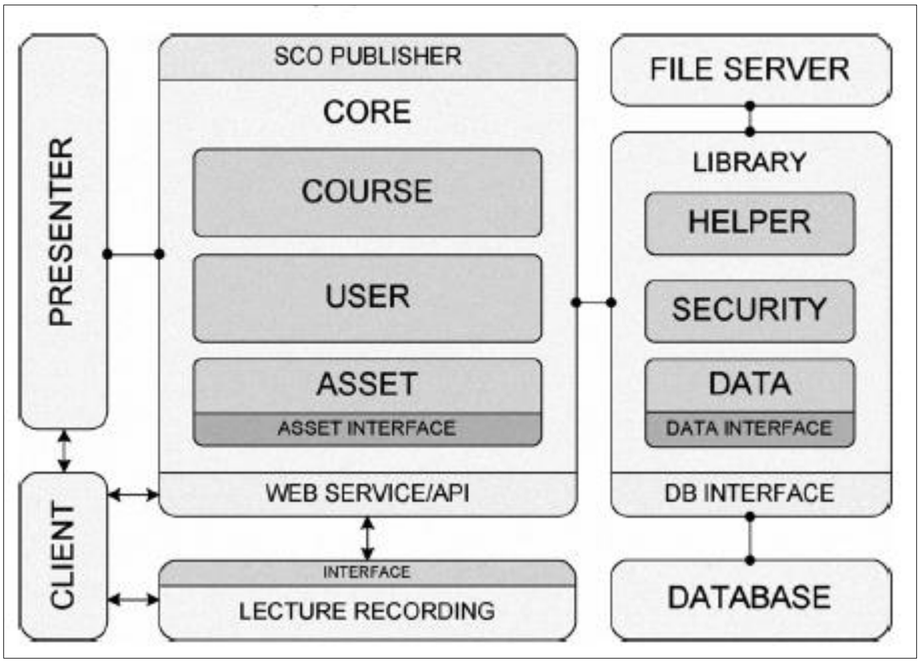


Figure 2.2: Complete implementation of the system [32]

Mobile learning/m-learning applications adds next level of iniquitousness to the learning process. Guerrero et al. [61] give two main studies carried out in the area of integrating the generated e-learning content with an existing LMS. They present a new architecture which allows LMS and Mobile Applications to work interoperable and the interoperability bet ween LMS and Mobile Applications also two-way. Figure 1.3 depicts the Architecture to integrate LMS with external mobile applications.

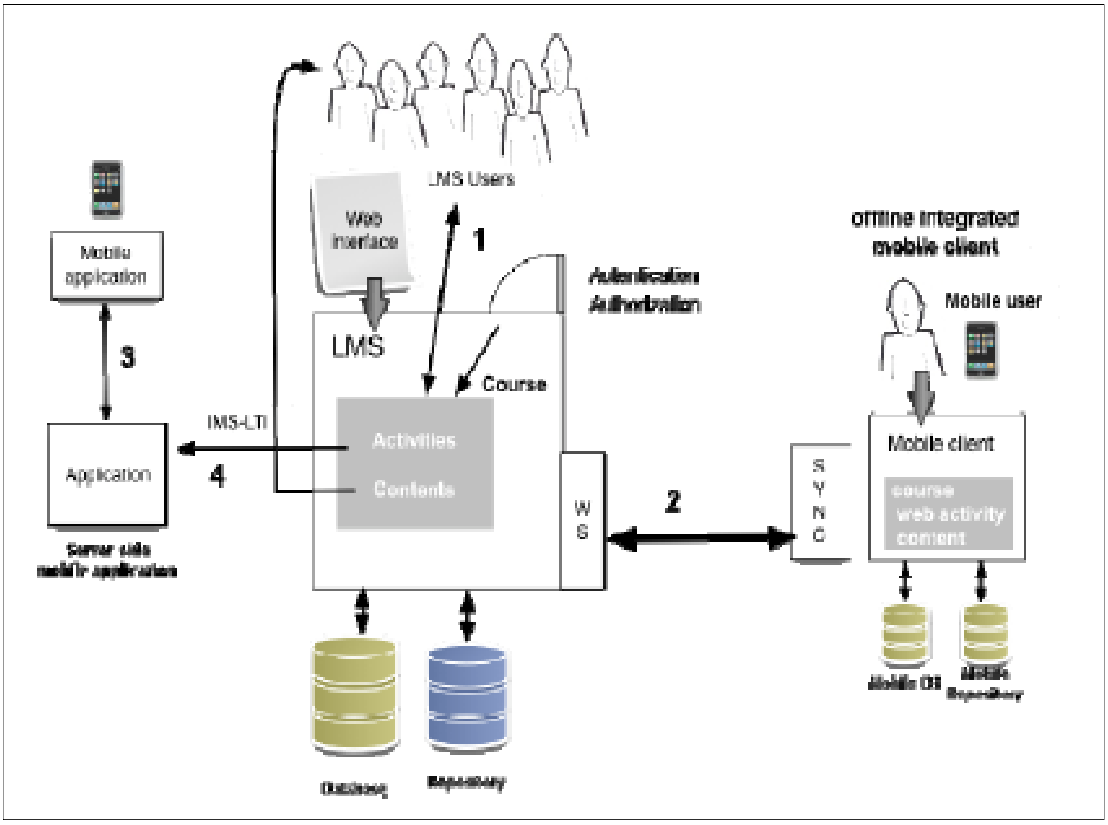


Figure 2.3: Architecture to integrate LMS with external mobile applications [58]

But in above mentioned two LMS integrations related studies, they have not focused on the Instructional Design process which is a key area in e-learning.

When it comes to the implementations, authors present two m-learning projects: (1) Moodbile: The Moodle Mobile Client, customized LMS which can be accessed via web browsers on mobile devices. (2) CLAYMobile.

Above mentioned studies done in the area of LMS integrations are mainly focused on integrating with mobile applications. However, in [62] Kautsar et.al present a tool for lecturers to upload their e-Learning content to LMS automatically. Normally lecturers/Content Designers create learning contents via accessing the Web interface of the Moodle. In order to do that an active internet connection is a must. But when it comes to a situation where the person who is developing e-Learning content is unable to access the Moodle server, this study proposes a method to do that outside of the Moodle server [62]. In order to do that they have used Moodle Web Service and access the Moodle via a REST Function call. In order to do that proper access mechanism is used via a token and it is passed in the REST function call as a parameter [63]. Moodle stores the content in Moodle tables: “mdl_course”, “mdl_course_section” are two tables used for storing course data and section data respectively. Section data could be Wikis, Forums, SCORM components etc. [62].

Below Figure 2.4 depicts how the course and section tables are organized in Moodle database.

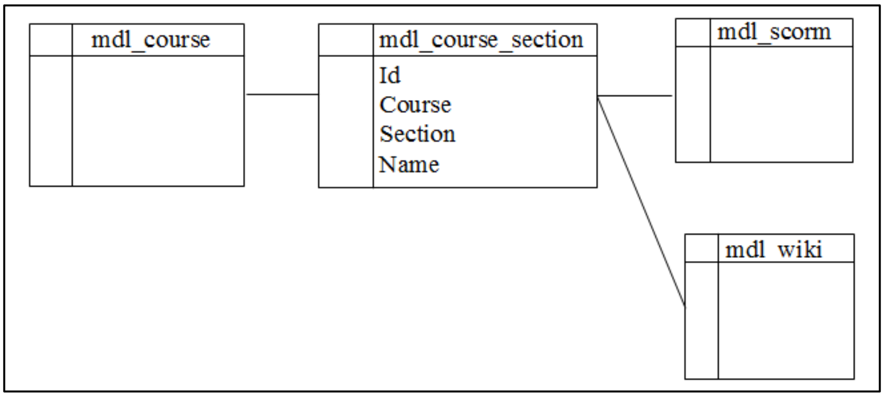


Figure 2.4: Moodle course tables relationships [59]

To upload the content, the application developed in [62] uses another REST function call. But there is no function available for creating sections, wikis and SCORM components. Hence, they have done that part via a direct database call.

The study done in [31] has addressed a similar problem but in that study Kautsar et al. were able to create the missing Web Service in their previous study. In that study they have done two plugins, one is to create learning activities and the other one to create a Web Service to write data into that table. As per the convention in Moodle plugin development [64], they have create a new table as “mdl_xlmsmemod”. Relationship between the plugin and core Moodle tables can be shown as below Figure 2.5.

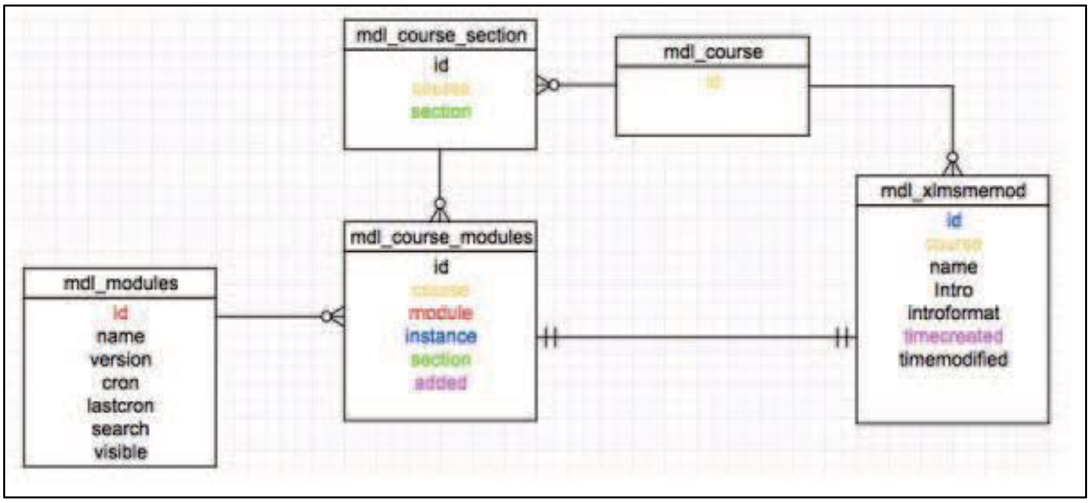


Figure 2.5: Relation Table Developed Plugin and Core Moodle Tables [31]

As mentioned earlier Kautsar et al. has developed another plugin "xlmsmemodws" to access the "xlmsmemod" plugin as "xlmsmemod" plugin has the Moodle database access [31]. Anyway, a web service function is not still available in Moodle REST API for creating new sections under a course. But there is a function available (core_course_edit_section) to edit the visibility of a section and also to delete it [65].

2.5 Theoretical Perspective Inspired Design of the Research

Blogs, e-learning and content extraction are the buzz words in the present world. There are lot of blogs, e-learning tools as well as text extraction methods. Content extraction from webpages, blogs is heavily studied area as there are lot of applications like sentiment analysis, content summarization and text classification. Blogs are also webpages and they also serve as learning materials. Existing e-learning tools also support integration with other systems. This study tries to combine all those aspects and presents a model to incorporate learning content from a blog with an e-learning system as an integration.

2.6 Chapter Summary

Purpose of this chapter was to review the research literature to identify research gaps and to highlight what needs to be researched in this study. The chapter commenced with a review of the use of blogs in e-learning and connection between blogs and e-learning. Next section described recent studies conducted in the area of extracting content from blogs. Then the chapter discussed modelling e-learning content and learning objects. Finally, the chapter described the methods used to integrate with LMSs.

Chapter 3

Methodology

This chapter discusses about the methodology adopted to conduct the current study. In this study, there are two main parts: content extraction from the blogs and creating e-learning content in LMS. The discussion begins with the design overview to show the high-level flow of the activities to be carried out, the proposed model, implementation details of the proposed model. And finally discusses about the dataset and evaluation criteria.

3.1 Design Overview

Different text extraction methods were discussed in the literature review chapter and among those methods, the appropriate methods were selected based on the previous studies conducted.

Below Figure 3.1 depicts how the flow of high-level activities carried out in this research.

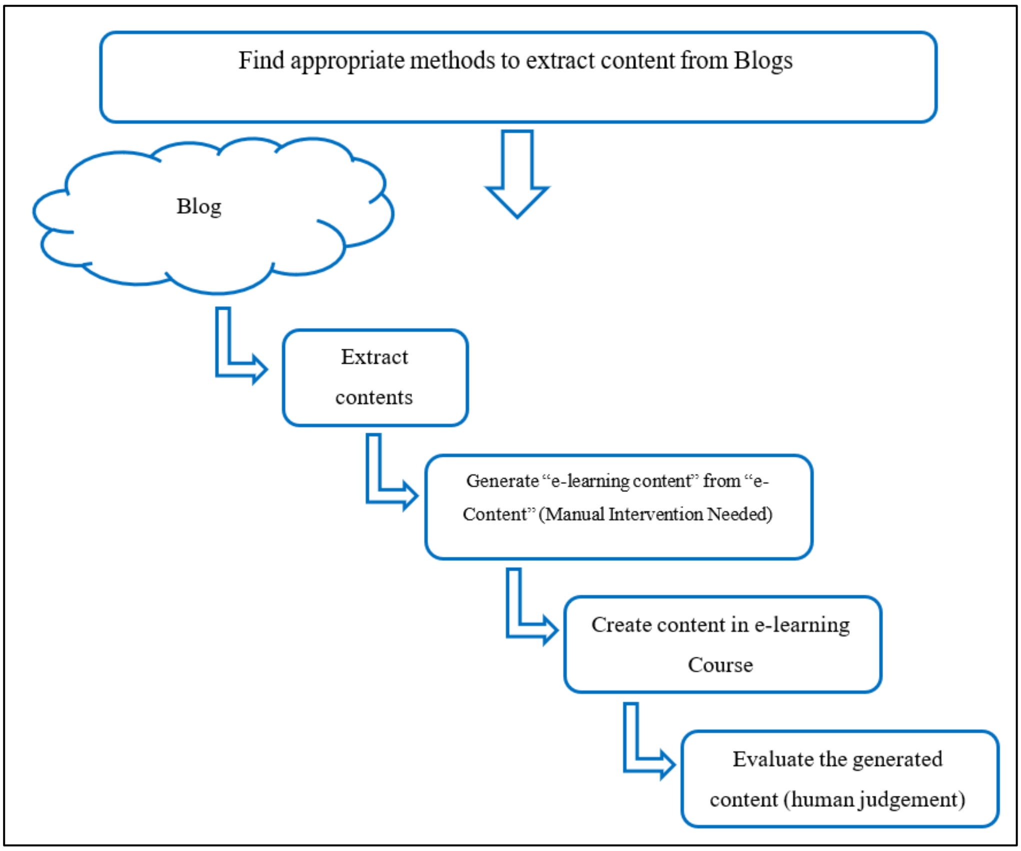

Find appropriate methods to extract content from Blogs
Blog
Extract contents
Generate "e-learning content" from "e-Content" (Manual Intervention Needed)
Create content in e-learning Course
Evaluate the generated content (human judgement)

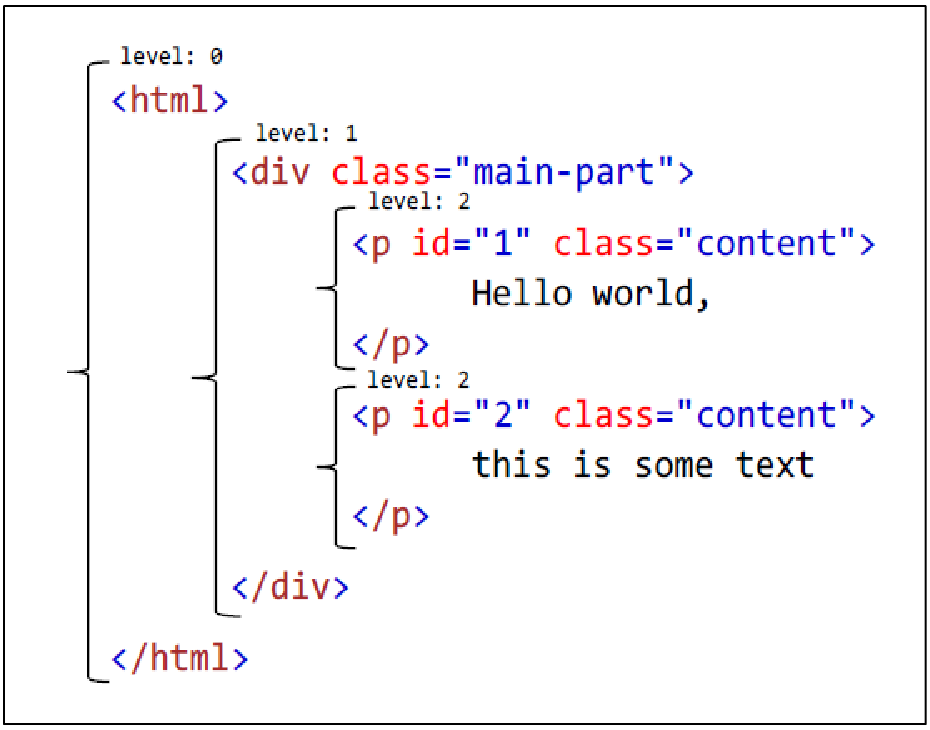
Figure 3.1: Research Design

Then the selected extraction methods were employed in the model to extract the content from Blogs. In the meantime, a separate survey was conducted to get inputs from e-Learning Instructional Designers to get their opinion on the user requirements of such model to extract the content from blogs and automatically integrate them with an LMS. Finally, the developed model was run against the selected blogs and the created e-Learning content was evaluated using human judgement.

3.1.1 Extracting Content

3.1.1.1 General layout of a HTML page

Generally, web and blog pages are written in Hypertext Markup Language (HTML). The layout of a basic HTML page can be shown as below Figure 3.2.



```
<html>
    <div class="main-part">
        <p id="1" class="content">
            Hello world,
        </p>
        <p id="2" class="content">
            this is some text
        </p>
    </div>
</html>
```

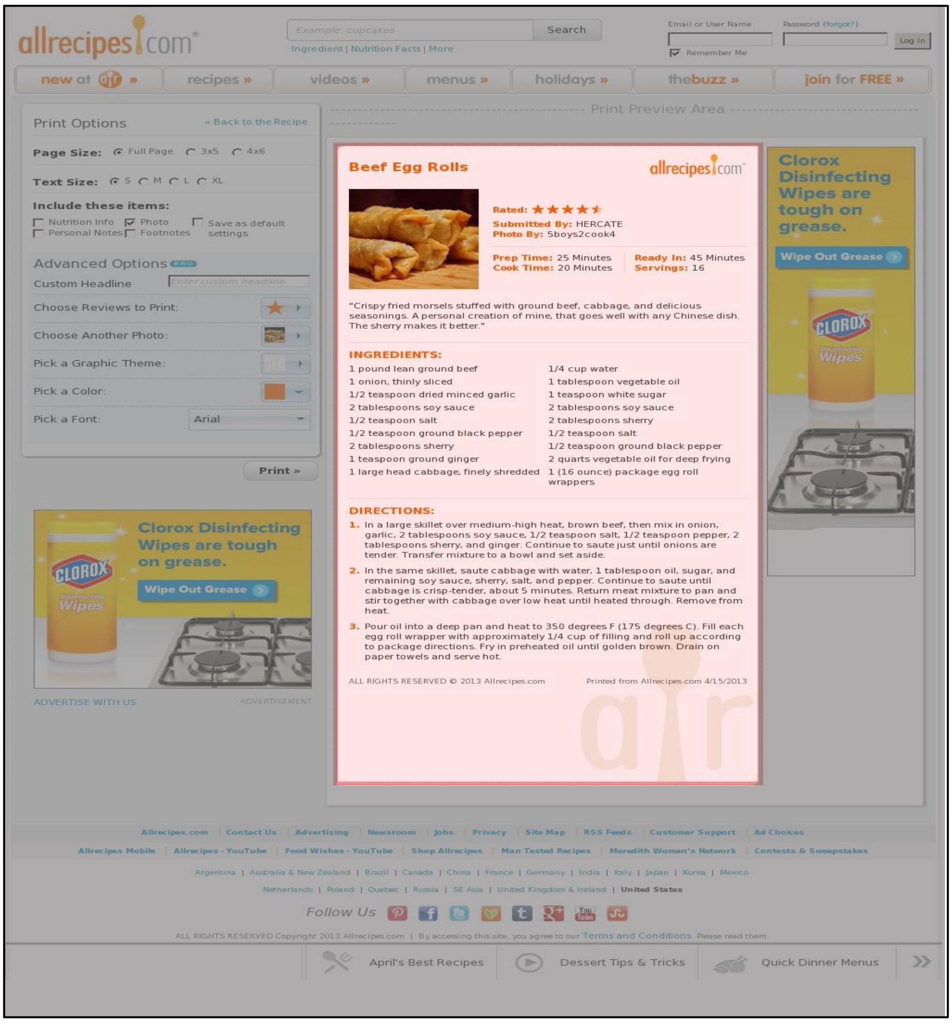
Figure 3.2: Example of an HTML document [15]

The whole HTML page can be considered as element. Under that element, there are one or more child elements. Tags in a HTML document are enclosed with angle brackets, < >. The <html> tag is the root of the document and it is in level 0. Then the next level is <div> element which has the level 1. Text in a HTML document is denoted with <p> tag, which stands for paragraph. Also, those text can be formatted with another tags <strong>, <u>, <em>, etc. [66].

3.1.1.2 Extracting text from a HTML page

Finding HTML tags in a web page seems to be a simple task. But when it comes to extracting the main content from the extracted HTML is not simple [66]. Boilerplate or the clutter removal from the original blog is a difficult problem. Most of the cases the DOM that contain the core of the article (core content) is scattered and lot of boilerplate (banners, ads, comments, contact info, related articles, etc…) is added in between [46].

Figure 3.3 shows how a typical extraction of the main content from a web page with a lot of clutter. The highlighted area is identified as the main content (meat) of the web page [67].


allrecipes.com
Example: cupcakes
Search
Ingredient | Nutrition Facts | More
Email or User Name
Password (forgot?)
Log In
☑ Remember Me
new at ar »
recipes »
videos »
menus »
holidays »
thebuzz »
join for FREE »
« Back to the Recipe
Print »
Clorox Disinfecting Wipes are tough on grease.
Wipe Out Grease
ADVERTISE WITH US
ADVERTISEMENT
Print Preview Area
Allrecipes.com | Contact Us | Advertising | Newsroom | Jobs | Privacy | Site Map | RSS Feeds | Customer Support | Ad Choices
Allrecipes Mobile | Allrecipes - YouTube | Food Wishes - YouTube | Shop Allrecipes | Man Tested Recipes | Meredith Women's Network | Contests & Sweepstakes
Argentina | Australia & New Zealand | Brazil | Canada | China | France | Germany | India | Italy | Japan | Korea | Mexico
Netherlands | Poland | Quebec | Russia | SE Asia | United Kingdom & Ireland | United States
Follow Us
April's Best Recipes
Dessert Tips & Tricks
Quick Dinner Menus

Print Options

Page Size: ● Full Page ○ 3x5 ○ 4x6

Text Size: ● S ○ M ○ L ○ XL

Include these items:

□ Nutrition Info ☑ Photo □ Save as default settings
□ Personal Notes □ Footnotes

Advanced Options PRO

Custom Headline: Enter custom headline

Choose Reviews to Print:

Choose Another Photo:

Pick a Graphic Theme:

Pick a Color:

Pick a Font: Arial

Beef Egg Rolls

Rated: ★★★★½

Submitted By: HERCATE
Photo By: 5boys2cook4

Prep Time: 25 Minutes
Cook Time: 20 Minutes

Ready In: 45 Minutes
Servings: 16

"Crispy fried morsels stuffed with ground beef, cabbage, and delicious seasonings. A personal creation of mine, that goes well with any Chinese dish. The sherry makes it better."

INGREDIENTS:

1 pound lean ground beef
1 onion, thinly sliced
1/2 teaspoon dried minced garlic
2 tablespoons soy sauce
1/2 teaspoon salt
1/2 teaspoon ground black pepper
2 tablespoons sherry
1 teaspoon ground ginger
1 large head cabbage, finely shredded
1/4 cup water
1 tablespoon vegetable oil
1 teaspoon white sugar
2 tablespoons soy sauce
2 tablespoons sherry
1/2 teaspoon salt
1/2 teaspoon ground black pepper
2 quarts vegetable oil for deep frying
1 (16 ounce) package egg roll wrappers

DIRECTIONS:

1. In a large skillet over medium-high heat, brown beef, then mix in onion, garlic, 2 tablespoons soy sauce, 1/2 teaspoon salt, 1/2 teaspoon pepper, 2 tablespoons sherry, and ginger. Continue to saute just until onions are tender. Transfer mixture to a bowl and set aside.
2. In the same skillet, saute cabbage with water, 1 tablespoon oil, sugar, and remaining soy sauce, sherry, salt, and pepper. Continue to saute until cabbage is crisp-tender, about 5 minutes. Return meat mixture to pan and stir together with cabbage over low heat until heated through. Remove from heat.
3. Pour oil into a deep pan and heat to 350 degrees F (175 degrees C). Fill each egg roll wrapper with approximately 1/4 cup of filling and roll up according to package directions. Fry in preheated oil until golden brown. Drain on paper towels and serve hot.

ALL RIGHTS RESERVED © 2013 Allrecipes.com Printed from Allrecipes.com 4/15/2013

ALL RIGHTS RESERVED Copyright 2013 Allrecipes.com | By accessing this site, you agree to our Terms and Conditions. Please read them.

Figure 3.3: An example of extracting the main content from a webpage [64]

3.1.1.3 Text Extractor Libraries

There are a lot of proprietary and open source text extractors available, based on the studies done in [14] [46], [68] following text extractors were short listed.

- Readability [49]
  - Based on the oldest and most used algorithms for text extraction. This uses heuristic based algorithm. This was ported to different languages like java, JavaScript, Node.js, Objective C, PHP, Python, etc [69]

- Boilerpipe [51]
  - Java based extractor and again uses heuristic based algorithm like in Readability. This has also been ported to different languages including python [46]

- NewsPaper [53]
  - Python based extractor and based on a another library called Goose [54], Goose was written in Scala [70]. Unlike Readability and Boilerpipe, Newspaper has in built support for extracting some other data like published date, author, key words, article summary, main image and video URLs [46]

Below Table 3.1 compares the capabilities of the selected extractor libraries.

Extractor Library	**Title Extraction**	**Format of the Extracted Content**	**Other Features**
Readability	Yes	HTML	
Boilerpipe	Yes	HTML/HTML Fragment/Text/JSON/debug*	different extractors** tuned for different purposes
NewsPaper	Yes	HTML/Text	extract “author”, “published date”, “key words”, “article summary”, “main

			image” and “video URLs”

Table 3.1: Comparison of the capabilities of Readability, Boilerpipe and NewsPaper extractor libraries [51], [53], [69]

* HTML: output the extracted main content as HTML; HTMLFragment: output only the HTML fragment regarded as main content; TEXT: out put the main content as plain text; JSON: out put the main content as json; debug: out put the debug information which are used for understanding the internal representation of the extracted content in boilerpipe.

** **ArticleExtractor**: full-text extractor tuned for news articles; **DefaultExtractor**: generic full-text extractor; **LargestContentExtractor**: like DefaultExtractor but keeps only the largest block of content. Recommended for no-article style pages with only one main block. **KeepEverythingExtractor**: extract the whole page as content.

In the Literature Review chapter, another two extractors called Mercury [50] and Goose [54] were mentioned. But those two were not taken into consideration as Mercury was based on Readability and NewsPaper was based on Goose [46].

3.1.1.4 Selection of Extractor Libraries

As per the previous studies and analysis done in [14], [46], [55], [68] the shortlisted extractors were selected. As mentioned in the section 1.6, the extractors with HTML support for output are selected to build the model. As shown in above Table 3.1, all three extractors can output the extracted content as HTML. Hence those three extractors were used in the proposed.

3.1.1.5 Selection of Blog posts

In the literature, researchers use different data sets to evaluate their models done for extracting content from webpages. Study done in [13] employs 140 webpages from popular web sites in Yahoo search engine. Some of them uses manually annotated dataset which consists of web pages downloaded from the internet [15]. Meanwhile [71] uses 20,000 web pages for the evaluation.

In this study, the researcher tries to solve the problem in a workplace: how to extract e-learning content from blog and integrate them with an LMS. Hence the selection of the blogs is also done in the same context and blog posts were selected from internal company blog hosted in [72].

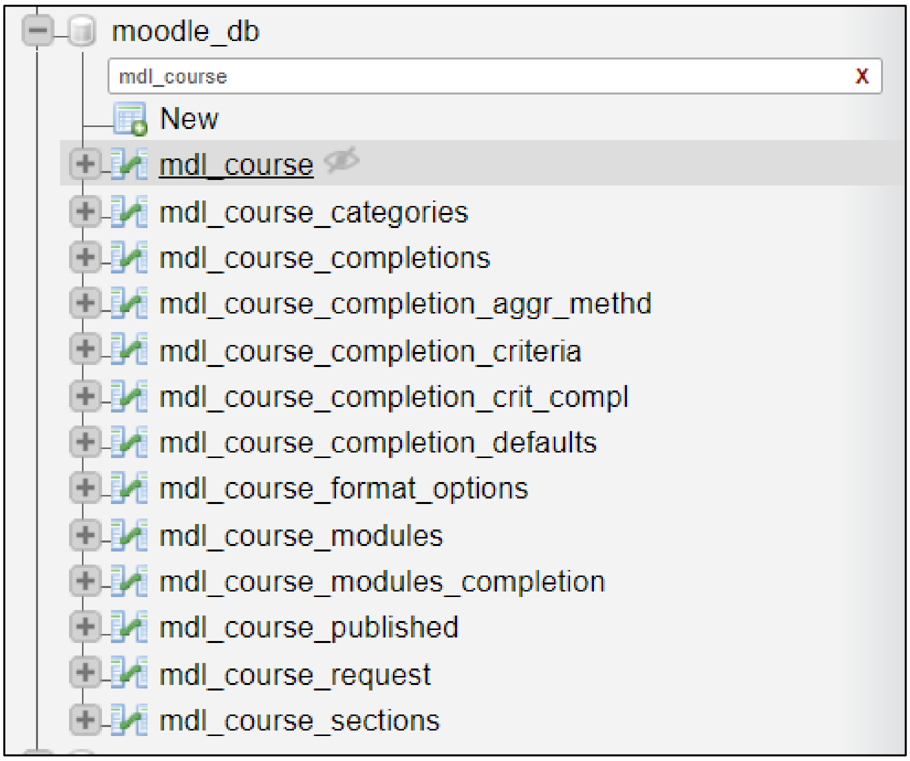
3.1.2 Integrate with LMS

3.1.2.1 Selection of LMS

Moodle is a free and open-source LMS, which is written in PHP. According to the studies done in [26, p. 6], [29], [30], it is evident that the Moodle is the worldwide popular LMS. Specially in regions like Europe, Latin America and Oceania, Moodle is leading with more than 50% of the market share. Also Moodle provides comprehensive set of REST API Functions [65] and that is another reason for selecting Moodle as the LMS for the integration part.

3.1.2.2 Integration with Moodle

When integrating extracted learning content with the Moodle, first we need to identify how the database structure is handled in Moodle. According to [31], learning content in Moodle stored in three kinds of tables. Courses are stored in Moodle course table: "mdl_course", course



mdl_course
mdl_course_categories
mdl_course_completions
mdl_course_completion_aggr_methd
mdl_course_completion_criteria
mdl_course_completion_crit_compl
mdl_course_completion_defaults
mdl_course_format_options
mdl_course_modules
mdl_course_modules_completion
mdl_course_published
mdl_course_request
mdl_course_sections

Figure 3.4: Moodle course tables

sections are stored in “mdl_course_sections” table and the learning activities are stored in individual tables and linked to a Moodle course section. There are 12 kinds of activity formats and each stored in a separate table. “mdl_scorm”, “mdl_forum”, “mdl_wiki” and “mdl_quiz” are some example of such tables [31]. Figure 3.4 shows the list of available Moodle course tables.

Below Figure 3.5 depicts the architecture of Moodle learning content under a course.

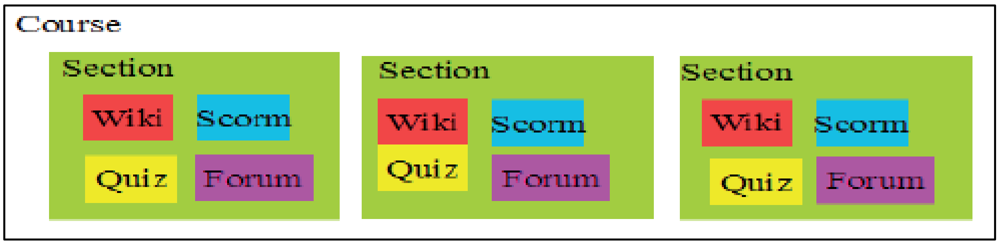


Figure 3.5: Architecture of Moodle learning content under a course [31]

Complete database schema information of the Moodle database can be found in [73].

Instead of accessing Moodle database directly (using username and password via an SQL client or another client application), the same can be done via Moodle REST API functions [31], [65]. In order to access the Moodle database via REST API functions, there are some configurations to be done in Moodle as an administrator. Complete set of steps for enabling the REST API Function calls via an external web service can be found in [63], [74].

After properly setting up the Web Service, it is possible to test the integration. For that we can use any of available Web Service functions [65]. In order to do that we can use any API client that support REST API calls. In this study, Postman is used for testing the REST API calls. More information about Postman can be found in [75]. Sample REST API call used for creating a new Course in Moodle is shown in below Table 3.2.

URI
http://localhost/moodle/webservice/rest/server.php?wstoken=38091c6777f2ac297655b12c1fba35e7&wsfunction=core_course_create_courses&moodlewsrestformat=json&courses[0][fullname]=DEMOCourse2&courses[0][shortname]=DEMO2&courses[0][categoryid]=13&courses[0][idnumber]=2

Table 3.2: Sample URI to create a course in Moodle via REST Function call

In the URI shown in Table 3.2, there are 4 main parts:

1. Moodle server address

2. Access token provided by Moodle when setting up external web service [63], [74]
3. REST Function name
4. Data as parameters

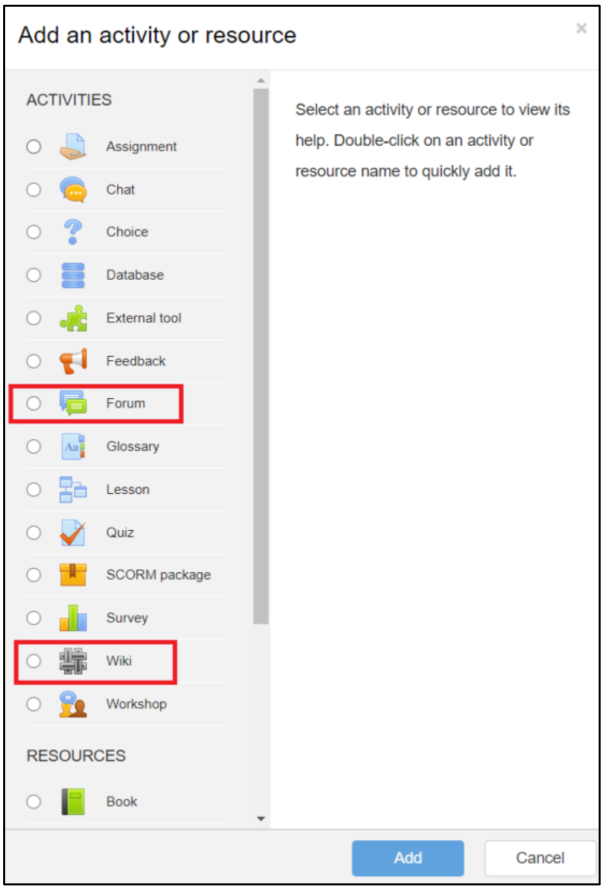
With the available Moodle REST API Functions [65], the extracted content can be integrated to e-learning course via REST API Functions as:

- Discussion in a Forum (function: “mod_forum_add_discussion”)
- Post under a discussion in a Forum (function: “mod_forum_add_discussion_post”)
- Page in a Wiki (function: “mod_wiki_new_page”)

Comprehensive description of the above REST API functions can be found in [65].

When it comes to Forum discussion, discussion post and wiki page, all three can show HTML page as the content. For the simplicity, and as it is a commonly used method of sharing the knowledge inside the company (in the Blog itself discussions can be initiated), the extracted content from the blogs are integrated with the LMS as Forum discussion and eventually used for evaluating the proposed model as well.

Figure 3.6 shows how to add a new activity to a course. As highlighted, Form and Wiki are two activities that we can access via Moodle REST API functions.



Select an activity or resource to view its help. Double-click on an activity or resource name to quickly add it.

Figure 3.6: Adding activity to course in Moodle

Still there are some REST API Functions missing in Moodle REST API [31]. For an example, there are no REST API Functions available for:

- Create/update Section in a course
- Create/update Forum in a course
- Create/update SCORM content in a course
- Create/update a Page in a course

Creating new Moodle plugins might be another solution as Kautsar et al. done in [31]. But this is not taken into the scope of this study due to the time constraints.

3.1.3 Usability of the Proposed Model

There is another aspect of usability of the proposed model in this study. When considering the typical e-Learning process, Instructional Designers are the ones who are responsible for designing specific e-learning materials and activities [24, p. 22,23]. Hence a survey has been conducted to find out the potential requirements of a such system. A questionnaire has been distributed among five selected Instructional Designers in e-Learning center of University of Colombo School of Computing (UCSC) after getting their consent. Consent letter and the questionnaire can be found in Appendix A and Appendix B respectively.

After getting the responses of the participants, the survey results were analyzed and following comments and suggestions were identified.

1. Getting the extracted content as HTML rather than text (4/5 participants)
2. Setting title from the blog as default title of the course content (3/5 participants)
   - Possibility to edit the title as well (1/3 participants)
3. Including original images from the blog (4/5 participants)
   - Possibility of adding custom images as well (1/4 participants)
4. Other requirements (open ended question) – (3/5 participants)
   - Ability to edit extracted content – (2/3 participants)
   - Ability to put the original source as a reference automatically (considering copyright issue) – (2/3 participants)
   - Ability to get the extracted blog content verified by a Subject Matter Expert (SME) via an e-mail or something like that – (1/3 participants)
   - Ability to extract content from multiple blogs – (1/3 participants)
   - Ability to keep history of the blogs that have been used for generating e-learning content for the future reference – (1/3 participants)
   - Process should be easy to understand for the person who is doing the extraction and course content generation – (1/3 participants)

Among those requirements, all others were implemented in the prototype used for analyzing the proposed model except last four points under Other requirements. They were not implemented due to the time constraints and kept as future improvements to the model.

3.2 Proposed Model

Proposed model focuses on extracting content from a blog post as HTML using selected extractor libraries (Readability, Boilerpipe and NewsPaper), generating e-Learning content from the extracted HTML (with manual intervention), transferring the generated e-Learning content to Moodle via REST API call utilizing Moodle REST API service functions to create forum discussion under already created form.

This process is repeated for all the selected blogs and per selected extractor library. Finally, the generated forum discussions are evaluated using pre-defined Evaluation Criteria, which is presented under section 3.4. In order to present and evaluate the proposed model, a prototype is developed.

When talking about the technical implementation of the prototype, it can be divided in to three main components.

1. Content Extractor Service
2. Learning Content Generation Component
3. Learning Content Integration Component

Following Figure 3.7 shows the interaction of the above three components and inputs/outputs of each component.

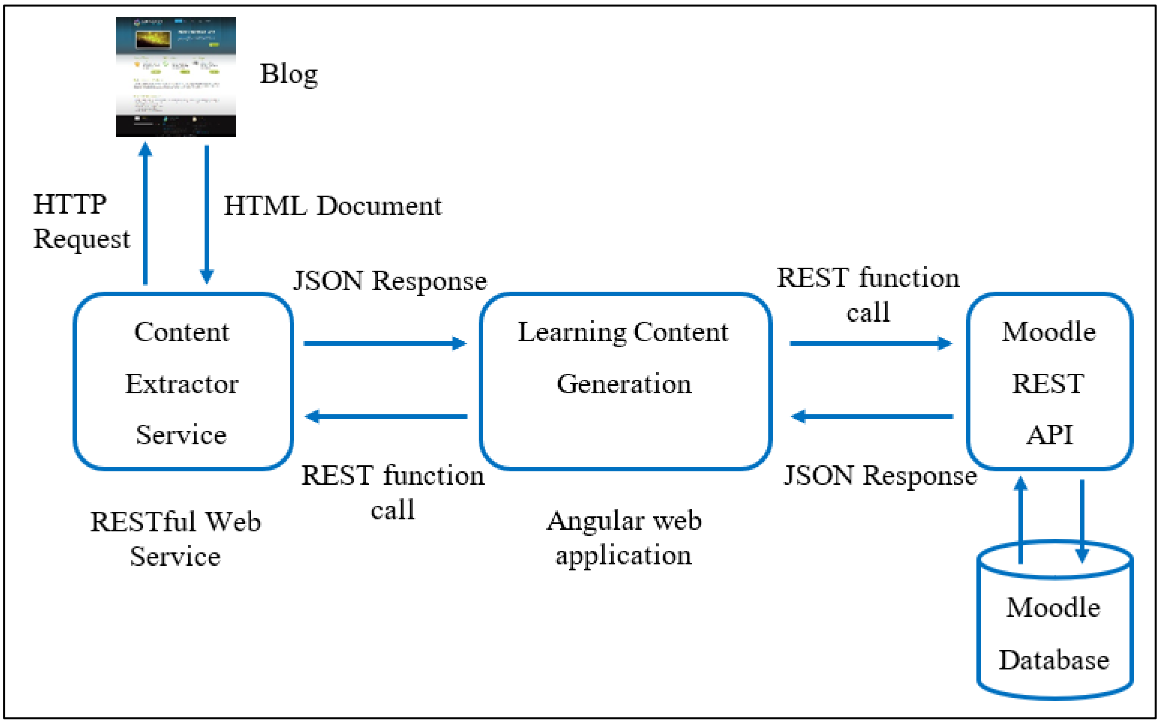
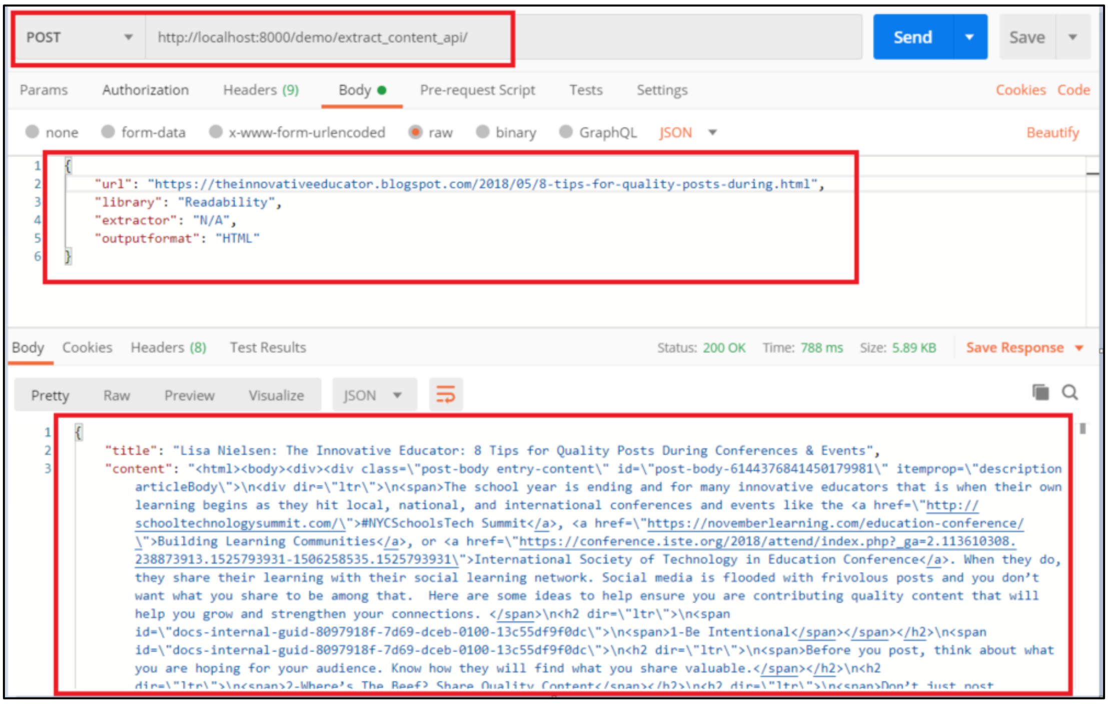


Figure 3.7: Component Architecture of the Prototype

3.2.1 Content Extractor Service

Content extractor service is a Python Django web service [76] based on Django REST framework [77]. Main function of this is to extract HTML content from a blog page based on the input from the Learning content generation component. In order to facilitate that, the Content extractor service exposes a service end point which can be accessed via REST protocol.

The request which comes from the Learning content generation component is a REST API call. Below Figure 3.8 shows a sample REST call made, parameters in the request body and the output. Any REST supported API client can be used to test the output. In this example Postman API client [75] is used. In the below Figure 3.8, REST API call, request body and the output JSON are highlighted.



```
{
    "url": "https://theinnovativeeducator.blogspot.com/2018/05/8-tips-for-quality-posts-during.html",
    "library": "Readability",
    "extractor": "N/A",
    "outputformat": "HTML"
}
```

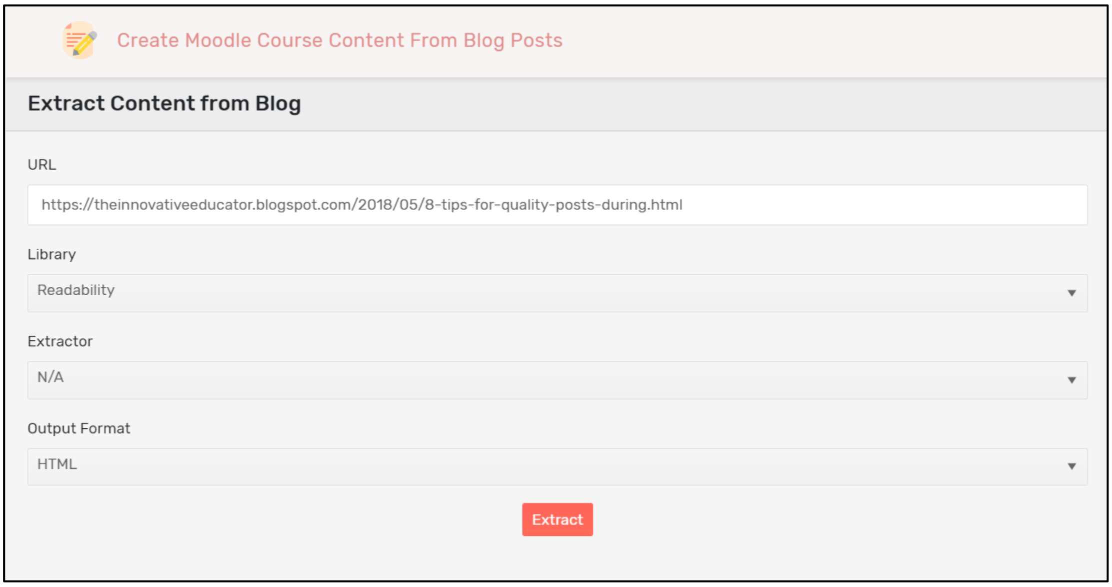
Status: 200 OK Time: 788 ms Size: 5.89 KB

```
{
    "title": "Lisa Nielsen: The Innovative Educator: 8 Tips for Quality Posts During Conferences & Events",
    "content": "<html><body><div><div class=\"post-body entry-content\" id=\"post-body-6144376841450179981\" itemprop=\"description
        articleBody\">\n<div dir=\"ltr\">\n<span>The school year is ending and for many innovative educators that is when their own
        learning begins as they hit local, national, and international conferences and events like the <a href=\"http://
        schooltechnologysummit.com/\">#NYCSchoolsTech Summit</a>, <a href=\"https://novemberlearning.com/education-conference/
        \">Building Learning Communities</a>, or <a href=\"https://conference.iste.org/2018/attend/index.php?_ga=2.113610308.
        238873913.1525793931-1506258535.1525793931\">International Society of Technology in Education Conference</a>. When they do,
        they share their learning with their social learning network. Social media is flooded with frivolous posts and you don’t
        want what you share to be among that.  Here are some ideas to help ensure you are contributing quality content that will
        help you grow and strengthen your connections. </span>\n<h2 dir=\"ltr\">\n<span
        id=\"docs-internal-guid-8097918f-7d69-dceb-0100-13c55df9f0dc\">\n<span>1-Be Intentional</span></span></h2>\n<span
        id=\"docs-internal-guid-8097918f-7d69-dceb-0100-13c55df9f0dc\">\n<h2 dir=\"ltr\">\n<span>Before you post, think about what
        you are hoping for your audience. Know how they will find what you share valuable.</span></h2>\n<h2
        dir=\"ltr\">\n<span>2-Where’s The Beef? Share Quality Content</span></h2>\n<h2 dir=\"ltr\">\n<span>Don’t just post
```

Figure 3.8: Content Extractor Service - sample REST API call

3.2.2 Learning Content Generation Component

Learning Content Generation Component is the interface between Content extractor service and Learning content integration component. Also, this is the place where the user interaction happens. This component is an Angular web application. From here, user can input the URL of the selected blog page where he/she needs to extract the content from, extractor library and the intended output format of the extracted content. When user starts the application, he or she can see a web page as shown in below Figure 3.9.



Extract Content from Blog

URL
https://theinnovativeeducator.blogspot.com/2018/05/8-tips-for-quality-posts-during.html

Library
Readability

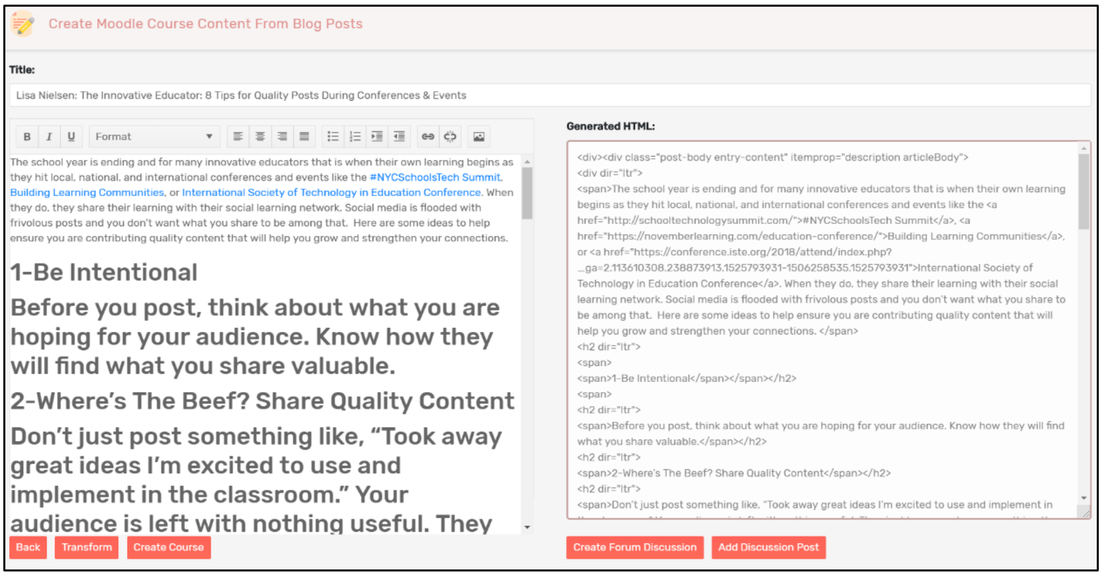
Extractor
N/A

Output Format
HTML

Figure 3.9: Learning Content Generation - Start Page

After entering “URL”, “Library” and the “Output Format”, user can press the “Extract” button. That is the triggering point for Learning content generation component to send a REST API call to Content Extractor Service. Then the Content Extractor Service sends a response with extracted content (HTML or Text depending on the values user has selected). This is the same response shown in above Figure 3.8.

Once the response is received, another page will be shown with the extracted content. If the user has selected HTML as the output format, both the extracted HTML and the preview will be shown as in Figure 3.10 given below.



Create Moodle Course Content From Blog Posts

Title:

Lisa Nielsen: The Innovative Educator: 8 Tips for Quality Posts During Conferences & Events

The school year is ending and for many innovative educators that is when their own learning begins as they hit local, national, and international conferences and events like the #NYCSchoolsTech Summit, Building Learning Communities, or International Society of Technology in Education Conference. When they do, they share their learning with their social learning network. Social media is flooded with frivolous posts and you don't want what you share to be among that. Here are some ideas to help ensure you are contributing quality content that will help you grow and strengthen your connections.

1-Be Intentional

Before you post, think about what you are hoping for your audience. Know how they will find what you share valuable.

2-Where's The Beef? Share Quality Content

Don't just post something like, “Took away great ideas I'm excited to use and implement in the classroom.” Your audience is left with nothing useful. They

Generated HTML:

```
<div><div class="post-body entry-content" itemprop="description articleBody">
<div dir="ltr">
<span>The school year is ending and for many innovative educators that is when their own learning
begins as they hit local, national, and international conferences and events like the <a
href="http://schooltechnologysummit.com/">#NYCSchoolsTech Summit</a>, <a
href="https://novemberlearning.com/education-conference/">Building Learning Communities</a>,
or <a href="https://conference.iste.org/2018/attend/index.php?
_ga=2.113610308.238873913.1525793931-1506258535.1525793931">International Society of
Technology in Education Conference</a>. When they do, they share their learning with their social
learning network. Social media is flooded with frivolous posts and you don't want what you share to
be among that.  Here are some ideas to help ensure you are contributing quality content that will
help you grow and strengthen your connections. </span>
<h2 dir="ltr">
<span>
<span>1-Be Intentional</span></span></h2>
<span>
<h2 dir="ltr">
<span>Before you post, think about what you are hoping for your audience. Know how they will find
what you share valuable.</span></h2>
<h2 dir="ltr">
<span>2-Where's The Beef? Share Quality Content</span></h2>
<h2 dir="ltr">
<span>Don't just post something like, "Took away great ideas I'm excited to use and implement in
```

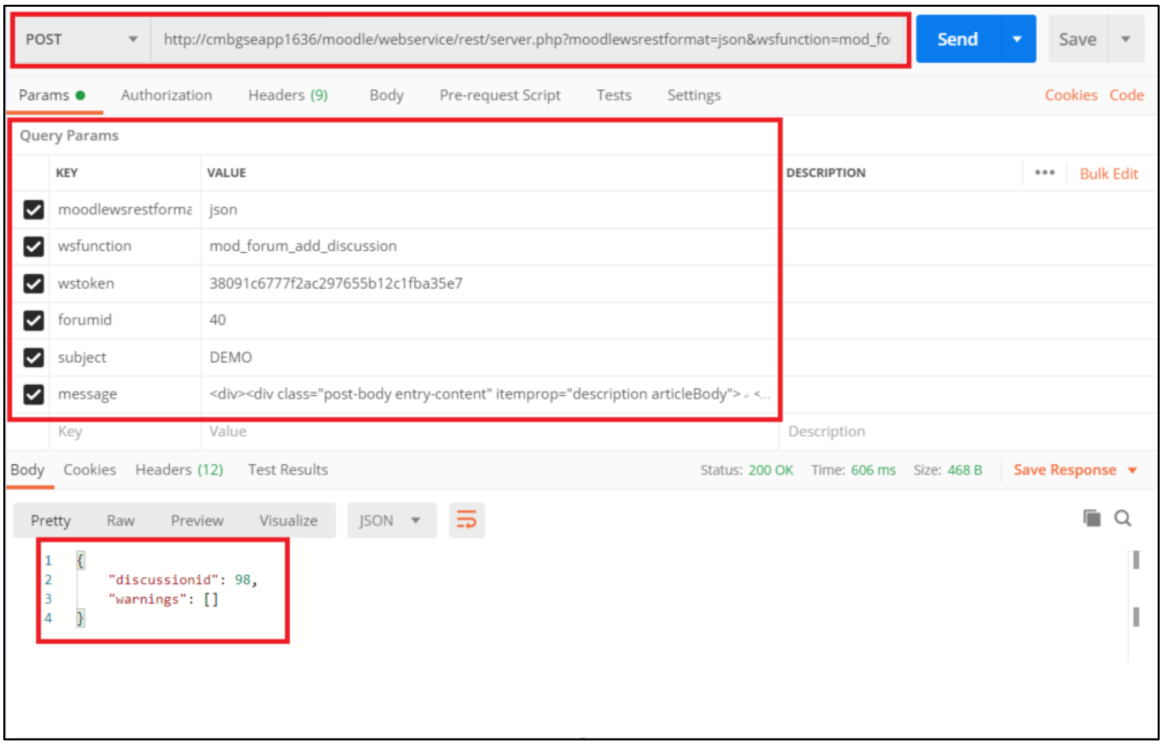
Figure 3.10: Learning Content Generation - Extracted Content

Then the generated content can be sent to Moodle. Optionally user can create a Moodle Course from this page as well (using “Create Course” option).

3.2.3 Learning Content Integration Component

Learning Content Integration Component is the “Middleman” between Learning Content Generation Component and Moodle. This is anyway an integral part of the Angular web application. Form this component the generated content is sent to Moodle via Moodle REST API.

Proposed model creates the e-Learning content as Forum Discussions under a Moodle course. Hence there should be previously created Course. It can be done either using the Prototype (“Create Course” option) or directly via Moodle web interface. Next step is creating a Forum under selected course (already discussed in section 3.1.2.2). After creating a Forum, then we can transfer the generated e-Learning content to Moodle. This can be done using “Create Forum Discussion Option”. This can also be demonstrated using Postman. Below Figure 3.11 shows a sample API call made - REST API call, request parameters and the output JSON are highlighted.



POST http://cmbgseapp1636/moodle/webservice/rest/server.php?moodlewsrestformat=json&wsfunction=mod_fo

Query Params

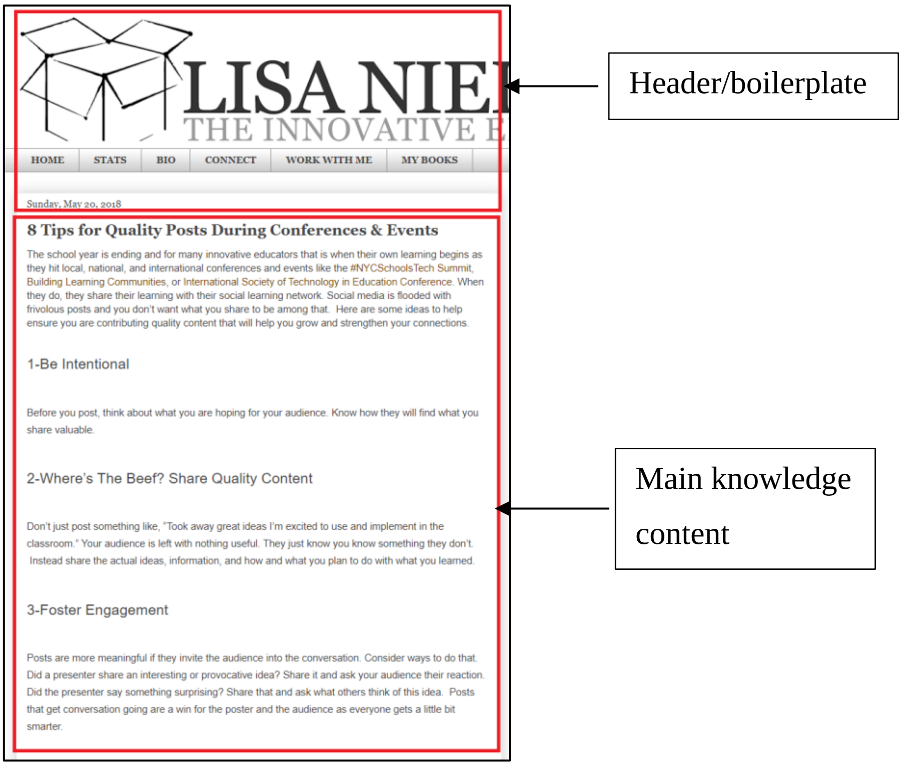
	KEY	VALUE	DESCRIPTION
☑	moodlewsrestforma	json	
☑	wsfunction	mod_forum_add_discussion	
☑	wstoken	38091c6777f2ac297655b12c1fba35e7	
☑	forumid	40	
☑	subject	DEMO	
☑	message	<div><div class="post-body entry-content" itemprop="description articleBody"> ↵ <...	
	Key	Value	Description

Status: 200 OK Time: 606 ms Size: 468 B

```
{
    "discussionid": 98,
    "warnings": []
}
```

Figure 3.11: Learning Content Integration Component - sample REST API call

Finally, the transferred course content can be found under the relevant Forum in Moodle. Below Figure 3.12 and 3.13 show the original Blog page and the content created in Moodle.



Sunday, May 20, 2018

8 Tips for Quality Posts During Conferences & Events

The school year is ending and for many innovative educators that is when their own learning begins as they hit local, national, and international conferences and events like the #NYCSchoolsTech Summit, Building Learning Communities, or International Society of Technology in Education Conference. When they do, they share their learning with their social learning network. Social media is flooded with frivolous posts and you don't want what you share to be among that. Here are some ideas to help ensure you are contributing quality content that will help you grow and strengthen your connections.

1-Be Intentional

Before you post, think about what you are hoping for your audience. Know how they will find what you share valuable.

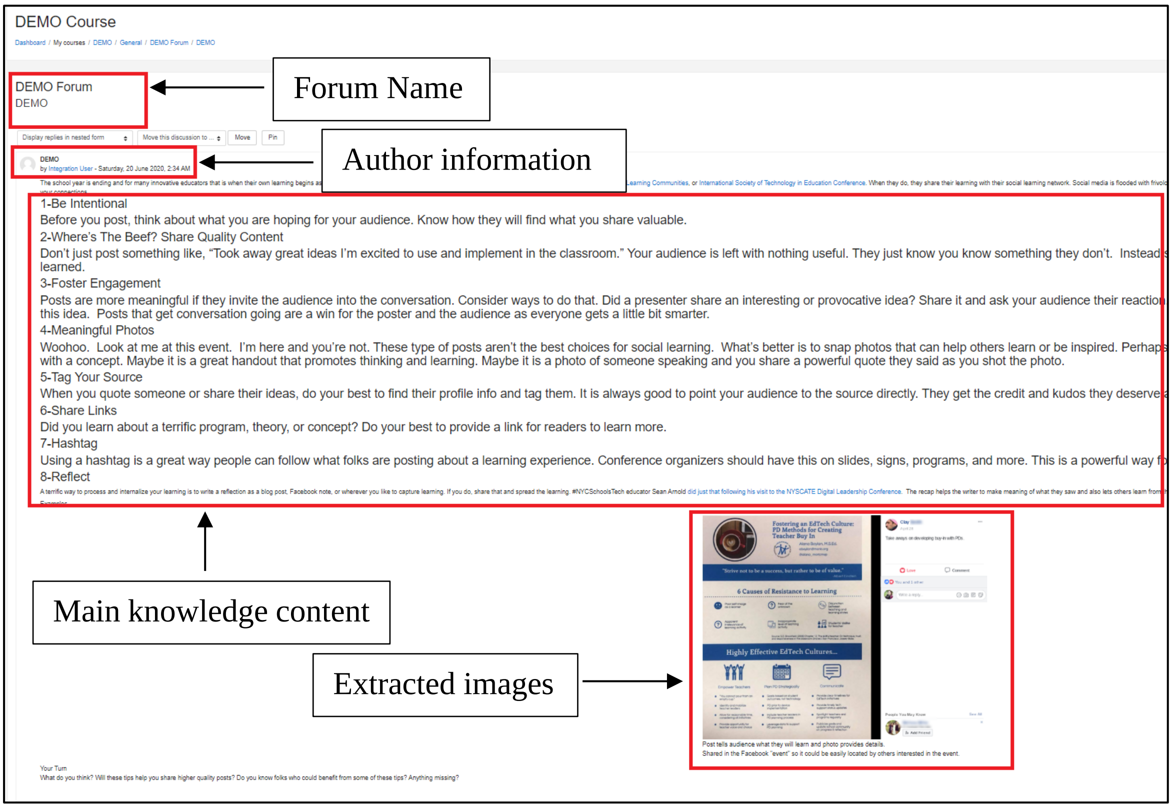
2-Where's The Beef? Share Quality Content

Don't just post something like, "Took away great ideas I'm excited to use and implement in the classroom." Your audience is left with nothing useful. They just know you know something they don't. Instead share the actual ideas, information, and how and what you plan to do with what you learned.

3-Foster Engagement

Posts are more meaningful if they invite the audience into the conversation. Consider ways to do that. Did a presenter share an interesting or provocative idea? Share it and ask your audience their reaction. Did the presenter say something surprising? Share that and ask what others think of this idea. Posts that get conversation going are a win for the poster and the audience as everyone gets a little bit smarter.

Figure 3.12: Original Blog Page


Display replies in nested form
Move this discussion to ...
Move
Pin
Forum Name
Author information
Main knowledge content
Extracted images
Fostering an EdTech Culture: PD Methods for Creating Teacher Buy In
"Strive not to be a success, but rather to be of value."
6 Causes of Resistance to Learning
Highly Effective EdTech Cultures...
Empower Teachers
Plan PD Strategically
Communicate
Take aways on developing buy-in with PDs.
Love
Comment
People You May Know
See All
Add Friend

DEMO Course

Dashboard / My courses / DEMO / General / DEMO Forum / DEMO

DEMO Forum
DEMO

DEMO
by Integration User - Saturday, 20 June 2020, 2:34 AM

The school year is ending and for many innovative educators that is when their own learning begins as

Learning Communities, or International Society of Technology in Education Conference. When they do, they share their learning with their social learning network. Social media is flooded with frivol

your connections.

1-Be Intentional
Before you post, think about what you are hoping for your audience. Know how they will find what you share valuable.
2-Where's The Beef? Share Quality Content
Don't just post something like, "Took away great ideas I'm excited to use and implement in the classroom." Your audience is left with nothing useful. They just know you know something they don't. Instead
learned.
3-Foster Engagement
Posts are more meaningful if they invite the audience into the conversation. Consider ways to do that. Did a presenter share an interesting or provocative idea? Share it and ask your audience their reaction
this idea. Posts that get conversation going are a win for the poster and the audience as everyone gets a little bit smarter.
4-Meaningful Photos
Woohoo. Look at me at this event. I'm here and you're not. These type of posts aren't the best choices for social learning. What's better is to snap photos that can help others learn or be inspired. Perhaps
with a concept. Maybe it is a great handout that promotes thinking and learning. Maybe it is a photo of someone speaking and you share a powerful quote they said as you shot the photo.
5-Tag Your Source
When you quote someone or share their ideas, do your best to find their profile info and tag them. It is always good to point your audience to the source directly. They get the credit and kudos they deserve
6-Share Links
Did you learn about a terrific program, theory, or concept? Do your best to provide a link for readers to learn more.
7-Hashtag
Using a hashtag is a great way people can follow what folks are posting about a learning experience. Conference organizers should have this on slides, signs, programs, and more. This is a powerful way f
8-Reflect
A terrific way to process and internalize your learning is to write a reflection as a blog post, Facebook note, or wherever you like to capture learning. If you do, share that and spread the learning. #NYCSchoolsTech educator Sean Arnold did just that following his visit to the NYSCATE Digital Leadership Conference. The recap helps the writer to make meaning of what they saw and also lets others learn from
Examples

Post tells audience what they will learn and photo provides details.
Shared in the Facebook "event" so it could be easily located by others interested in the event.

Your Turn
What do you think? Will these tips help you share higher quality posts? Do you know folks who could benefit from some of these tips? Anything missing?

Figure 3.13: Content Integrated with Moodle Course – Forum Discussion

3.3 Data Set

As already mentioned in section 3.1.1.5, the model was evaluated using 32 recent blog posts in the company internal blog, both simple and complex. Selected dataset comprises of 5 Simple blog posts and 27 complex blog posts. Categorizing the blogs as simple and complex has done using below criteria.

- Simple – only contains texts (paragraphs, headings), hyperlinks, no bulleted or numbered lists
- Complex – contains texts (paragraphs, headings), hyperlinks, bulleted and/or numbered lists, embedded images, videos, code segments, emojis

3.4 Evaluation Criteria

Process from extracting content from a blog to transferring it to Moodle is repeated for all the selected blogs (simple and complex) per extractor library (Readability, Boilerpipe and NewsPaper). Then the quality of the generated forum discussions is compared with the original blog posts. Similar to the study done in [13], human judgment is used for determining the quality of the generated forum discussion and categorized in to three levels: “Perfect”, “Satisfactory” and “Failed”. Quality of the generated forum discussions is assessed according to the criteria shown in below Table 3.3.

Evaluation Criteria	**Quality Level**
Extraction of the Title from the original Blog post	Perfect/Satisfactory/Failed
Extraction of the Content from the original Blog post	Perfect/Satisfactory/Failed
Usability of the generated forum discussion in Moodle	Perfect/Satisfactory/Failed
Overall Quality	Perfect/Satisfactory/Failed

Table 3.3: Evaluation Criteria

Using human judgment to evaluate something is very subjective and always prone to giving a bias result. In order to minimize the impact of the human bias, the created forum discussions in the Moodle were evaluated as a blinded experiment [78]. For that, two voluntary participants were selected from the workplace (two IFS employees who use the internal blog in their day-to-day activities). They have given the selected list of the blog posts from the company internal blog and asked them to assess the quality of the created content against the original blog posts. Results were collected as MS Excel document and kept as two separate files.

For the simplicity of the evaluation process, a numerical weight is assigned for each quality level as shown in below Table 3.4.

Quality Level	Numerical Weight
Perfect	2.0
Satisfactory	1.0
Failed	0.0

Table 3.4: Numerical Weight per Quality Level

Then those numerical weights were summed up and the average weight is taken for calculating the overall quality. Mapping between average weight and overall quality level is given below Table 3.5.

Average	Overall Quality Level
2.0	Perfect
1.0 – 1.99	Satisfactory
0.0 – 0.99	Failed

Table 3.5: Average Numerical Weight and Overall Quality Level Mapping

3.5 Chapter Summary

This chapter discussed the methodology used for the research. First, the chapter presented the overview of the design and the proposed model. Then detailed description about the model is given. Content extraction from blogs, different extractors, selecting blogs, selecting an LMS, integration details and the usability assessment of the proposed model were discussed in this chapter. Then the proposed model is explained with implementation details as well. Finally, the selected dataset and evaluation criteria were explained.

Chapter 4

Results and Analysis

This chapter presents the results and includes a descriptive analysis of the emerged results according to the pre-defined evaluation criteria given in section 3.4.

This section presents the results for the selected extractors mentioned in 3.1.1.3. All together there were 6 extractor instances* tested by two voluntary evaluators.

1. Readability
2. Boilerpipe DefaultExtractor
3. Boilerpipe ArticleExtractor
4. Boilerpipe LargestContentExtractor
5. Boilerpipe KeepEverythingExtractor
6. NewsPaper

***** Boilerpipe has 4 inbuilt extractors.

Above 6 extractors were run against the selected 32 blog posts and the results were collected, according to the evaluation criteria, corresponding numerical weights were given to calculate the total score, and then the average score was calculated.

4.1 Overall Results per Evaluation Criteria

This section presents the collected data per each evaluation criteria.

- Extraction of the Title from original blog post
- Extraction of the Content from original blog post
- Usability of the generated forum discussion in Moodle

For one criterion, results from both Evaluator 1 and Evaluator 2 are presented in a same bar chart.

4.1.1 Extraction of the Title from Original Blog Post

Figure 4.1 shows the results for extraction of title from the original blog posts.

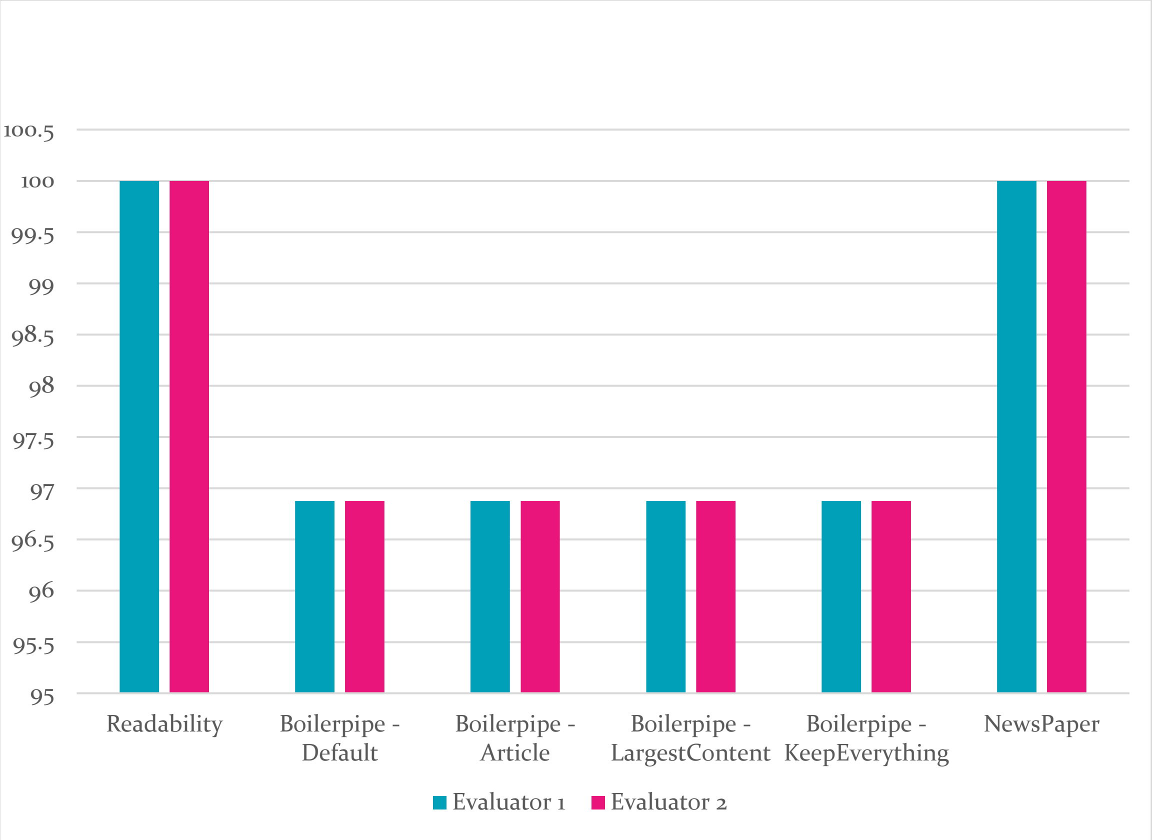


Figure 4.1: Extraction of the Title vs Extractor

According to the Figure 4.1, both Readability and NewsPaper libraries have been 100% successful in extracting the title from all 32 blog posts. Also, both the evaluators have got the dame results. Anyway, this is something expected as extracting the title is something binary.

4.1.2 Extraction of the Content from Original Blog Post

Figure 4.2 shows the results for extraction of content from the original blog posts.

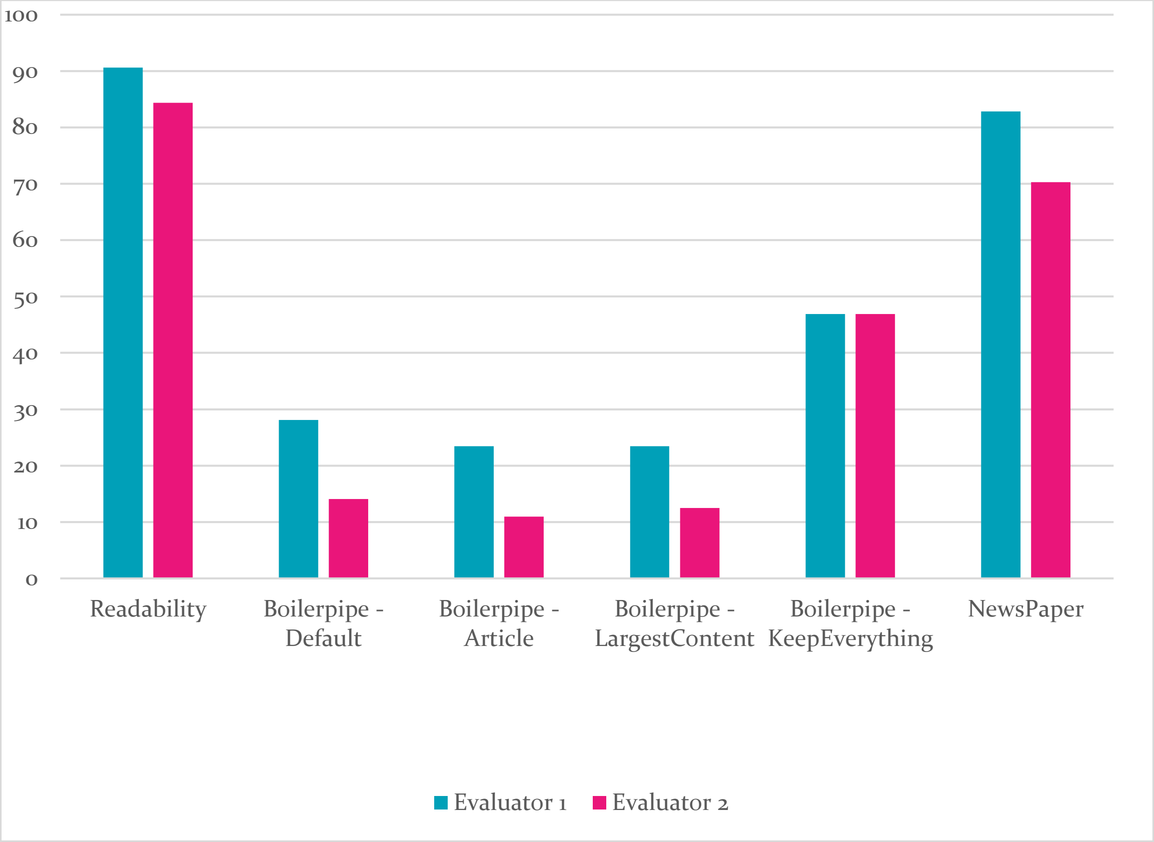


Figure 4.2: Extraction of the Content vs Extractor

Unlike in extracting titles, there are slight changes in the results from both evaluators. But still, Readability and Newspaper are performing well. If we compare those two extractors, Readability outperforms NewsPaper.

4.1.3 Usability of the Generated Forum Discussion in Moodle

Testing the usability aspect of the generated content in Moodle is more subjective. Figure 4.3 shows the results for extraction of content from the original blog posts.

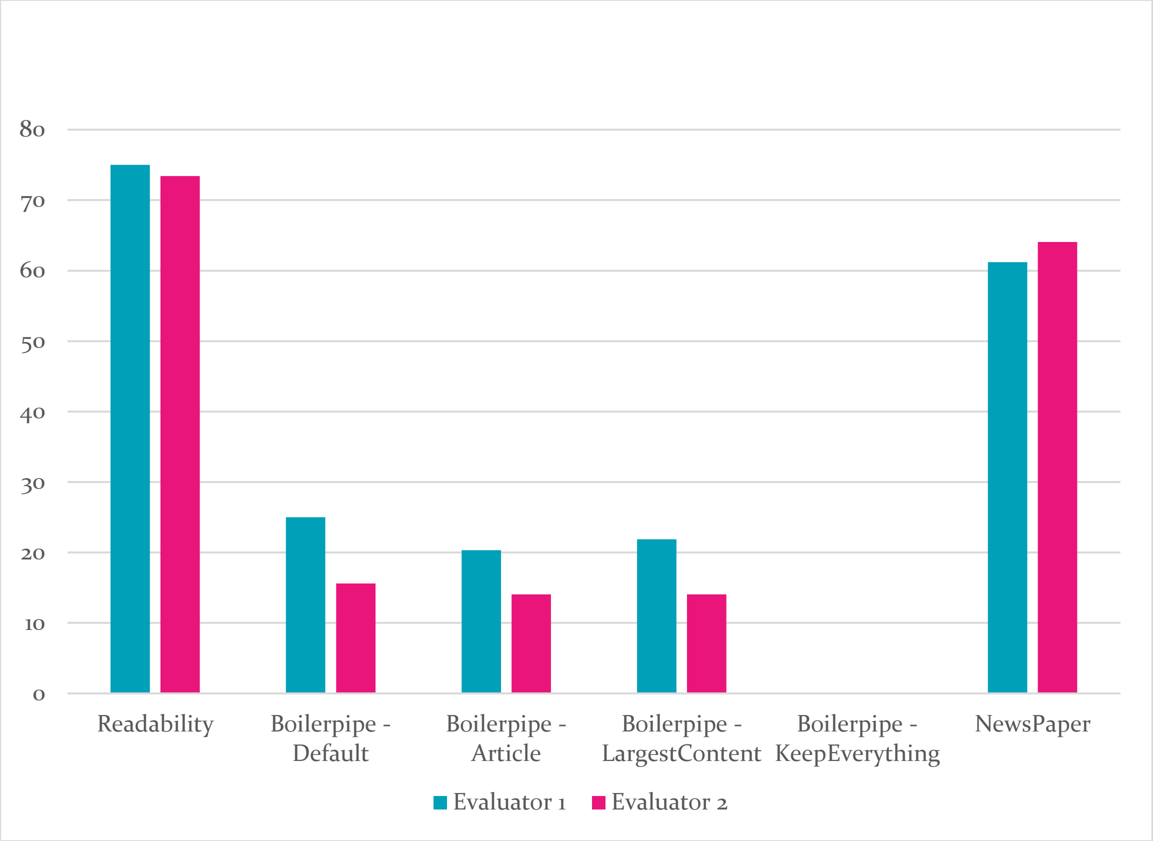


Figure 4.3: Usability of the Generated Forum Discussion vs Extractor

4.2 Overall Results per Extractor

This section presents the collected data per each extractor.

- Readability
- Boilerpipe - Default
- Boilerpipe - Article
- Boilerpipe - LargestContent
- Boilerpipe - KeepEverything
- NewsPaper

For one extractor, the results from both Evaluator 1 and Evaluator 2 are presented in the same bar chart according to the evaluation criteria and the overall score.

4.2.1 Readability Extractor

Figure 4.4 shows the performance of Readability Extractor per each evaluation criterion. Overall score is generated as the average of all the evaluation criteria and shown in the same bar chart.

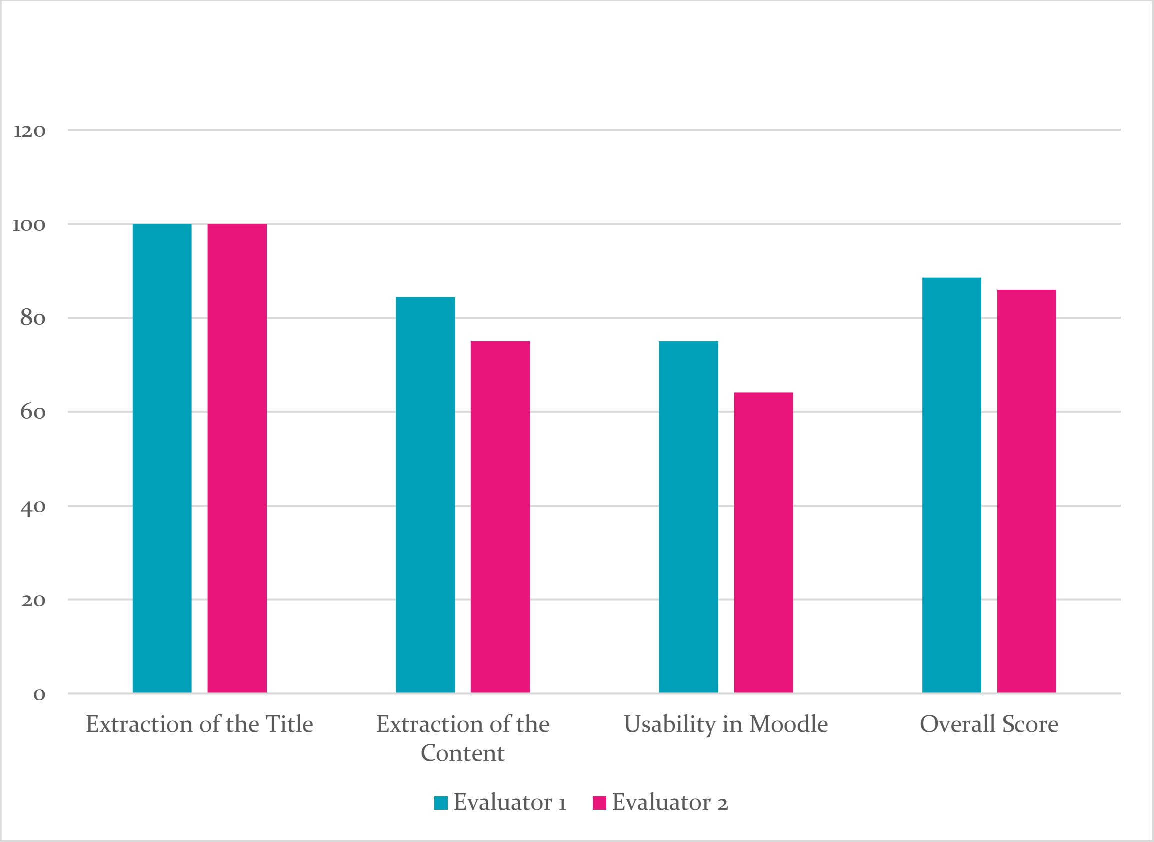


Figure 4.4:Performance of the Readability Extractor vs Evaluation Criteria

4.2.2 Boilerpipe – Default Extractor

Figure 4.5 shows the performance of Boilerpipe – Default Extractor per each evaluation criterion. In here also the overall score is generated as the average of all the evaluation criteria and shown in the same bar chart.

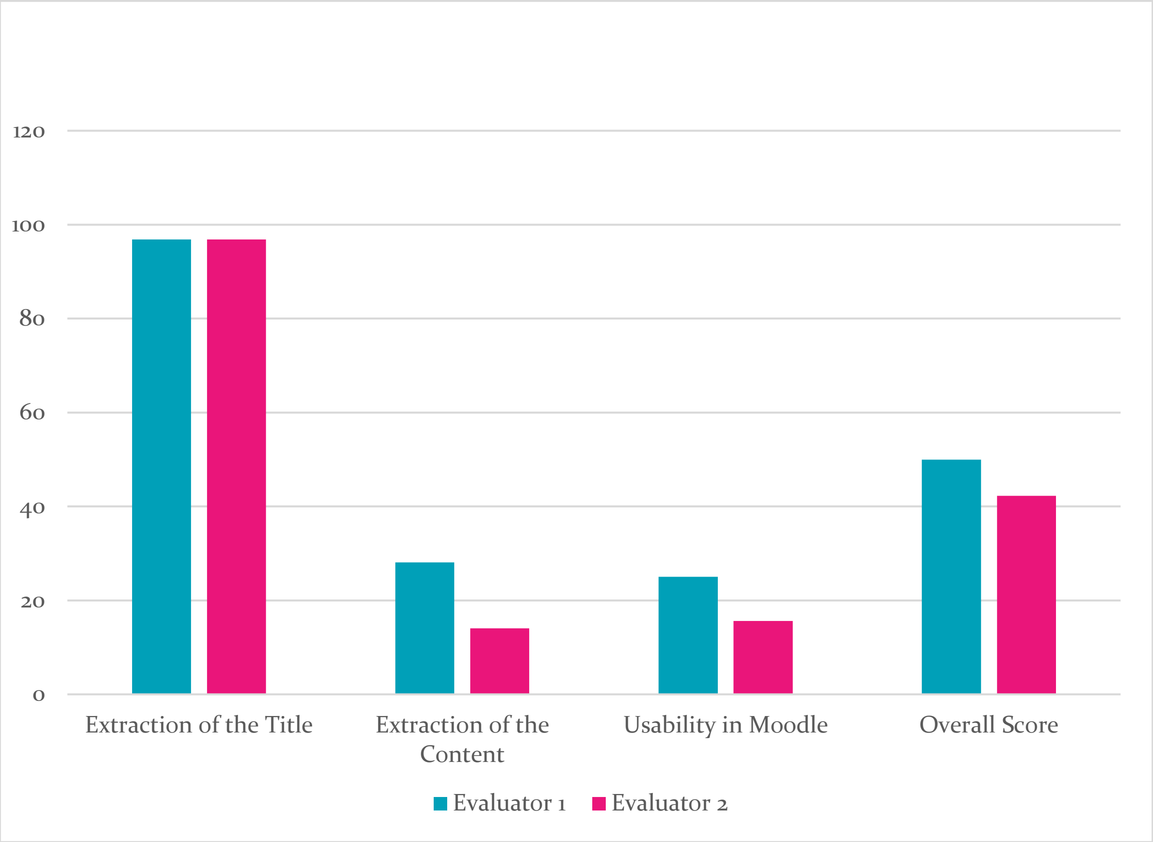


Figure 4.5: Performance of the Boilerpipe - Default Extractor vs Evaluation Criteria

4.2.3 Boilerpipe – Article Extractor

Figure 4.6 shows the performance of Boilerpipe – Article Extractor per each evaluation criterion. In here also the overall score is generated as the average of all the evaluation criteria and shown in the same bar chart.

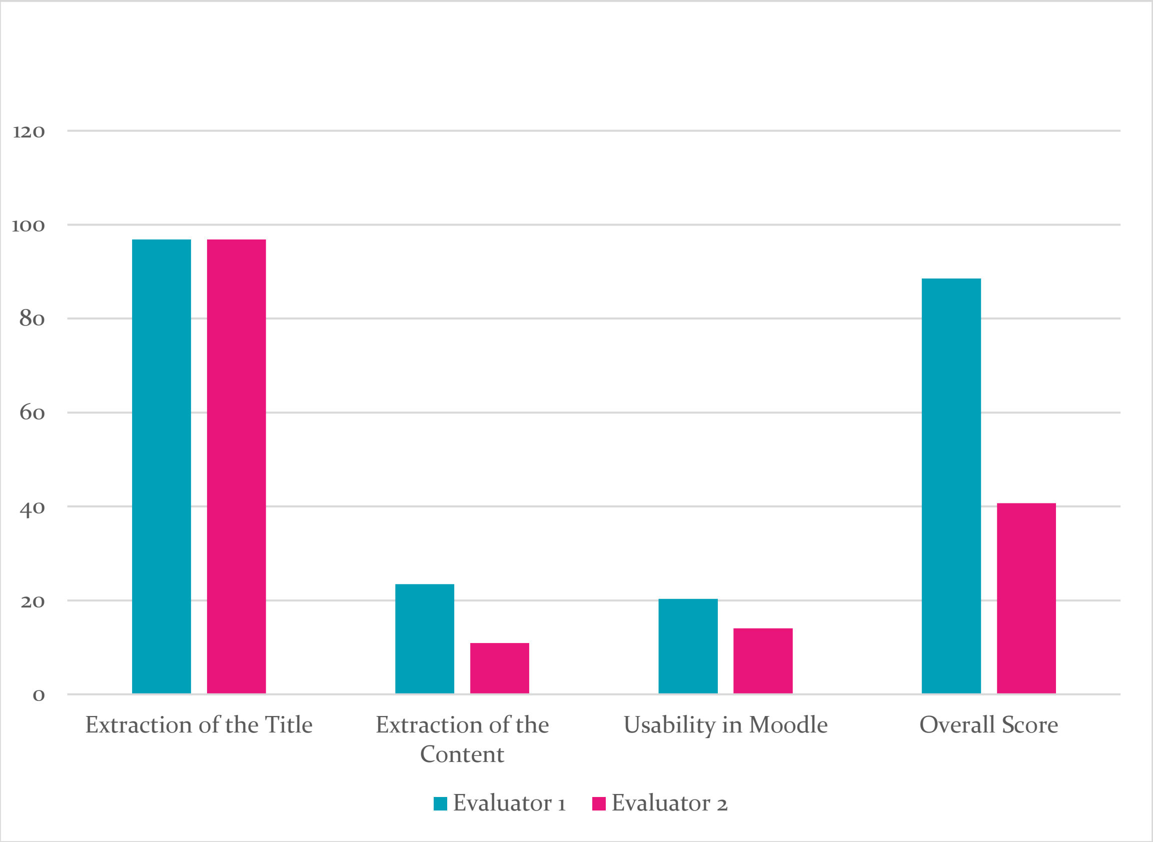


Figure 4.6: Performance of the Boilerpipe - Article Extractor vs Evaluation Criteria

4.2.4 Boilerpipe – Largest Content Extractor

Figure 4.7 shows the performance of Boilerpipe – Largest Content Extractor per each evaluation criterion. In here also the overall score is generated as the average of all the evaluation criteria and shown in the same bar chart.

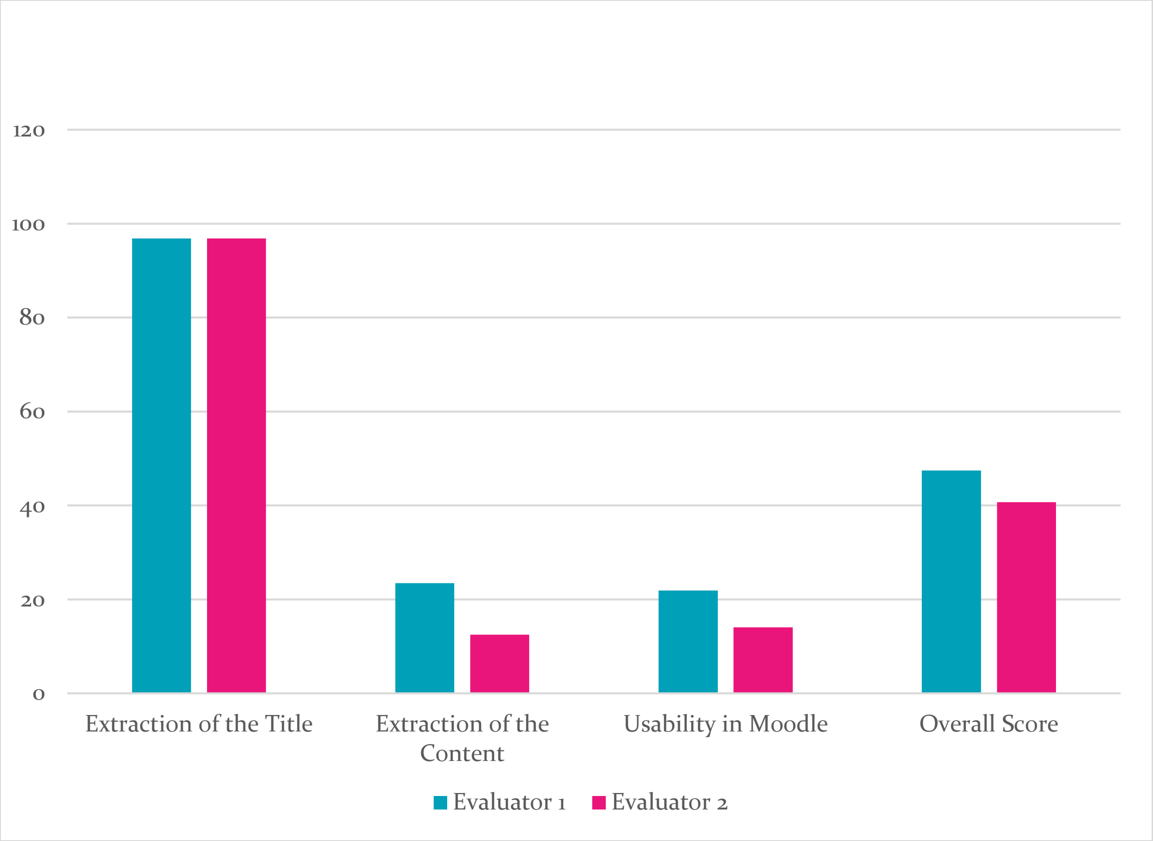


Figure 4.7: Performance of the Boilerpipe - Largest Content Extractor vs Evaluation Criteria

4.2.5 Boilerpipe – Keep Everything Extractor

Figure 4.8 shows the performance of Boilerpipe – Keep Everything Extractor per each evaluation criterion. In here also the overall score is generated as the average of all the evaluation criteria and shown in the same bar chart.

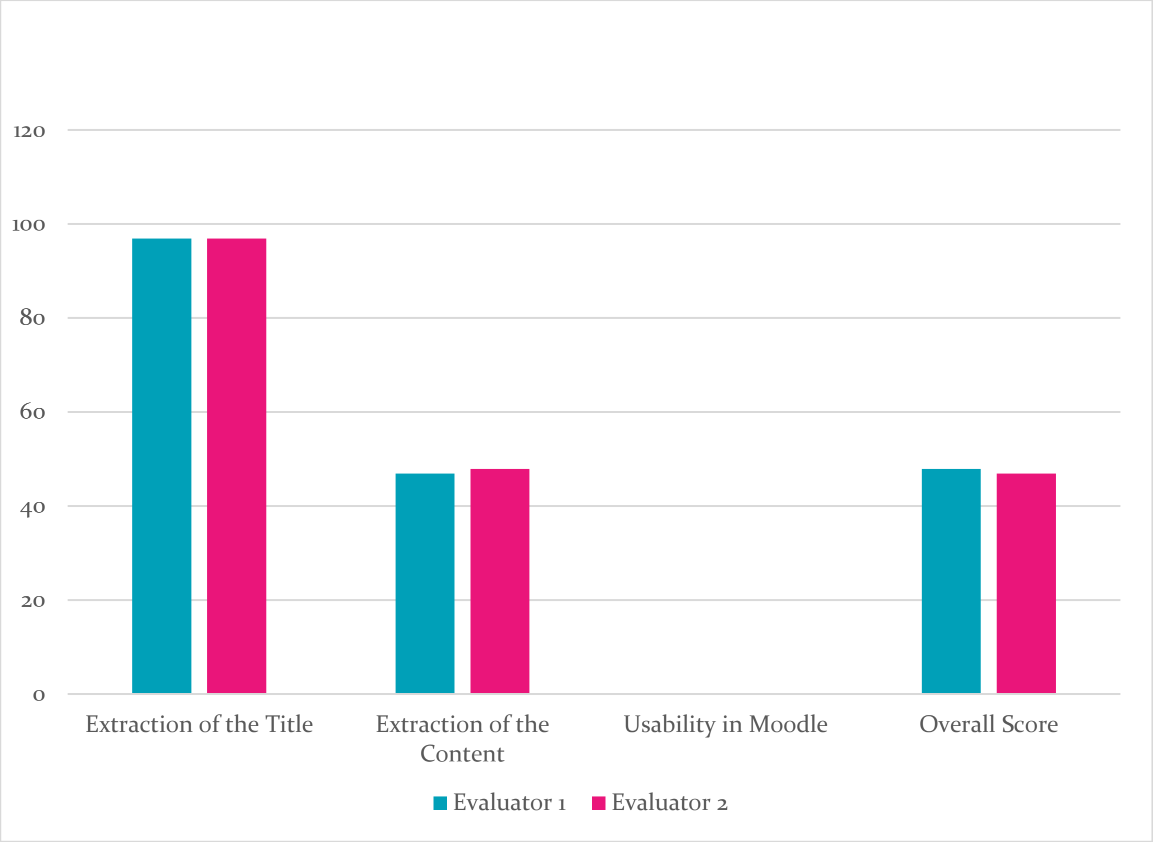


Figure 4.8: Performance of the Boilerpipe - Keep Everything Extractor vs Evaluation Criteria

4.2.6 NewsPaper Extractor

Figure 4.9 shows the performance of Boilerpipe – News Paper Extractor per each evaluation criterion. In here also the overall score is generated as the average of all the evaluation criteria and shown in the same bar chart.

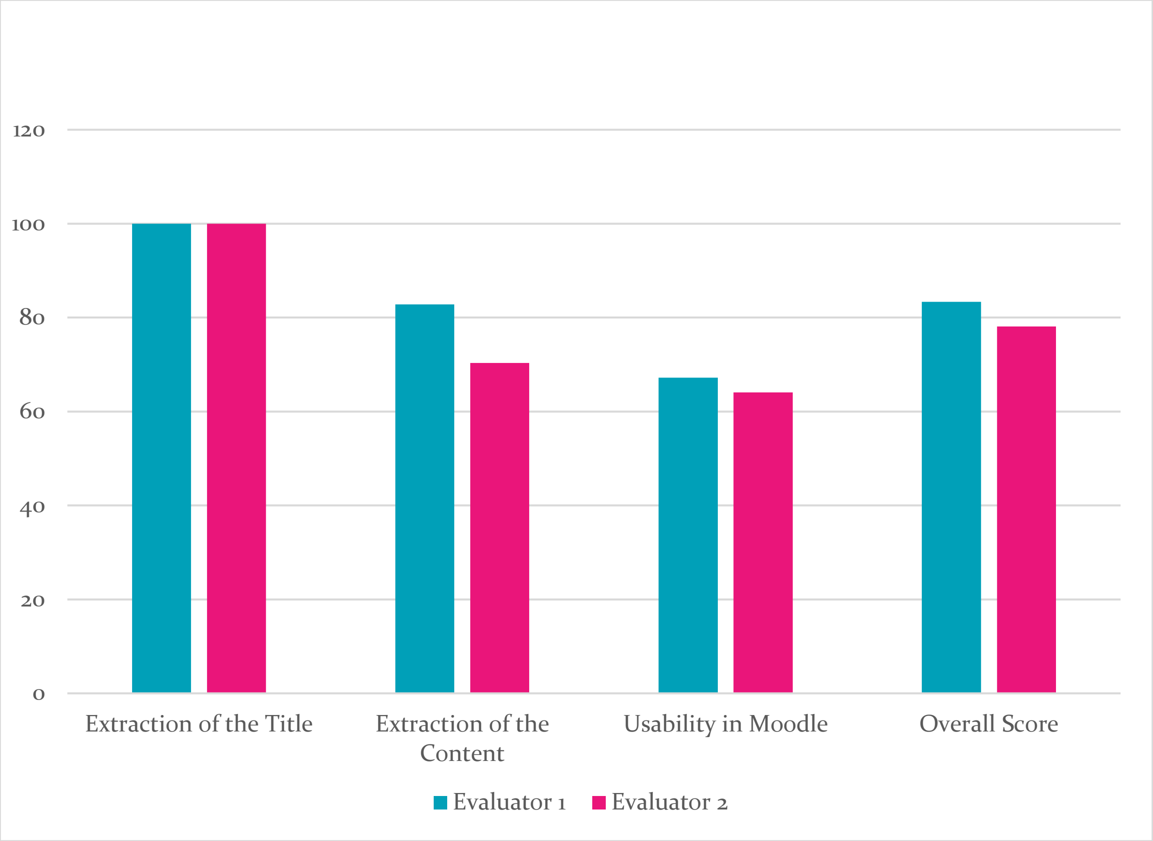


Figure 4.9: Performance of the News Paper Extractor vs Evaluation Criteria

4.3 Chapter Summary

This chapter presents the results gained from the proposed model running against the selected blog posts. Then the results were analyzed in different angles. Firstly, the selected extractors were evaluated against the evaluation criteria, extracting title text, extracting content and finally the created learning content in Moodle Forum discussions were evaluated based on the overall quality.

Chapter 5

Discussion

This chapter discusses about how the objectives of the research are achieved based on the results and the analysis done in section **Results and Analysis**. The objectives of the current study are selecting best appropriate method to extract knowledge content from blog posts, get the extracted content verified by human user and integrating the verified learning content with an existing LMS.

5.1 Selecting Best Appropriate Extraction Method

For the current study 6 extractors were selected and those extractors were evaluated against 32 selected blog posts using following criteria:

1. Extractor's capability of extracting title of the blog
2. Extractor's capability of extracting main content of the blog

From the analysis of the derived results, following facts were gathered regarding the best appropriate method for extracting content from blogs.

- When extracting the titles from blog posts, almost all the extractors perform well. Statistics reveals that all six extractors were successfully gave the correct results more than 95% of the cases. Out of all 6 extractors both Readability and NewsPaper outperform and gave 100% results for the selected blog posts.

- Both Readability and NewsPaper libraries outperform others in the content extraction as well. Readability gave the best results more than 85% of the cases whereas NewsPaper was successful more than 70% of the cases. All other four extractors were below the 50% mark.

All in all, Readability has a slight edge over Newspapers as its unique capabilities in identifying code blocks, emojis as well. On the other hand, NewsPaper also has unique features such as extracting keywords, summary of the content, author name, published date, main image URL, and embedded videos. But for the current study those features were not considered.

5.2 Verify the Extracted Knowledge Content

Verification part is vital as there is no guarantee that the extracted content is correct and relevant. This model proposes the intervention of a human user for that. After the extraction, content is presented to the person who is responsible for transferring the knowledge content from the blogs to an LMS. If we put this in the context of our initial problem, the same person who is responsible for the blog can be the person who is verifying the extracted content.

5.3 Integrate the Extracted Knowledge Content with Moodle

Among all the available LMSs, Moodle has been selected for the integration. And the proposed model was integrated with Moodle and the required operations were performed via Moodle REST API functions.

Both the quality of the verification and the Performance of the integration part is evaluated by the overall quality of knowledge content integrated to a Moodle course.

- Here also both Readability and NewsPaper libraries perform well compared to the other four extractors in Boilerpipe library.
- Among Readability and NewsPaper libraries, Readability shows its eminent performance with the quality of the extracted content as well.

All in all, Readability is the leading extractor among six of the selected extractors while NewsPaper is also exhibiting decent overall performance which is next to Readability.

Chapter 6

Conclusion and Future Work

This chapter concludes the work that has been done in the current study and gives the major contributions of the research and implications for future work.

The current study assesses the proposed model using selected blog posts in the same workplace and derives following conclusions.

- Among selected extractors, Readability is the one which is giving higher success rate
- Extracted content can be verified by a human user (Subject Matter Expert or Technical Expert) before transferring the content to LMS
- Extracted and verified knowledge content can be integrated with the selected LMS (Moodle) using the Moodle REST API functions and a simple web client developed in the prototype

6.1 Major contributions

This study addresses a problem in a workplace, “Automatic creation of e-learning content using blog posts”. Tech companies often use internal and external blogs to allow their employees to share the expert knowledge within the company instead of incorporating this knowledge with existing e-learning tools like virtual learning environments as it takes an extra effort and cost to integrate the knowledge in the blogs with the e-learning tools. In this study the researcher proposes a model to select the best appropriate method to extract the content from blogs, get the extracted content verified by a human user and finally a method to integrate knowledge content with an existing e-leaning system (Moodle).

6.2 Limitations

Only 32 selected blog posts were used as the test set and only two human evaluators were involved. Using human judgment to evaluate something is very subjective and always prone to giving a bias result.

Due to the limitation in Moodle REST API (Moodle REST API doesn’t contain an API function to create course activities and sections via a REST call), currently the user must create a Forum or wiki page before sending the learning content to Moodle.

6.3 Future Work

The current study touches two main areas in Computer Science discipline, Natural Language Processing and e-learning. Hence there are ample amount of future work can be done related to both the areas. Below I have listed some of identified future work as continuation of the current study.

- The proposed model is based on the selected three extractors written in Python. As highlighted in the Discussion, these extractors were not specifically built for extracting content from the blogs, those primarily built for extracting core content from general web pages based on heuristic methods. When analyzing the results, it is evident that none of the selected extractors was able to extract the comment section in the blogs. But blog comments are very crucial component in tech industry. Different web pages have their own format as those are built for serving different purposes. Hence it is impractical to come up with a “universal” extractor. Both heuristic and machine learning methods can be employed in tandem to develop a customized extractor for blogs and such an extractor will serve well as there is a learning part for the extractor.

- Currently the developed model supports only one blog page at a time. When creating e-learning course content, the Subject Matter Experts refer more than one blog page, hence the model can be extended to support more than one blog page. This can also be done with minimal effort by introducing an internal database to hold the multiple blog posts. Further, when considering situations where the LMS integration with an external tool is not allowed (due to some company policies, security considerations), if the extracted learning content can be developed as an SCORM package, then an authorized user can easily import that package to Moodle, further that package can be distributed among other Moodle instances as well (assuming a company maintain different LMSs in different sites).

- Due to the time constraints, the model was evaluated only with 32 blog posts with the help of two voluntary evaluators. Also, the evaluation is based on the human judgement.

As we all know humans tend to bias, when it comes to some qualitative concepts like completeness, aesthetics, appearance and things like that. If the test set can be increased (may be already available data set) and if the model can be tested by set of human users, a much better results can be drawn. Evaluating qualitative attributes like completeness, organization, aesthetics might be tricky, a statistical evaluation process can be introduced to use standard evaluation matrices like Precision, Recall and F1 score.

- Currently the proposed model integrates the content with Moodle as it is widely used LMS. But there are few other popular LMSs as well. Since the developed prototype supports the integration with a REST end point, integration with other LMSs REST end points can be done with a minimal effort.

- Due to the limitation in Moodle REST API (Moodle REST API doesn’t contain an API function to create course activities and sections via a REST call), currently the user must create a Forum or wiki page before sending the learning content to Moodle. This also can be rectified by creating a Moodle web service plugin as done in [31]. The new plugin can be developed to enable REST end points for creating sections in the course, creating forums. wikis and pages.

- Plugin development in Moodle is very famous in the open source community. There are lot of resources available (documentation, forums, etc.) for creating such plugins. There is a limitation in the current model as this model is creating learning content under already available activities (Forum and Wiki). A new activity plugin (SCORM compatible) can be designed and developed to create the learning content in our own format.

- User still has to copy and paste the blog post URL in the prototype. A browser plugin can be developed to directly invoke the Text extraction and process.

References

[1] “Outline of natural language processing,” *Wikipedia*. Aug. 09, 2018, Accessed: Aug. 16, 2018. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Outline_of_natural_language_processing&oldid=854129981.

[2] “Blog,” *Wikipedia*. Aug. 11, 2018, Accessed: Aug. 15, 2018. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Blog&oldid=854491039.

[3] “How To Use Blogs In the Classroom,” *eLearning Industry*, Sep. 26, 2013. https://elearningindustry.com/how-to-use-blogs-in-the-classroom (accessed Aug. 15, 2018).

[4] S. Luján-Mora and S. de Juana-Espinosa, “The Use of Weblogs in Higher Education: Benefits and Barriers,” 2007, pp. 1–7, Accessed: Aug. 15, 2018. [Online]. Available: http://desarrolloweb.dlsi.ua.es/blogs/use-of-weblogs-in-higher-education-benefits-and-barriers.

[5] D. R. Garrison, *E-Learning in the 21st Century : A Framework for Research and Practice*. Routledge, 2011.

[6] “Build eLearning Courses with Materials You Already Have,” *LearnUpon*, Jan. 09, 2018. https://www.learnupon.com/blog/build-courses-with-materials/ (accessed May 22, 2018).

[7] S. Sarawagi, “Information Extraction,” *Found. Trends® Databases*, vol. 1, no. 3, pp. 261–377, Nov. 2008, doi: 10.1561/1900000003.

[8] Y. (eugene Agichtein, Y. (eugene Agichtein, A. Professor, and L. Gravano, *ABSTRACT Extracting Relations from Large Text Collections*. 2005.

[9] P. Viola and M. Narasimhand, *Learning to Extract Information from Semistructured Text using a Discriminative Context Free Grammar*. .

[10] A. Carlson, J. Betteridge, R. C. Wang, E. R. Hruschka, and T. M. Mitchell, “Coupled semi-supervised learning for information extraction,” 2010, p. 101, doi: 10.1145/1718487.1718501.

[11] V. Gurusamy and S. Kannan, “Preprocessing Techniques for Text Mining,” Oct. 2014.

[12] S. Gupta, G. Kaiser, D. Neistadt, and P. Grimm, “DOM-based content extraction of HTML documents,” in *Proceedings of the 12th international conference on World Wide Web*, Budapest, Hungary, May 2003, pp. 207–214, doi: 10.1145/775152.775182.

[13] D. Cai, S. Yu, J.-R. Wen, and W.-Y. Ma, “Extracting Content Structure for Web Pages Based on Visual Representation,” in *Web Technologies and Applications*, vol. 2642, X. Zhou, M. E. Orlowska, and Y. Zhang, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 406–417.

[14] C. Kohlschütter, P. Fankhauser, and W. Nejdl, “Boilerplate detection using shallow text features,” in *Proceedings of the third ACM international conference on Web search and data mining - WSDM ’10*, New York, New York, USA, 2010, p. 441, doi: 10.1145/1718487.1718542.

[15] C. K. Nguyen, L. Likforman-Sulem, J.-C. Moissinac, C. Faure, and J. Lardon, “Web Document Analysis Based on Visual Segmentation and Page Rendering,” in *2012 10th IAPR International Workshop on Document Analysis Systems*, Gold Coast, Queenslands, TBD, Australia, Mar. 2012, pp. 354–358, doi: 10.1109/DAS.2012.95.

[16] M. Piotrowski, “Document-oriented e-learning components,” Jan. 2009.

[17] P. Gaur, “Research Trends in E-Learning,” *Media Commun. NIU J. Media Stud. ISSN 2395-3780*, vol. 1, pp. 29–41, Sep. 2015.

[18] “What is Elearning Courseware?,” *ProProfs Learning and Knowledge Management Resources*, Mar. 03, 2015. https://www.proprofs.com/c/lms/what-is-elearning-courseware/ (accessed Aug. 16, 2018).

[19] D. G. Cooper, *Research into Cognitive Load Theory and Instructional Design at UNSW*. .

[20] C. Mierlo, H. Jarodzka, F. Kirschner, and P. Kirschner, “Cognitive Load Theory in E-Learning,” *Encycl. Cyber Behav.*, vol. 1, pp. 1178–1211, Jan. 2012, doi: 10.4018/978-1-4666-0315-8.ch097.

[21] “E-Learning & Instructional Design 101,” *The Rapid E-Learning Blog*, Feb. 15, 2011. http://blogs.articulate.com/rapid-elearning/e-learning-instructional-design-101/ (accessed Aug. 16, 2018).

[22] “Cognitive Load Theory And Instructional Design,” *eLearning Industry*, Feb. 05, 2014. https://elearningindustry.com/cognitive-load-theory-and-instructional-design (accessed Aug. 16, 2018).

[23] “Instructional design,” *Wikipedia*. Jul. 19, 2018, Accessed: Aug. 16, 2018. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Instructional_design&oldid=851082146.

[24] B. Ghirardini and Organisation des Nations Unies pour l'alimentation et l'agriculture, *E-learning methodologies: a guide for designing and developing e-learning courses*. Rome: Food and Agriculture Organization of the United Nations, 2011.

[25] "Everything you need to know about LMSs » NEO LMS." https://www.neolms.com/info/everything_about_lms (accessed Jun. 17, 2020).

[26] A. Al-Ajlan and H. Zedan, "Why Moodle," in *2008 12th IEEE International Workshop on Future Trends of Distributed Computing Systems*, Oct. 2008, pp. 58–64, doi: 10.1109/FTDCS.2008.22.

[27] "Blackboard Learn," *Wikipedia*. May 27, 2020, Accessed: Jun. 17, 2020. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Blackboard_Learn&oldid=959176034.

[28] "Moodle," *Wikipedia*. Jun. 16, 2020, Accessed: Jun. 17, 2020. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Moodle&oldid=962866801.

[29] adminuser, "6th Annual LMS Data Update | edutechnica." http://edutechnica.com/2018/10/06/6th-annual-lms-data-update/ (accessed Jun. 17, 2020).

[30] "Academic LMS Market Share: A view across four global regions," *e-Literate,* Jun. 28, 2017. https://eliterate.us/academic-lms-market-share-view-across-four-global-regions/ (accessed Jun. 17, 2020).

[31] I. A. Kautsar, Y. Musashi, S.-I. Kubota, and K. Sugitani, "Developing Moodle plugin for creating learning content with another REST function call," in *2014 IEEE Global Engineering Education Conference (EDUCON)*, Apr. 2014, pp. 784–787, doi: 10.1109/EDUCON.2014.6826183.

[32] P. Viola and M. Narasimhand, *Learning to Extract Information from Semistructured Text using a Discriminative Context Free Grammar*. .

[33] K. Jung, K. In Kim, and A. K. Jain, "Text information extraction in images and video: a survey," *Pattern Recognit.*, vol. 37, no. 5, pp. 977–997, May 2004, doi: 10.1016/j.patcog.2003.10.012.

[34] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," *Found. Trends® Inf. Retr.*, vol. 2, no. 1–2, pp. 1–135, Jul. 2008, doi: 10.1561/1500000011.

[35] A. F. R. Rahman, H. Alam, and R. Hartono, "Content Extraction from HTML Documents," p. 4.

[36] Y. Ren and J. Tian, "Data Extraction Based on Page Structure Analysis," *MATEC Web Conf.*, vol. 139, p. 00118, 2017, doi: 10.1051/matecconf/201713900118.

[37] E. Halcomb and L. Hickman, "Mixed methods research," *Fac. Sci. Med. Health - Pap. Part A*, pp. 41–47, Jan. 2015, doi: 10.7748/ns.29.32.41.e8858.

[38] "7 Tips To Use Blogs In eLearning Course Design," *eLearning Industry*, Jan. 25, 2016. https://elearningindustry.com/7-tips-use-blogs-in-elearning-course-design (accessed May 22, 2018).

[39] "Using Blogs in your teaching | E-Learning Unit." https://elearning.qmul.ac.uk/enhancing-your-teaching/using-social-media/using-blogs-in-your-teaching/ (accessed May 22, 2018).

[40] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manag.*, vol. 24, no. 5, pp. 513–523, Jan. 1988, doi: 10.1016/0306-4573(88)90021-0.

[41] M. Rajman and R. Besançon, "Text Mining - Knowledge extraction from unstructured textual data," in *Advances in Data Science and Classification*, Springer, Berlin, Heidelberg, 1998, pp. 473–480.

[42] M. Grineva, M. Grinev, and D. Lizorkin, "Extracting Key Terms from Noisy and Multitheme Documents," in *Proceedings of the 18th International Conference on World Wide Web*, New York, NY, USA, 2009, pp. 661–670, doi: 10.1145/1526709.1526798.

[43] G. Rizzo and R. Troncy, "NERD: A Framework for Unifying Named Entity Recognition and Disambiguation Extraction Tools," in *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Stroudsburg, PA, USA, 2012, pp. 73–76, Accessed: May 22, 2018. [Online]. Available: http://dl.acm.org/citation.cfm?id=2380921.2380936.

[44] S. C. Sood and S. H. Owsley, "TagAssist: Automatic Tag Suggestion for Blog Posts," p. 7.

[45] Shri Shivaji Science & Arts College, Chikhali and S. Sirsat, "Extraction of Core Contents from Web Pages," *Int. J. Eng. Trends Technol.*, vol. 8, no. 9, pp. 484–489, Feb. 2014, doi: 10.14445/22315381/IJETT-V8P285.

[46] "Extracting clean data from blog and news articles," *Ujeebu blog*, Aug. 09, 2019. https://ujeebu.com/blog/how-to-extract-clean-text-from-html/ (accessed Jun. 16, 2020).

[47] "Heuristic," *Wikipedia*. Jun. 05, 2020, Accessed: Jun. 17, 2020. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Heuristic&oldid=960834734.

[48] "javascript - What algorithm does Readability use for extracting text from URLs?," *Stack Overflow*. https://stackoverflow.com/questions/3652657/what-algorithm-does-readability-use-for-extracting-text-from-urls (accessed Jun. 17, 2020).

[49] A. van Cranenburgh, *readability: Measure the readability of a given text using surface characteristics.* .

[50] "Mercury Web Parser by Postlight," *Mercury by Postlight*. https://mercury.postlight.com/web-parser/ (accessed Jun. 17, 2020).

[51] "boilerpipe," Aug. 08, 2018. https://boilerpipe-web.appspot.com/ (accessed Aug. 08, 2018).

[52] *dragnet-org/dragnet*. dragnet-org, 2020.

[53] L. Ou-Yang, *codelucas/newspaper*. 2020.

[54] X. Grangier, *goose-extractor: Html Content / Article Extractor, web scrapping*. .

[55] "Compare Diffbot to AlchemyAPI, Embedly, Readability, and Open-Source Article Extractors," *Diffbot*. https://www.diffbot.com/benefits/comparison/ (accessed Jun. 17, 2020).

[56] Z. Wang, X. Wang, and X. Wang, "Research and Implementation of Web-Based E-Learning Course Auto-generating Platform," in *Technologies for E-Learning and Digital Entertainment*, Jun. 2008, pp. 70–76, doi: 10.1007/978-3-540-69736-7_8.

[57] D. Abbakumov, "The Solution of the 'Cold Start Problem' in e-Learning," *Procedia - Soc. Behav. Sci.*, vol. 112, pp. 1225–1231, Feb. 2014, doi: 10.1016/j.sbspro.2014.01.1287.

[58] E. Diaz-Aviles, M. Fisichella, R. Kawase, W. Nejdl, and A. Stewart, "Unsupervised Auto-tagging for Learning Object Enrichment," in *Towards Ubiquitous Learning*, Sep. 2011, pp. 83–96, doi: 10.1007/978-3-642-23985-4_8.

[59] D. A. Wiley and Agency for Instructional Technology, Eds., *The instructional use of learning objects*, 1st ed. Bloomington, Ind: Agency for Instructional Technology : Association for Educational Communications & Technology, 2002.

[60] D. H. Anh, G. G. D. Nishantha, Y. Hayashida, and P. Davar, "A Flash-based lecture recording system and its integration with LMS," in *2010 The 12th International Conference on Advanced Communication Technology (ICACT)*, Feb. 2010, vol. 2, pp. 1425–1429.

[61] M. J. C. Guerrero, M. Á. C. González, M. A. Forment, and F. J. G. Peñalvo, "APPLICATIONS OF SERVICE ORIENTED ARCHITECTURE FOR THE INTEGRATION OF LMS AND m-LEARNING APPLICATIONS," 2009, pp. 54–59, doi: 10.5220/0001836900540059.

[62] I. Kautsar, S.-I. Kubota, Y. Musashi, and K. Sugitani, "A Supportive Tool for Lecturers to Upload LMS Learning Contents Automatically," Oct. 2013.

[63] "Using web services - MoodleDocs," Jul. 12, 2018. https://docs.moodle.org/35/en/Using_web_services (accessed Jul. 12, 2018).

[64] "Tutorial - MoodleDocs." https://docs.moodle.org/dev/Tutorial (accessed Jun. 15, 2020).

[65] “Web service API functions - MoodleDocs.” https://docs.moodle.org/dev/Web_service_API_functions (accessed Jun. 13, 2020).

[66] “13. How to get text from web pages — NLP 0 documentation.” http://www.tulane.edu/~howard/NLP/webpages.html (accessed Jun. 16, 2020).

[67] S. Wu, J. Liu, and J. Fan, “Automatic Web Content Extraction by Combination of Learning and Grouping,” in *Proceedings of the 24th International Conference on World Wide Web - WWW ’15*, Florence, Italy, 2015, pp. 1264–1274, doi: 10.1145/2736277.2741659.

[68] M. Tung, “Diffbot Leads in Text Extraction Shootout,” *Diffblog*, Jun. 14, 2011. https://blog.diffbot.com/diffbot-leads-text-extraction-shootout/ (accessed Jun. 18, 2020).

[69] masukomi (a k a K. Rhodes), *masukomi/arc90-readability*. 2020.

[70] “The Scala Programming Language.” https://www.scala-lang.org/ (accessed Jun. 19, 2020).

[71] “Web based Content Extraction and Retrieval in Web Engineering,” *Int. J. Recent Technol. Eng.*, vol. 8, no. 2S11, pp. 71–80, Nov. 2019, doi: 10.35940/ijrte.B1013.0982S1119.

[72] “Blogger.com - Create a unique and beautiful blog. It’s easy and free.” https://www.blogger.com (accessed Jun. 20, 2020).

[73] “Database schema introduction - MoodleDocs.” https://docs.moodle.org/dev/Database_schema_introduction (accessed Jun. 15, 2020).

[74] “Moodle REST Web Services tutorial – example – instructions – guidelines | Pavlos Spanidis.” http://www.spanidis.eu/?p=27 (accessed Jun. 15, 2020).

[75] “How to Use Postman API Client: GraphQL, REST, & SOAP Supported,” *Postman*. https://www.postman.com/product/api-client/ (accessed Jun. 19, 2020).

[76] “Home - Django REST framework.” https://www.django-rest-framework.org/ (accessed Jun. 20, 2020).

[77] “Representational state transfer,” *Wikipedia*. Jun. 13, 2020, Accessed: Jun. 20, 2020. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Representational_state_transfer&oldid=962367235.

[78] “Blinded experiment,” *Wikipedia*. Jun. 05, 2020, Accessed: Jun. 21, 2020. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Blinded_experiment&oldid=960844257.

Appendices

Appendix A: Consent Letter

K.A.S.N. Wijerathna,
Bannaggama,
Nikadalupotha,
Kurunegala,
Sri Lanka.

Date: ……………….

To: ………………………

………………………

Dear Sir/Madam,

I am Sujith Nishantha Wijerathna (K.A.S.N. Wijerathna), student of Master of Computer Science (2015 batch) bearing registration number 2015MCS080.

I am kindly requesting your participation in my Master's Research Study that I am conducting titled: *"Automatic e-Learning course content generation using existing blog posts".* As a part of this research, I aimed to design e-Learning Course content using content in blog posts.

As the Instructional Designers in your University, your inputs would be very beneficial for my study. Hereby, I would like to kindly invite you to participate in this research study by providing consent for this data gathering and completing the attached survey questionnaire.

I assure that I will not use your name or any other sensitive data in the analysis and reporting of results. I will only use pseudonyms that cannot be identified with you if we need to describe a statement or provide a citation from the transcript.

You can choose whether to give your consent for participating in this study or not. Even if you volunteer to participate and give your consent, you may withdraw at any time without penalty or loss of benefits to which you might otherwise be entitled.

If you have any questions or concerns about this research, please feel free to contact me (sujithkasn@gmail.com). If you agree to participate in this study having understood the procedures described above, include your signature and the date and complete the questionnaire attached to this form.

Your sincere response in this regard is highly appreciated. Your responses will be used only for academic purposes and will be treated with utmost confidentiality.

Thank you for your time. Your kind consideration in this regard is highly appreciated.

Yours sincerely,

K.A.S.N. Wijerathna.
MCS Student,
UCSC.

Contact Details
Official Address: IFS
Orion Towers, Orion City IT Park,
752, Dr. Danister De Silva Mawatha,
Colombo 9,
SRI LANKA

Mobile: +94718061720

E- Mail: sujithkasn@gmail.com

Consent Form

I have read and understand the aim of this data gathering and the conditions of this study. I have had all my questions answered. I know that if I participate, I may withdraw at any time without penalty. I hereby acknowledge the above and give my voluntary consent for participation in this study.

Signature: ……………………………………….

Date: …………………………………………….

Name (optional): ……………………………………………………………………………

Contact: phone or email: …………………………………………………………………

Appendix B: Questionnaire

Part I- Personal Information

Please tick (√) or fill appropriately.

1. Name (optional): ………………………….……………………………………………

2. Job Position: …………………………………………………………………………

3. Previous employment/s (if any): …………………………...……………………….

4. Years of Experience in the current employment: …………………

5. Age (optional): ………………………

6. Gender (optional):

Male: □ Female: □

Part II- Designing e-Learning content

Tick (√) or fill appropriately.

7. When designing e-learning courses, what type of content do you use?

Text documents: □

Audio: □

Video: □

Examinations: □

Surveys: □

Web: □ ……………………………………

Other: □

8. When designing e-Learning content, do you use templates?

Yes: □ No: □

9. When designing e-Learning content, do you follow any guidelines/standards?

Yes: □ No: □

10. If your answer for above (9) is “Yes”, what are those guidelines/standards?

……………………………………………………………………………………………..

……………………………………………………………………………………………..

……………………………………………………………………………………………..

……………………………………………………………………………………………..

Suppose that there is a tool/system to automatically extract content from blog posts and pre-process them for e-Learning course content. Below I have mentioned a list of requirements for such a tool/system.

Please tick (√) or fill appropriately.

11. I need to have the extracted content formatted as HTML rather than having plain text.

Yes: □ No: □

12. I need to have the default title of the extracted content generated as the title of the Blog post.

Yes: □ No: □

13. I need to see the extracted content formatted according to different templates.

Yes: □ No: □

14. I need to have the images of the original blog post in the extracted content.

Yes: □ No: □

15. I need to see the subtitles of the original blog post in the extracted content and the sub-title should be distinguishable.

Yes: □ No: □

16. Do you have any other requirements from such a tool/system that automatically extract content from blogs and pre-process them for e-Learning course content?

Yes: □ No: □

17. If your answer for above (16) is “Yes”, what are those requirements?

a. ………………………………………………………………………………………

b. ………………………………………………………………………………………

c. ………………………………………………………………………………………

d. ………………………………………………………………………………………

e. ………………………………………………………………………………………

18. Do you have any other comments about the tool/system that automatically extract content from blogs and pre-process them for e-Learning course content?

………………………………………………………………………………………………..

………………………………………………………………………………………………..

………………………………………………………………………………………………..

………………………………………………………………………………………………..

-Thank you very much for your kind cooperation-

Appendix C: Required Tools and Libraries

Visual Studio Code

Python 3.6

Django Web framework

Extractor libraries

Node.js

Angular 8

Postman