Bauhaus-Universität Weimar
Faculty of Media
Degree Programme Computer Science for Digital Media

Detecting Web Functions

Master's Thesis

Ademola Eric Adeuwmi

1. Referee: Prof. Dr. Benno Stein

Submission date: September 30, 2021

Declaration

Unless otherwise indicated in the text or references, this thesis is entirely the product of my own scholarly work.

Weimar, September 30, 2021

. . . .
Ademola Eric Adeuwmi

Abstract

Since the emergence of the World Wide Web (WWW), the interlinked hypertext documents (Webpages), which the WWW is composed of, fulfills essential purposes for its users. These Webpages can be separated into entertainment, giving information, data collection/forms, link/suggestion, discussion, or e-commerce. Webpages that belong to the same category have similar visual features, such as element structure, which users employ to navigate the webpage. Machines need to understand the structure of web pages for better classification. Web genre classifies Webpages by purpose; for machines to understand web genres, we need to train them on the Webpages' information using machine learning techniques. This thesis aims to build a classifier that automatically determines a Web page's functions by training a random forest model using a ground-truth dataset. The Webpages segmented into sections that fulfill a specific purpose. Header and body sections of a webpage carry information that aids a user in inferring its function. An annotator marks a web page's function by looking at the various segments and determining its purpose; however, machines need to analyze Webpages differently to know their function. Despite several applications in web content analysis, automatically determining the genre for web pages remains challenging. Colored sections highlight annotator mouse over indicating which section of that web page is the main reason they agree that a webpage belongs to any class. After annotation, we determine each worker's level of understanding using MACE (multi annotator competence estimation). We then use the annotation data to build a random forest classifier. The result shows that while most of the webpages are classified as information webpages. The majority were not e-commerce websites.

Contents

Acknowledgements

I thank Johannes Kiesel for being a patient mentor, and To my wife Olabisi Adewumi, thank you for the emotional support and encouragement. My gratitude goes to all the friends I made in Weimar, especially Sebastian Laverde, who always assists. Special thanks to Dominic Omachonu and his family for their support throughout my study in Germany.

Chapter 1

Introduction

Given the influx of technology and the ease of sharing ideas on the world wide web (internet) [1], webpage authors have various options in the function their webpages perform. Various webpage functions make detecting web page functions increasingly researched. Web page function is the authors' intentions towards the content of their web page. Contents of web pages make up semantically coherent visual segments [Sanoja and Gançarski, 2014]. Each segment has a purpose that reflects the author's intention and is relevant to downstream tasks such as targeted ads [Wikipedia contributors, 2021c].

The aim of our study is automatically detecting functions of the most extensive collection of annotated web pages [Couvillion, b] with high accuracy by use of machine learning classification techniques.

This thesis aims to accurately predict a web page's function using the webpage segmentation technique for ground truth data annotation. First, we need to build a model that represents web pages as decision trees in a random forest then classify them. [Wikipedia].

The research questions include; Can we annotate webpages of the largest publicly available dataset using a multiclass multi-label tool to identify segmented webpages consistently? [Zenodo] Secondly, can we train our model with the largest publicly available dataset with consistently identified segments? Will that will improve prediction accuracy?

This chapter introduces webpage function identification using a mental segmentation technique. Next are the use cases of Web functions, further explanation about our dataset, and automatic web function identification in the second chapter. The third chapter discusses the operationalization of webpage functions and defines our six distinct functions out of many possible functions. The fourth chapter discusses the base dataset and the annotation interface we improved, building a model, training the dataset, testing the model, and

[1] Wikipedia [2020]

comparing results with the ground truth data. Chapter five discusses experimental setup, features, random forest classification, and evaluation. Chapter six presents the classification performance results for binarized and multi-label predictions, then compares the ground truth to the prediction; lastly, in chapter seven, we conclude this thesis and discuss future work.

Chapter 2

Background

Genre is a crucial categorical concept. Many research communities are trying to exploit it for several uses, such as document management, the automatic generation of genre-specific documents, and information filtering. There is no agreement on the genre labels, on the genre systems, and taxonomy. However, according to a survey conducted by Marina Santini [Santini], says genres refer to a distinctive category of the discourse of any type, spoken or written, with or without literary aspirations. This study aims to collect a list and a short description of all the published works on automatic genre identification/classification in one document.

Web genre is a type of genre explored in multimedia [Wikipedia contributors, 2020]. There are several web genres: message boards, personal home pages, chat groups, virtual worlds, amongst others. Automatic web genre identification will dramatically improve search engine results, allowing the user to specify the desired Web genre along with a set of keywords. Furthermore, the American Society for information science and technology [Kwasnik] wrote a Bulletin with the topic, "Identifying document genre to improve web search effectiveness." This bulletin describes the notion of genre as a classification of communication by form or purpose, then discusses genres on the web and how a webpage visitor can be confused if they did not follow the evolution of that genre. This bulletin also surveyed the advancement in artificial intelligence and natural language processing, predicting its application in Automatic webpage function identification. This research sought to find out if detecting webpage function improves web search results concluding with promising results and plans of integrating web genre identification into a search engine for richer search results.

2.1 Use Case for Web Function Detection

The web is a massive but chaotic database of information. The web is chaotic because human productivity in generating web information has surpassed the ability to process it [Kwasnik et al., 2000]. This chaos makes processing web information of web pages a crucial step to detecting web page functions. Detecting web page functions is helpful in several downstream tasks such as targeted advertisements [Wikipedia contributors, 2021c]. Detecting webpage function is a critical factor in improving the often unsatisfactory results of search engines, as the user would be able to specify the desired Web genre along with a set of keywords [Rehm, 2002].

2.1.1 Targeted ads

Targeted advertising is a form of advertising, including online advertising, directed towards an audience with certain traits based on the product or person the advertiser is promoting. [Wikipedia contributors, 2021c]. Search engine marketing uses search engines to reach target audiences. Search engines benefit from the detection of web page functions by managing web pages based on their organization [Wikipedia contributors, 2021b]. Search engine optimization supports directed information retrieval ranked in terms of factors other than topical relevance. Experiments by Dewdney et al. [Kennedy and Shepherd, 2005b] have shown that the inclusion of genre information as part of the query can significantly improve search result precision. In addition, many website administrators use advertisements for revenue generation. Drawing the line between these different conventional and subtle advertisements is essential. A paper by Finn et al. [Finn et al., 2002] researches three approaches to classifying documents by genre, which are the traditional bag of words techniques, part-of-speech statistics, and hand-crafted shallow linguistic features. He focused on how well the classifier generalized from the training corpus based on a new corpus.

Documents recognized as advertisements are less likely to be swayed by questionable reasoning [Martin's]. For example, a webpage from a news site containing a fitness article might also have "noise information" from the advertisement of airlines. Therefore a search engine attempting to index the page's total content might choose keywords based upon the noise information instead of text related to the page's primary topic – fitness article. Thus, correctly detecting web page functions before indexing makes search engines rank pages better. Webpages that appear in search results where the query terms searched for are indexed as the primary function are likely to provide a much better experience for a search engine user. However, web pages of search query terms where the primary function is not indexed or unknown [Wikipedia contributors, 2021b] provide less user experience comparatively.

2.2 Genre Classification Systems

Genre classification is the process of grouping objects together based on defined similarities such as subject, format, style, or purpose [Abbott and Kim, 2008]. Genre classification creates conceptual links between different objects that can enhance browsing functionality and can be further developed into personalized retrieval or marketing tools. Genres develop from society to help authors accomplish a general purpose. For example, academic essays help academics demonstrate their knowledge to an instructor or as a way of passing information to a society or the community. In contrast, informative articles in newspapers, magazines, and newsletters help writers share information and ideas with their readers. In contrast to academic essays, opinion columns and letters to the editor are often used by writers to advance arguments. Although the notion of genre classification is still shrouded in ambiguity, it seems clear that we are striving towards a document typology that is different from topical classification.

It has also been observed that documents in a particular genre usually share a general purpose [Abbott and Kim, 2008]. Those documents tend to use similar writing conventions, such as the level of formality or the type of evidence used to support a point. For example, newspaper obituaries are usually severe and formal, while e-mail messages are informal and relaxed. Scholarly articles almost always refer to the source of evidence offered to support their points, while letters to the editor sometimes provide no evidence at all. Also, documents in a particular genre often use similar design elements. Academic essays, for example, are usually written with wide margins and double-spaced lines, while magazine articles often use columns and make extensive use of color and illustrations.

The Clearinghouse [Martin's], an open access publishing collaborative, published an article that explains how genres are shaped by the social, cultural, and disciplinary contexts from which they emerge. The article explains that when writers and readers form a community such as an academic discipline, a professional association, or a group that shares an interest in a particular topic or activity, they develop distinctive ways of communicating with one another. Over time, community members will agree to the type of evidence generally accepted to support arguments and the style and how documents should be designed and organized. As the needs of a community evolve, the genre will adapt as well. Articles in magazines for automobile or motorcycle enthusiasts differ in ways from articles in magazines about contemporary music. Scholarly articles are written by sociologists, civil engineers, and chemists similarly use evidence or organization in distinct ways.

Analyzing genres of webpages further, we discover that the web is fluid, unstable, and fast-paced. In addition, we found out that genres on the web are instantiated in webpages [Santini, 2007]. Web genres are relatively new and become more popular with the advent of the World Wide Web. Musical and literature, in contrast, are genres that have been in existence for a longer time.

A second study was conducted by A. Kennedy [Kennedy and Shepherd, 2005a] to distinguish home webpages from non-home webpages using a neural-net classifier. The paper defined home webpages as webpages that describe the interests and ambitions of a person, where those ambitions do not include making a profit through selling some product or service. First, Kennedy observed that a similar growth had matched the growth of the World Wide Web in the variety of cybergenres (web genres) found on the web. He explained that this growth includes replicating existing genres onto the web, the evolution of existing, and the spontaneous appearance of new genres. Kennedy found out that expanding and evolving web genres make it challenging to automatically identify a web page's genre. Additionally, he emphasized that it is difficult to know the boundaries of a genre and to know when one has crossed from one genre into another genre or when a web page represents the emergence of a new genre. Finally, the study results indicate that Kennedy's classifier can distinguish home pages from non-home pages. Within the home page genre, it can distinguish personal from corporate home pages. Organization home pages, however, were more difficult to distinguish from personal and corporate home pages. Thus we agree that there exist several classification systems for genres and web genres alike.

2.3 Web Genre Datasets

Based on research and the application of web page segmentation techniques, existing algorithms cannot be evaluated against large annotated datasets because they are primarily created for information retrieval and accessibility enhancements [Zenodo]. Furthermore, several datasets have been created for web page segmentation, but none has become a standard benchmark. Zenodo Instead, most algorithms come with a new dataset for their evaluation. Issues that prevent the reuse of the existing datasets include missing data sources such as unavailability of screenshots, bias due to heuristic annotations, no ground truth annotations, availability of only a few specific sample websites. As such, none of the previously published datasets combines completeness, reliability, diversity, and scale.

Additionally, vast annotation datasets lack diversity since their annotation process presumes all web pages to be homogeneous. While webpage segmentation algorithms attempt to use the same cues as humans to obtain a segmentation, the information they use to identify these cues and their interpretation varies. They require evaluation as performed in a thesis by Mayer [Zenodo] to determine which approach comes closest to human performance. Some algorithms create the visual representation of the web page visitors interact with and store the elements that make up the page as a Document Object Model (DOM) tree. The hierarchy of elements and information about webpage visual appearance (their style) contained therein is used by such algorithms to identify the abovementioned cues. However, the DOM does not contain information about the segments on a webpage. Other algorithms only use the rendered appearance of a web page in the form of a screenshot and segment purely visually. In contrast, our dataset Webis-seg-20 was generated from creating, resizing, and moving rectangles on a screenshot to specific segments, which annotators marked based on the function of the webpages to create our final dataset webis-webfunc-21. webis-webfunc-21 was created by integrating an annotation tool that allows crowd-sourced annotators to give further insights into important segments by marking web pages based on their function.

2.4 Automatic Web Genre Identification

When we identify web genres, we need to do it automatically because automating genre identification allows large-scale genre identification, which is relevant for comparative analysis of genre identification techniques. A paper published by [Rehm, 2002] argues for systematic analysis of academic webpages by creating a database system of over 100,000 HTML documents. Then introducing notions of web genre types constitutes the basic framework for specific compulsory and optional modules. The analysis of a 200 document sample illustrates Rehm's notion of Web genre hierarchy, into which Web genre types and modules are embedded.

Subsequently, A. Kennedy [Kennedy and Shepherd, 2005a] researched the automatic identification of home webpages. He planned to incorporate webpage genres into the search process in order to improve search results. Training a neural net classifier, he was able to distinguish home webpages from non-home webpages.

Shepherd and Watters [Shepherd and Watters, 1998] classified 96 randomly selected websites on the basis of content, form, and functionality. They

Table 2.1: Proportions of Cybergenres.

Cybergenre	S & W	C & W
Homepage	0.40	0.10
Brochure	0.17	0.06
Resource	0.35	0.82
Catalog	0.05	0.02
Game	0.03	0.00

used a much coarser-grained set of criteria and grouped the 96 sites into 5 major categories consisting of: home page, brochure, resource, catalog, and game. No search engines were among the 96 randomly selected websites. This classification was much coarser-grained than that of Crowston and Williams [Kevin Crowston, 2000], Shepherd and Watters proceeded to map Crowston and Williams' 48 genres into the 5 cybergenres they discovered with the results shown in Table 2.1

The column headed "S & W" represents the proportion of each cybergenre in Shepherd and Watters' sample of 96 websites. The column headed "C & W" represents the proportion of each cybergenre after mapping the 48 genres of Crowston and Williams's into the five cybergenres.

Although this was not done statistically, there appear to be significant differences in the proportions of each cybergenre. Shepherd and Watters indicate that while these differences may be due to several reasons, they believe the main reason may well be the enormous change on the web over the two years between the studies (1997–1999).

In 2001 [Roussinov et al., 2001], did a more extensive study of the genre on the web with 184 users. The web pages were tracked, and the respondents were asked to report the purpose or task that they were performing when viewing that page. There were 1234 web pages altogether. The interviewers coded the web pages with the addition of new genres as needed. There were 116 different genres identified. The respondents were asked to assign their web pages to the appropriate genres. Only 1076 web pages were successfully assigned to genre categories, with 49.63% between the interviewers and the respondents.

These studies reveal that the number of web genres seems to be growing and that it is often difficult to determine the genre of a web page [Kennedy and Shepherd, 2005a].

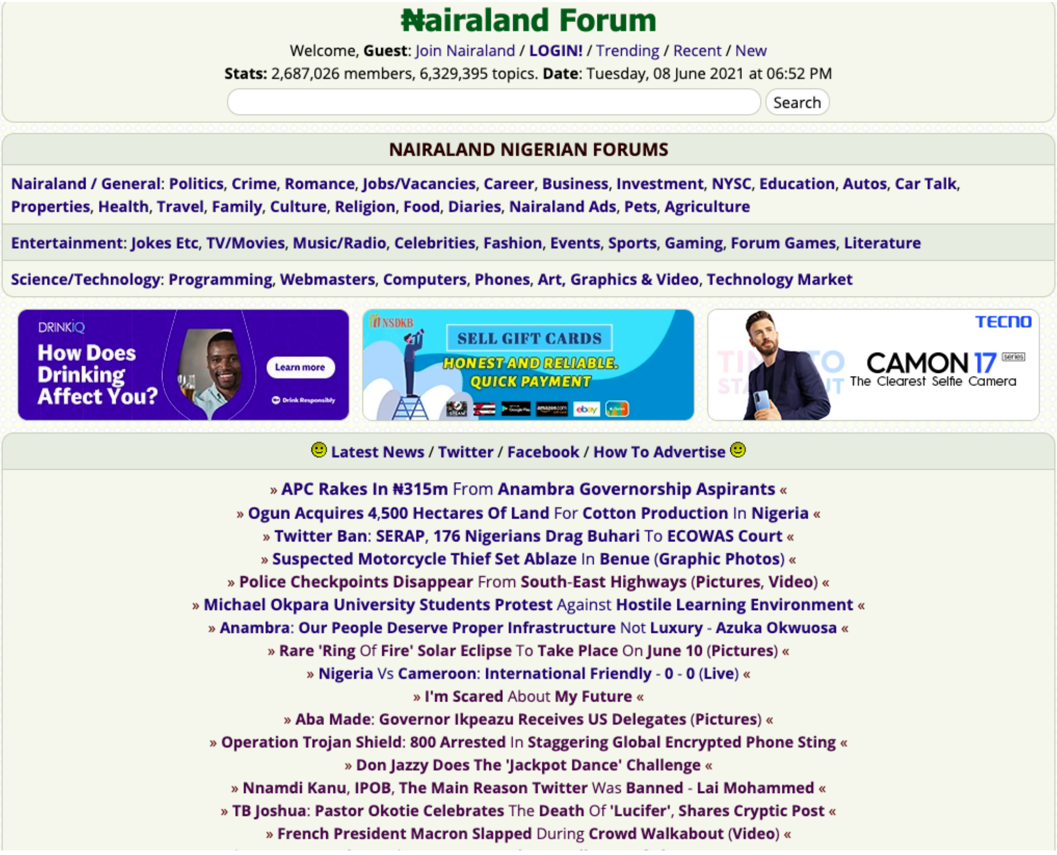
Chapter 3

Operationalizing Web Page Functions

In the previous chapter, we reviewed previous work on detecting webpage functions, which we followed with web functions classification systems and then web functions datasets. After that, we presented different automatic web genre identification approaches, emphasizing the importance of detecting web page functions with use cases such as targetted ads. Now in chapter three, we will define a webpage function as a way to identify, categorize or describe a webpage, [Boisvert]. First, we state that a web genre is a way of categorizing multimedia notions based on their similarity in the construction of media texts [Wikipedia contributors, 2020]. In summary, an author's intention towards the audience of a web page we regard as the function. Webpages can perform any one or more functions and belong to any one or more web genre concurrently. Some webpages are used for eCommerce, while others are used as discussion forums and entertaining the user. There are several webpage functions, just as there are several web genres.

In this thesis, we consider six functions a webpage can perform. We chose only these six web page functions because annotating six labels is faster for annotators since the more web page functions we annotate. The more elaborate the annotation task will reduce the inter-annotator agreement because there are too many webpage functions to choose from. Also, annotating for more webpage functions takes a longer time and costs more money. These six webpage functions are the cross-grained levels of webpage functions. However, we do not distinguish between personal levels in the following webpage functions.

Discussion webpages Discussion webpages (forums) allow users to hold conversations in the form of posted messages; they differ from chat rooms in that messages are often longer than one line of text and are at least temporarily



₦airaland Forum

Welcome, **Guest**: Join Nairaland / **LOGIN!** / Trending / Recent / New

Stats: 2,687,026 members, 6,329,395 topics. **Date:** Tuesday, 08 June 2021 at 06:52 PM

Search

NAIRALAND NIGERIAN FORUMS

Nairaland / General: Politics, Crime, Romance, Jobs/Vacancies, Career, Business, Investment, NYSC, Education, Autos, Car Talk, Properties, Health, Travel, Family, Culture, Religion, Food, Diaries, Nairaland Ads, Pets, Agriculture

Entertainment: Jokes Etc, TV/Movies, Music/Radio, Celebrities, Fashion, Events, Sports, Gaming, Forum Games, Literature

Science/Technology: Programming, Webmasters, Computers, Phones, Art, Graphics & Video, Technology Market

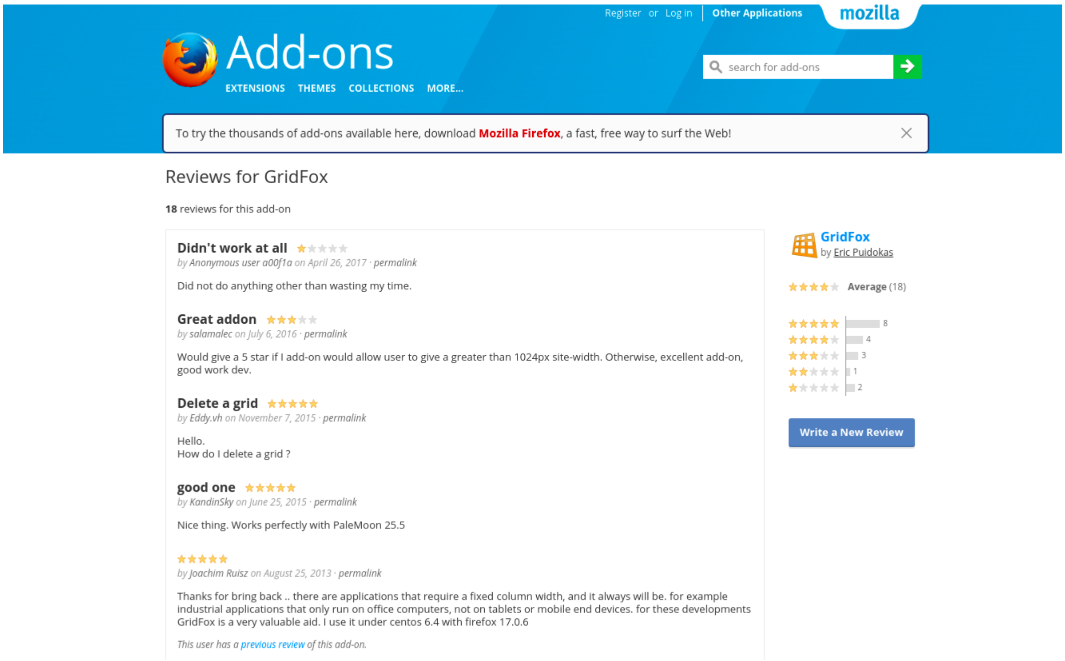
Latest News / Twitter / Facebook / How To Advertise

» **APC Rakes In ₦315m** From **Anambra Governorship Aspirants** «
» **Ogun Acquires 4,500 Hectares Of Land** For **Cotton Production** In **Nigeria** «
» **Twitter Ban: SERAP, 176 Nigerians Drag Buhari** To **ECOWAS Court** «
» **Suspected Motorcycle Thief Set Ablaze** In **Benue (Graphic Photos)** «
» **Police Checkpoints Disappear** From **South-East Highways (Pictures, Video)** «
» **Michael Okpara University Students Protest** Against **Hostile Learning Environment** «
» **Anambra: Our People Deserve Proper Infrastructure** Not **Luxury - Azuka Okwuosa** «
» **Rare 'Ring** Of **Fire' Solar Eclipse** To **Take Place** On **June 10 (Pictures)** «
» **Nigeria** Vs **Cameroon: International Friendly - 0 - 0 (Live)** «
» **I'm Scared** About **My Future** «
» **Aba Made: Governor Ikpeazu Receives US Delegates (Pictures)** «
» **Operation Trojan Shield: 800 Arrested** In **Staggering Global Encrypted Phone Sting** «
» **Don Jazzy Does The 'Jackpot Dance' Challenge** «
» **Nnamdi Kanu, IPOB, The Main Reason Twitter** Was **Banned - Lai Mohammed** «
» **TB Joshua: Pastor Okotie Celebrates** The **Death** Of **'Lucifer', Shares Cryptic Post** «
» **French President Macron Slapped** During **Crowd Walkabout (Video)** «

Figure 3.1: Discussion webpages show several topics that have been posted by users and approved by moderators.

archived [Wikipedia contributors, 2021a]. Discussion forums have an input field and a send button that allow interactivity. They also act as centralized locations for topical discussions involving multiple users, some of which are moderators who may need to approve other users' entries before such entries become publicly visible. Figure 3.1 depicts a forum some examples include `https://whatsapp.com` and `https://discord.com`.

Suggestion/Link Webpages Suggestion Webpages allow users to make an idea or put forward a consideration. This category of web pages usually has sections that display information to the user. They differ from information webpages because suggestion webpages do not provide detailed information; however, they nudge a user to take action based on an idea put forward. Figure 3.2 shows reviews for the GridFox add-on. Examples include `https://www.infoplease.com`



To try the thousands of add-ons available here, download **Mozilla Firefox**, a fast, free way to surf the Web!

Reviews for GridFox

18 reviews for this add-on

Didn't work at all ★☆☆☆☆
by Anonymous user a00f1a on April 26, 2017 · permalink

Did not do anything other than wasting my time.

Great addon ★★★☆☆
by salamalec on July 6, 2016 · permalink

Would give a 5 star if I add-on would allow user to give a greater than 1024px site-width. Otherwise, excellent add-on, good work dev.

Delete a grid ★★★★★
by Eddy.vh on November 7, 2015 · permalink

Hello.
How do I delete a grid ?

good one ★★★★★
by KandinSky on June 25, 2015 · permalink

Nice thing. Works perfectly with PaleMoon 25.5

★★★★★
by Joachim Ruisz on August 25, 2013 · permalink

Thanks for bring back .. there are applications that require a fixed column width, and it always will be. for example industrial applications that only run on office computers, not on tablets or mobile end devices. for these developments GridFox is a very valuable aid. I use it under centos 6.4 with firefox 17.0.6

This user has a previous review of this add-on.

GridFox
by Eric Puidokas

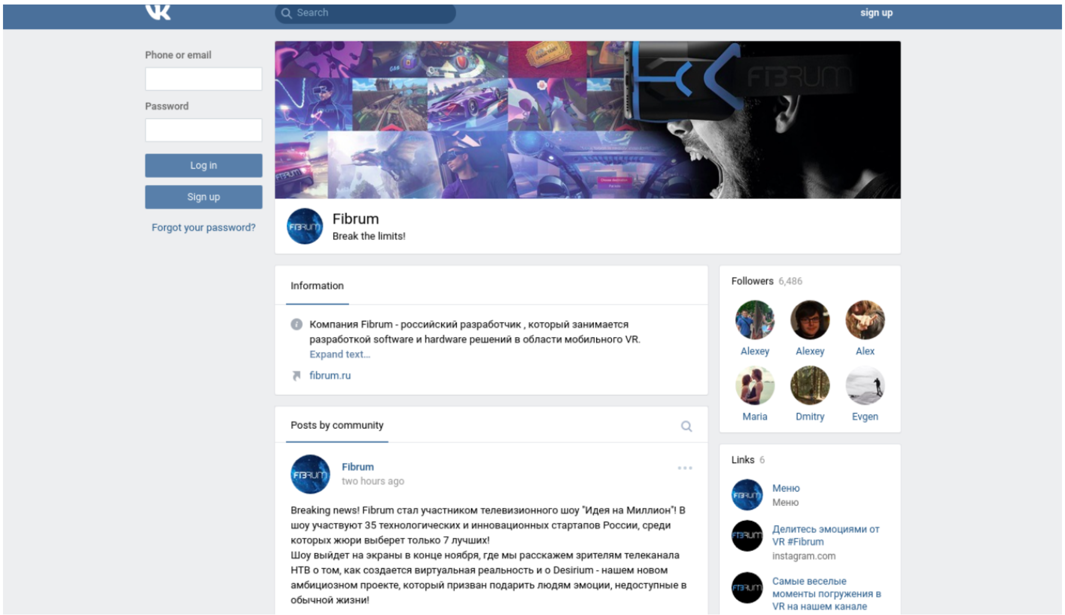
★★★★☆ **Average** (18)

★★★★★	8
★★★★☆	4
★★★☆☆	3
★★☆☆☆	1
★☆☆☆☆	2

Figure 3.2: Suggestion webpage: A review webpage for the GridFox app suggests to users what to expect based on the experiences of previous users. Reviews help to intend users compare their service with other competitors

Information Webpages Information webpages are large web portals, organized as multi-level integration of various resources and services, updated in real-time. Their primary purpose is to provide detailed information about a specific topic [Editorial]. Informational webpages contain enormous amounts of unique content available to users. Compared to suggestion webpages, information webpages have a complex navigation structure that allows easy information access and contains various interactive services. Figure 3.3 is an example of an informational webpage. Examples include weather and news webpages such as `https://www.accuweather.com` and `https://cnn.com`

Form Webpages Form webpage provides users with forms that should be filled out for relevant data collection by allowing user input. It is an interactive page that mimics a paper document or form, where users fill out particular fields [Carter]. The Data collected is then sent to a server for processing. Forms can resemble paper or database forms because web users fill out the forms using checkboxes, radio buttons, or text fields. Figure 3.4 is an example of a form webpage with form fields. In contrast to information websites which mainly give information, forms allow inputs and data collection. An example includes `https://facebook.com/register`



Fibrum
Break the limits!

Компания Fibrum - российский разработчик , который занимается разработкой software и hardware решений в области мобильного VR.
Expand text...

fibrum.ru

Fibrum
two hours ago

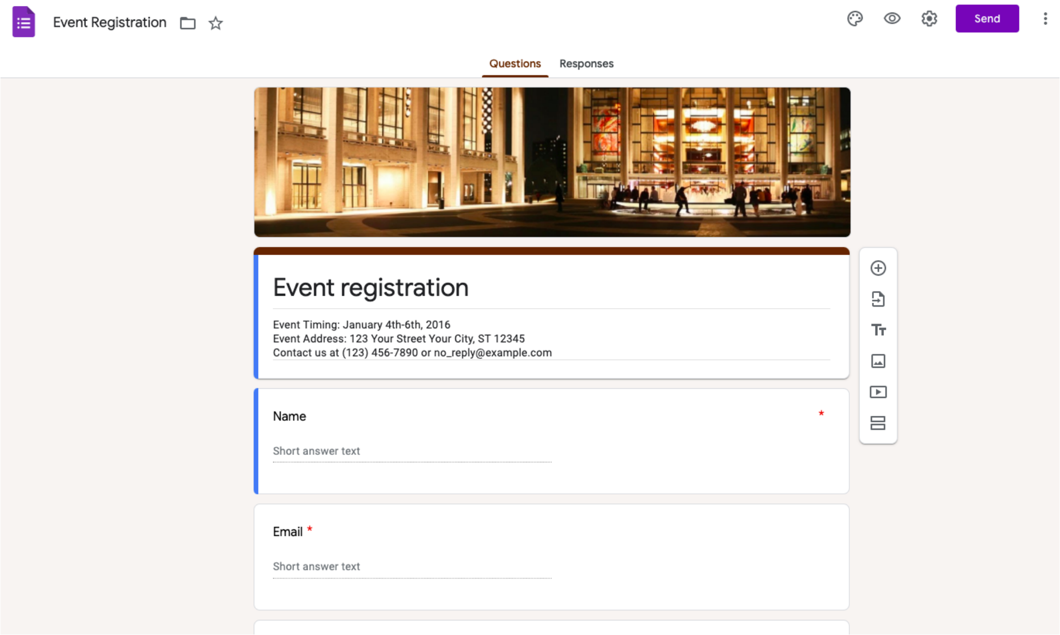
Breaking news! Fibrum стал участником телевизионного шоу "Идея на Миллион"! В шоу участвуют 35 технологических и инновационных стартапов России, среди которых жюри выберет только 7 лучших!
Шоу выйдет на экраны в конце ноября, где мы расскажем зрителям телеканала НТВ о том, как создается виртуальная реальность и о Desirium - нашем новом амбициозном проекте, который призван подарить людям эмоции, недоступные в обычной жизни!

Меню
Меню

Делитесь эмоциями от VR #Fibrum
instagram.com

Самые веселые моменты погружения в VR на нашем канале

Figure 3.3: Information webpages have relatively large segments that display information, this webpage also has a complex navigational structure in comparison with eCommerce webpages



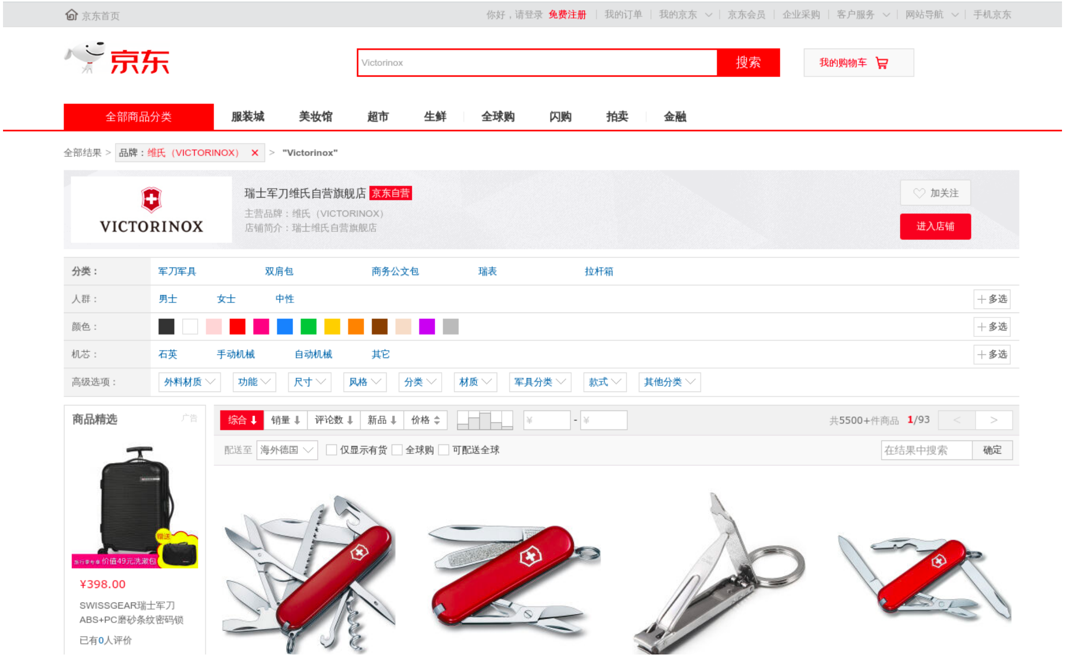
Event registration

Event Timing: January 4th-6th, 2016
Event Address: 123 Your Street Your City, ST 12345
Contact us at (123) 456-7890 or no_reply@example.com

Name *
Short answer text

Email *
Short answer text

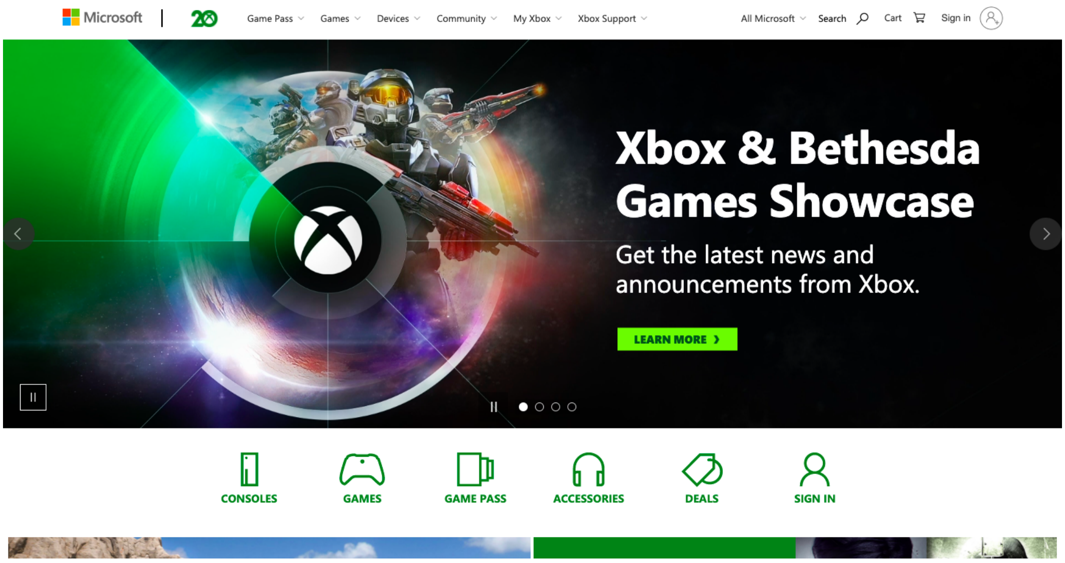
Figure 3.4: Form webpages consist of HyperText Markup Language (HTML) form elements that allow users to enter data. It sends data through a server to a database for further processing



The ecaptionEcommerce webpage features the shopping cart at the top right-hand corner and items to be sold in the main segment of the webpage. This webpage accepts online payments.

eCommerce Webpages E-commerce webpages are used to purchase products or sell via electronic transmission. They are online platforms that provide the means to exchange goods and services for payment. Electronic commerce draws on mobile business, electronic funds transfer, supply chain management, Internet marketing, and online transaction processing. There are several eCommerce business models such as business two business and customer to the business [builder, a]. The key difference between eCommerce and entertainment webpages is that entertainment often embeds video or audio while eCommerce webpages usually feature shopping carts and payment processing. Figure **??** is a Chinese eCommerce webpage. Examples include `https://amazon.com` and `https://ebay.com`

Entertainment Webpages Entertainment webpages provide interactive functionality and content to users by way of live video streaming, video chat communications, multi-player gaming, music, and video streaming, with social networking services such as social graph management, forums, reviews, ratings, and geolocation options [builder, b]. In comparison with information webpages, entertainment engages the user with the purpose of relaxation. Entertainment webpages feature bright images, animation, entertainment infor-



Xbox & Bethesda Games Showcase

Get the latest news and announcements from Xbox.

Figure 3.5: Entertainment webpages often feature bright colors and a photo gallery. There might be links to music or other games on an entertainment webpage.

Table 3.1: Characteristics of webpages with specific functions.

Characteristic	**Discuss**	**Entertain**	**Form**	**Inform**	**Link**	**Sell**
Appearance	Chat Box	Video Frame	Form Fields	News	Address Book	Shopping Cart
Interactive elements	Forums	Audio	Question-naire	IFrames	Anchor Tags	Checkout
Bandwidth Consumption	Midium	High	Low	Midium/Low	Low	High/Medium

mation, interactive chat rooms, online games, photo galleries, drawings, audios, and videos. Moreover, some entertainment also does not relax. Figure 3.5 shows an entertainment webpage. An example is `https://netflix.com` and `https://xbox.com`

Segments on a web page help users determine the function of that web page based on the textual composition and structure of the segments on that web page. It is important to note that webpages can have more than one function since segments give a clue about that web page's function, and one segment can perform more than one function. Also, a typical web page contains more than one segment, allowing users to quickly tell what the function of a web page is by looking at those segments and the composition and structure. A study conducted by Marina Santini [Santini, 2008] to investigate to what extent the classification of a web page by a single genre matches the users' perspective discovered that in an open communication space like the web, phenomena such as genre colonization and genre contamination are likely to occur. Because in the web, many communities meet, each with its genre system and repertoire (cf. Crowston and Williams, 2000).

We have established that users sometimes disagree in their opinion of a webpage function because a webpage often appears to be composite, especially in the visual organization of the spaces between segments, where different communicative purposes and several functions are included at the same time [Santini, 2008]. Furthermore, Johan Dewe et al. [Dewe] studied the difficulty of fitting a web page into a single genre. They found different reasons for that: the web page is multi-genre, without any genre, its genre conventions are unclear, or its genre taxonomy is fuzzy. Multi-genre is evident in our dataset, where annotators were asked to mark segments that influence their webpage function decision. Our annotation result indicates that in some cases, annotators correctly selected different webpage functions for the same webpage. However, webpage functions are more elaborate to web genres in the definition.

Chapter 4

Dataset

In the previous chapters, we discussed detecting webpage functions and webpage function classification systems, automatic webpage function identification, and operationalizing webpage functions. We touched on the similarities and differences between webpage functions and web genres. We also enumerated the six webpage functions that we are focusing on in this thesis. In this chapter, we present our dataset [1], which is the largest publicly available dataset as well as a novel method for obtaining ground truth. The data is achieved by fusing crowdsourced segmentations and post-processing segmentation fitting to DOM nodes. We annotated this dataset using our annotation interface, which is discussed later, after which we present the annotation analysis. How we handled the quality control after the annotation, subsequently we discussed the ground truth.

4.1 Base Datasets

In the course of this thesis, we created Webis-WebFunc-21, which is created from two base datasets, namely Webis-Web-Archive-17 [Kiesel et al., 2018] which comprises a total of 10,000 web page archives from mid-2017 that were carefully sampled from the Common Crawl to involve a mixture of high-ranking and low-ranking web pages. The dataset contains the web archive files, HTML DOM, and screenshots of each web page, as well as per-page annotations of visual web archive quality and Webis-WebSeg-20 [Kiesel et al., 2020] which comprises 42,450 crowdsourced segmentations for 8,490 web pages from the Webis-Web-Archive-17. Segmentations were fused from the segmentations of five crowd workers each. The Webis-WebSeg-20 consists of webpages that were initially 10,000 webpages from the Webis-Web-Archive-17, but this was

[1] https://doi.org/10.5281/zenodo.3354902

reduced to 8490 webpages, because 362 webpages had errors, 50 webpages had loading problems, 15 webpages were blank, and in 2 screenshots, the popup fills the entire screen `https://webis.de/publications.html#kiesel_2019b`. Having omitted the erring webpages, the number of remaining webpages is 9571. Of the 9571 left, 1061 did not have segments and extracted text, so they were omitted from our dataset. It is essential to remove webpages that have errors because our classifier relies on extracted text and segmentation to annotate screenshots of webpages. These segments give further insights into why annotators choose web page functions during annotation in contrast to Webis-WebSeg-20, which only provides information on webpages' segmentation. Thus our dataset (screenshots of webpages) comprises over 42,000 segmentations of the 8,490 webpages from 5,516 websites obtained via high and low ranking as per Alexa. Our dataset is outstanding because several methods have been proposed on webpage segmentation techniques over the past two decades. However, none were sufficient as these methods did not utilize sufficiently large and varied datasets. Also, the segmentations were predominantly Adhoc evaluations performed on small datasets within the context of specific use, which could not produce the results we need to build a good performing classifier [Couvillion, b]. Nevertheless, Webis-WebFunc-21 builds upon Webis-WebSeg-20, which uses a sufficiently large and varied dataset and is not used for a specific purpose.

4.2 Annotation

Dataset annotation is done because a ground truth is required to compare the results from our classifier predictions. Thus segmented webpages need to be annotated before the classification process. In this thesis, we use Amazon Mechanical Turk for annotation. Amazon Mechanical Turk (AMT) is a marketplace that helps companies and professionals alike to outsource various types of virtual processes and tasks, also known as Human Intelligence Tasks (HITS), to a distributed workforce. These tasks cannot be automated or carried out by bots because critical thinking is required. Crowdsourcing using AMT allowed us access to a talented workforce which are cheap and fast to evaluate systems and provide categorical annotations for our training data. In addition, there were no management overheads or resource allocation challenges during the process as the HITS were done remotely. Other benefits include indefinite scalability, and this allowed us to annotate the dataset in large batches after crosschecking to ensure we get the desired results from small batches.

In this thesis, we hired AMT users to mark the function of webpages by choosing segments to inspire that choice. This process of marking segments of a webpage to indicate its function is known as an annotation. There are six webpage functions that annotators can choose from, which we discussedchapter 3 In. The annotation task asks AMT users to select segments of each webpage screenshot that fulfill the six webpage functions as its primary function. We used bounding boxes to marked visually to allow the annotator to distinguish each web page segment boundary intuitively when they mouse over it. In summary, for each webpage, an annotator needs to mark which segments make that function as mainly. These webpage functions include; Discussion, Entertainment, Information, Forms/Data collection, Selling, and Linking/Suggestion.

This Crowd-labeling process collected annotations from AMT users and used them for estimating consensus values. However, inattentive annotators and spammers reduce the quality of consensuses. Therefore, we observed the annotators' behavior early on in the annotation process to improve the quality of consensuses.

The procedure for annotation is as follows. First, we created and uploaded 10 HITS with five pages each, making a total of 50 assignments in order to determine if annotators understood our segmentation tool. Figure 4.1 shows the process of creating a Human Intelligence Task on AMT. The first step is building the Human Intelligence Task HIT, which entails the task we want to achieve robots cannot do that. Annotation is the task of deciding the primary and not function of a web page. Next, we need to test the HIT to ensure it is bug-free and understood by AMT users. After testing, the task is posted on AMT publicly, which can be searched and accepted. After the acceptance, AMT users can now go ahead annotating webpages, after which the task is submitted when the process is complete. Administrators can now accept or reject HITS based on the suggestion of MACE and our acceptance and rejection guild lines. Twenty-seven workers annotated the test Batch. Then we followed with 90 HITS after it was evident the initial batch was well understood, with a total of 450 assignments which 147 workers annotated within a few days. We then did five more batches using 627 workers to annotate 1914 HITS to make up 8490 assignments. It took about 2.5 minutes on average to finish one HIT. The total time spent on annotation is 71 hours. Furthermore, each worker received financial rewards for their task. Good performing workers who understood the task were assigned more HITS. In contrast, annotators who did not understand the task or did it wrongly were rejected. We paid 0.25 USD for a HIT containing six web pages (including 1 test case, so actually five web pages for our purposes). So we get one web page annotation for 0.05 USD. In total, we paid 2,466.90 USD at a rate of 6.7 USD per hour. The following subsection discusses our annotation interface and the various decisions an annotator can make while building the ground truth.

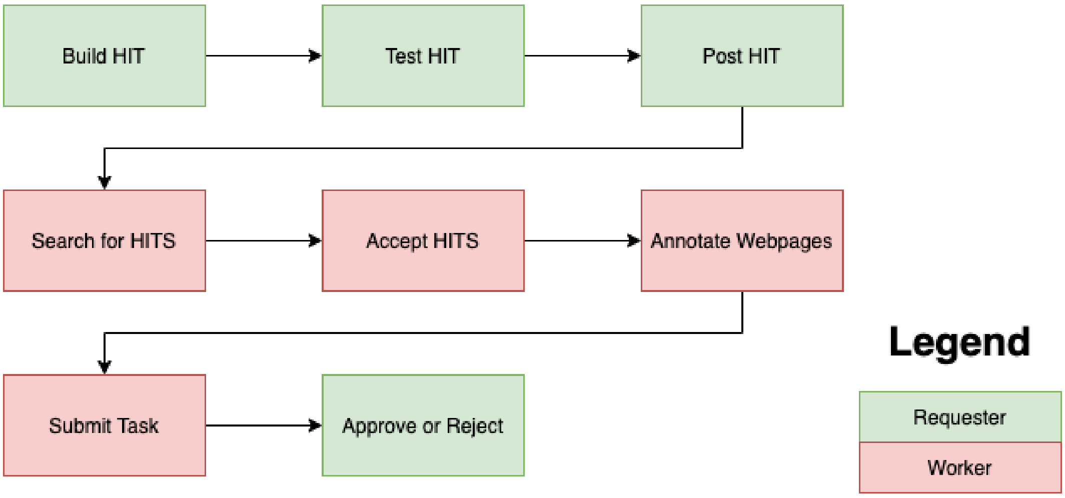
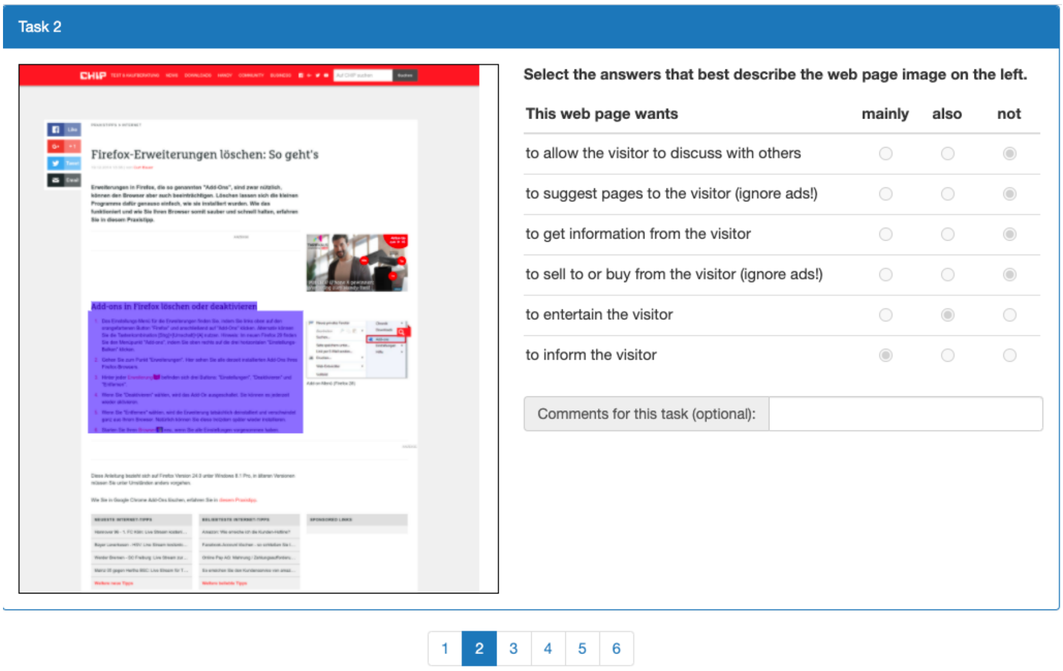


Figure 4.1: Crowd sourcing process diagram shows the steps we took from building a HIT to approving or rejecting the same HIT

4.2.1 Annotation Interface

In our annotation interface, MTURK users can choose between six functions as mainly, also, or not functions. These functions include; Ecommerce, Suggestion/Linking, Information, Entertainment, Discussion, and Data collection. To make a function, each annotator has to choose either mainly, also, or not. This choice represents that a webpage has its primary purpose as the chosen webpage function if the choice is ’mainly’ and the webpage is also another function if the option is ’also.’ In contrast, the function is not that purpose if a webpage is ’not’ that function. The annotation interface allows an annotator to mark what segment on each web page influenced their decision to choose if the function that webpage is mainly, or not anyone one of the six web page functions. These transparent bounding boxes appear in red when an annotator mouses over a segment of the webpage to be annotated. Once they click the segment, the segment’s color changes to blue, indicating they have decided on that segment because they choose the webpage function. The selected segment can be clicked again to remove the blue color if the AMT user changes their mind about that segment. One or more segments can be highlighted on each webpage when more than one segment influences their decision. Figure 4.2 shows a highlighted section from an annotator. Who has chosen only one segment as the reason why they choose the webpage as an Information webpage.



Task 2

Firefox-Erweiterungen löschen: So geht's
Add-ons in Firefox löschen oder deaktivieren

Select the answers that best describe the web page image on the left.

This web page wants	mainly	also	not
to allow the visitor to discuss with others	○	○	●
to suggest pages to the visitor (ignore ads!)	○	○	●
to get information from the visitor	○	○	●
to sell to or buy from the visitor (ignore ads!)	○	○	●
to entertain the visitor	○	●	○
to inform the visitor	●	○	○

Comments for this task (optional):

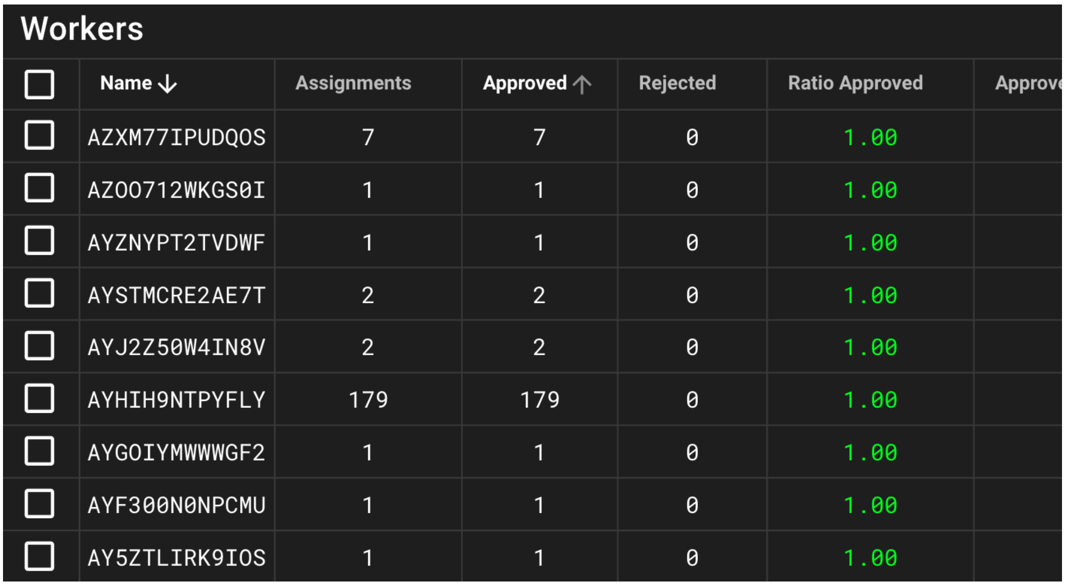
Figure 4.2: The highlighted blue segment in the middle left shows the segment the annotator chooses as the reason for marking this webpage as an Information webpage

4.2.2 Annotation Analysis

In crowdsourced annotation tasks, some annotators deliberately choose the wrong webpage functions or mark segments randomly without giving enough thought to the instructions of the task in order to maximize their pay by supplying quick answers. Furthermore, manual identification of each wrong or hastily done annotation is tedious. Therefore, we need an automated tool to detect fraudulent annotations accurately. To achieve this, we use Multi Annotator Competence Estimation MACE, which is an item-response model, to monitor redundancy in annotations [Hovye, 2013]. First, MACE identifies trustworthy annotators, then determines the correct underlining webpage function. Trustworthy annotators are then upgraded and allowed to annotate more webpages. MACE helps us determine the annotator's actions while deciding to either approve or reject a HIT. Figure 4.3 shows the number of approved annotations that help determine trustworthy annotations.

4.2.3 Quality Control

Providing ground truth labels for large datasets is often time-consuming. That is why we outsourced the annotation process of our relatively large dataset.



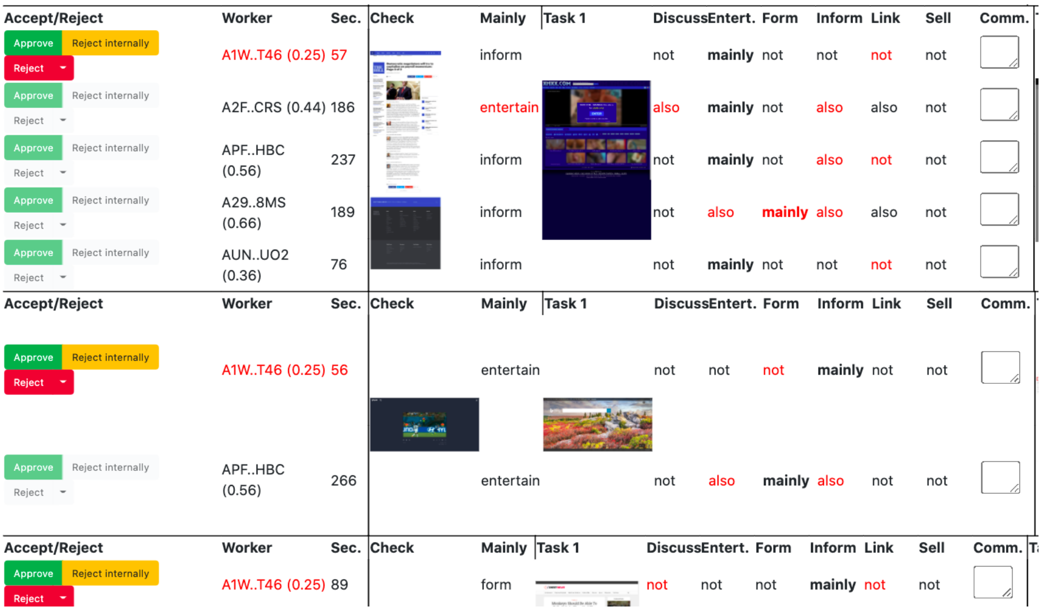
Workers

□	Name ↓	Assignments	Approved ↑	Rejected	Ratio Approved	Approve
□	AZXM77IPUDQOS	7	7	0	1.00	
□	AZOO712WKGS0I	1	1	0	1.00	
□	AYZNYPT2TVDWF	1	1	0	1.00	
□	AYSTMCRE2AE7T	2	2	0	1.00	
□	AYJ2Z50W4IN8V	2	2	0	1.00	
□	AYHIH9NTPYFLY	179	179	0	1.00	
□	AYGOIYMWWWGF2	1	1	0	1.00	
□	AYF300N0NPCMU	1	1	0	1.00	
□	AY5ZTLIRK9IOS	1	1	0	1.00	

Figure 4.3: Multi Annotator Competency Estimator MACE helped us determine which worker performed better than others. A worker with 179 assignments and 179 approved shows a worker who understands the task

However, employing expert annotators is expensive. More so, Crowd-sourcing the annotation process is a cost-effective and fast method for annotation, especially when expertise is not necessarily required or can be quickly acquired by following instructions. Using a multi-annotator competency estimator, which we discussed earlier, we distinguished good annotators from bad ones.

Because we rely on annotated data to train models, this implicitly assumes that the annotations represent the truth. However, this basic assumption can be violated in two ways: either because the annotators exhibit a particular bias (label bias) or because there is no single truth (bias in ground truth). To estimate an annotator's competency, we count inter-annotator agreement and calculate the harmonic mean's reciprocal. Workers which MACE flags as wrong on more than three webpages are rejected; up to 2 webpages wrong are rejected internally, rejected internally were annotators who understood the task but did not correctly select at least 3. The rest which does not fall into any of these categories was accepted were accepted. That is why we rejected 4 HITS out-rightly, rejected ten internally, and accepted the rest of 1900 HITS. 4.4 shows MACE implementation that allows easy detection, acceptance or rejection of users who disregard annotation instruction. Also, we considered annotation time because annotators who understood the task spent more time than annotators that did not. Annotators are expected to engage in deeper information processing when they make choices during the



Accept/Reject	Worker	Sec.	Check	Mainly	Task 1	Discuss	Entert.	Form	Inform	Link	Sell	Comm.
Approve Reject internally Reject	A1W..T46 (0.25)	57		inform		not	mainly	not	not	not	not	
Approve Reject internally Reject	A2F..CRS (0.44)	186		entertain		also	mainly	not	also	also	not	
Approve Reject internally Reject	APF..HBC (0.56)	237		inform		not	mainly	not	also	not	not	
Approve Reject internally Reject	A29..8MS (0.66)	189		inform		not	also	mainly	also	also	not	
Approve Reject internally Reject	AUN..UO2 (0.36)	76		inform		not	mainly	not	not	not	not	

Accept/Reject	Worker	Sec.	Check	Mainly	Task 1	Discuss	Entert.	Form	Inform	Link	Sell	Comm.
Approve Reject internally Reject	A1W..T46 (0.25)	56		entertain		not	not	not	mainly	not	not	
Approve Reject internally Reject	APF..HBC (0.56)	266		entertain		not	also	mainly	also	not	not	

Accept/Reject	Worker	Sec.	Check	Mainly	Task 1	Discuss	Entert.	Form	Inform	Link	Sell	Comm.
Approve Reject internally Reject	A1W..T46 (0.25)	89		form		not	not	not	mainly	not	not	

Figure 4.4: The image shows the interface used in deciding workers to accept and reject. The interface implements Multi Annotator Competency Estimator

annotation process [Couvillion, a]. Thus the artificial imposition of sections for better genre annotation increased workers' cognitive load and higher arousal level.

4.2.4 Ground Truth Creation

Ground truth refers to the reality we want our model to predict. It is the accuracy of the training set's classification for supervised learning techniques such as decision trees. The ground truth enables us to perform quality control; therefore, we create it by creating Human Intelligence Tasks (HITS), which AMT users then do. There are numerous approaches to Ground truth Creation, such as synthetic labeling, internal labeling programs, and crowdsourcing, amongst others.

In this thesis, we used crowdsourcing, as discussed earlier. Each webpage has one 'main' function, another 'also' function, and a 'not' function to make its function unambiguous. Annotation was carried out by AMT annotators who used our annotation tool to select sections on a webpage that support their decision of the function of a webpage. It is important to compare the performance of our model to the performance of the ground truth dataset. Below we show the level of accuracy between our model and the ground truth. The bar chart shows the level of similarity between our model and the ground

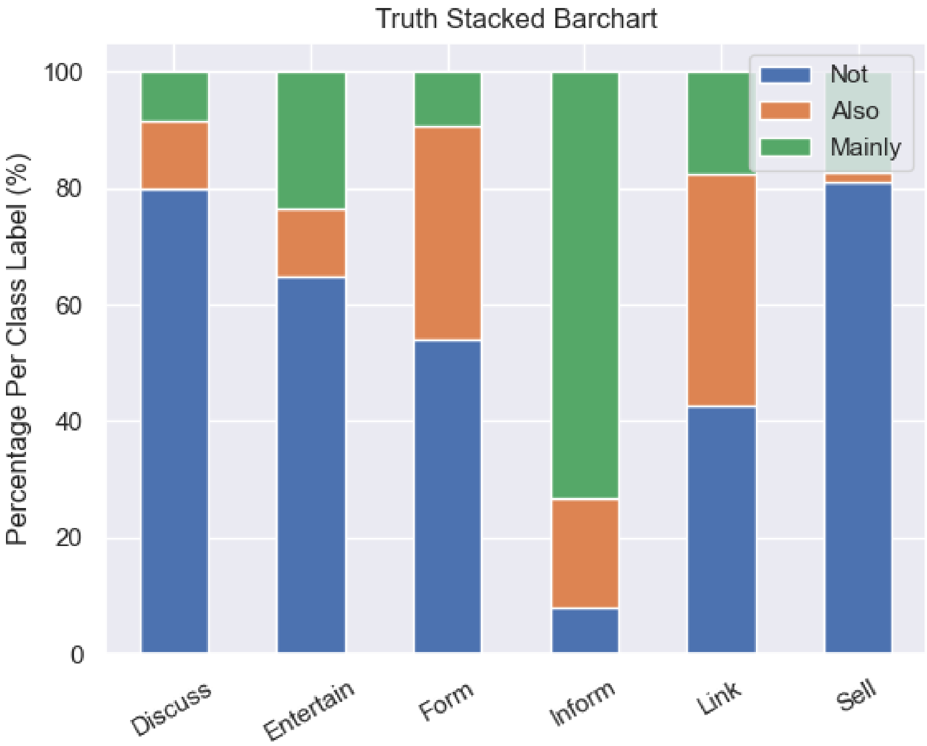


Figure 4.5: Ground truth bar chart which shows the distribution of the number of webpages per function

truth. Figure 6.1 shows the ground truth labels for each webpage function, while Figure 6.2 shows the prediction of our model.

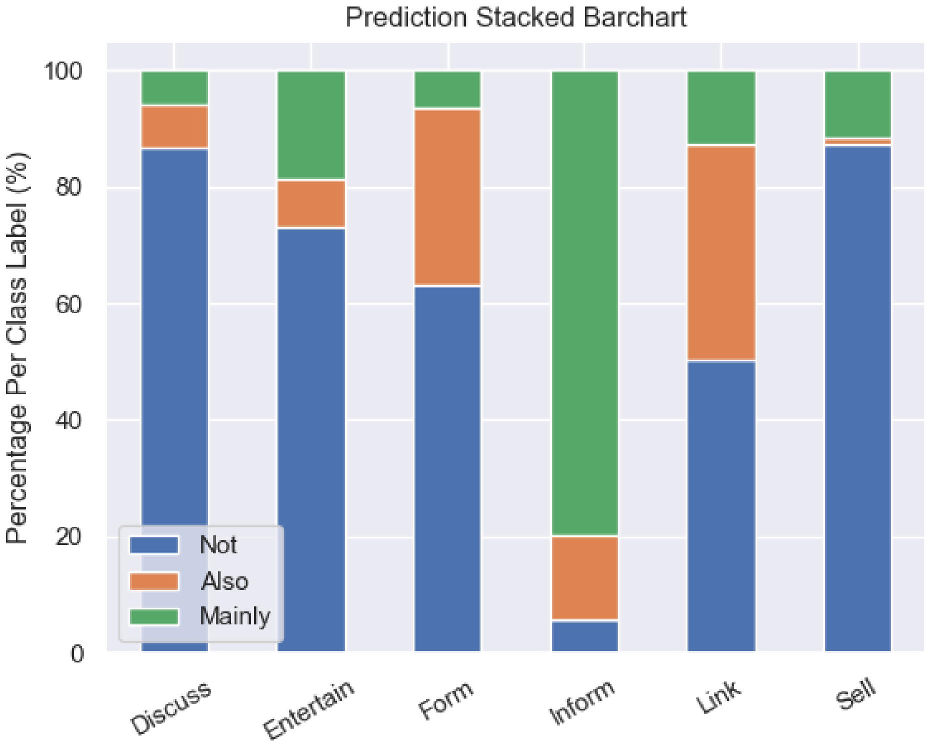


Figure 4.6: Our prediction bar chart shows the number of the webpages predicted for each function

Chapter 5

Experiments

In chapter 5, we present experiments done on our machine learning classifier, which predicts webpages' function automatically.

I took four steps in building the classifier. The first step is to utilize the cross folds validation technique with a split of web pages into five folds. These folds are contained in a comma-separated value (CSV) file containing each web page ID and corresponding fold number. By separating the webpages into test and train data based on their fold number. I used some webpages to train and the rest to testing, thereby predicting the label of webpages after training. The splitting of test and train data was done by a ratio of 2 to 8, respectively. Then for each webpage, I extracted the text contained therein and did the following preprocessing step.

The second step is preprocessing, which has two steps in itself. The first preprocessing step is the removal of stop words. The second preprocessing step is tokenization, which was achieved by separating each word into character tokens.

The third step I took in building the classifier was to convert each character token into vectors. [1]. Since webpages usually have more than one word, I found the average vector of each webpage token by summing up all vectors in each webpage and dividing it by the number of tokens contained in that webpage. The average vector is saved in a list to be used for training and testing.

Now having each webpage's average vector as a list and inserting each of these lists into another list, we have a list containing lists of average webpage vectors. It is necessary to have vectors of webpages in the list of lists because we use the exact representation for our ground truth. Thus, for each webpage, we create a list for our ground truth and insert each webpage's corresponding ground truth as a list which was discussed in chapter 4. In the annotation, we represent the label 'mainly' with the number '2', 'also' with the number '1'

[1] `https://www.datacamp.com/community/tutorials/python-scipy-tutorial`

and 'not' with the number '0' because our classifier can compute numbers and not words. Therefore we now have two lists of lists. One list has the vectors, and the other has the ground truth.

Because our ground truth has mainly, and not, we are not solving binary classification problem statements. In binary classification, any of the samples from the dataset takes only one label out of two classes, but we have three classes. In this thesis, the ground truth of each webpage can belong to either mainly, also, or not.

5.1 Features

In order to get the features of webpages, I followed the steps enumerated above to make vectors(weights) from the dataset. This weight measures the importance of an index term in our webpage based on the word2vec algorithm. Word2vec is a technique in natural language processing that uses a neural network to learn word associations from a corpus of text. I used it to learn a representation of words for our random forest classification task.

5.2 Random Forest classification

Learning to classify webpages can be achieved using any of the machine learning techniques. Supervised machine learning algorithms can be broken down into two distinct types, namely regression, and classification. Regression seeks to predict a continuous property, while classification is used to predict a particular label, as in our use case. There are several classification algorithms to choose from, such as logistic regression, naive Bayes, decision trees, random forests, K-Nearest Neighbors Algorithm, and Support Vector Machine Algorithm. I used random forest because it is a forest of decision trees splitting with oblique hyperplanes, ensuring the model gained accuracy as it grew without suffering from over-training. After all, the forests are randomly restricted to be sensitive to only selected feature dimensions [2], with the most trees predicted as the correct class. Finally, our ground truth is neither numerical nor ordinal because our label 'mainly' is not twice as much as 'also.' Therefore, we can not put a ratio on our labels because they are non-non-deterministic.

In classification problems, over-fitting is the generalization of unseen data, and this occurs when training data fits the testing data. I mitigated overfitting by doing a training and testing dataset split to avoid over-fitting, especially since the training dataset is susceptible to over-fitting [Wikipedia], due to

[2] [Wikipedia]

our relatively few webpages of 8686 in total. Using the fold cross-validation method, I had 6794 webpages for training and the remaining 1692 webpages for testing our random forest classifier.

Fixed splitting the train and test dataset was not enough to check overfitting because a random split might be biased, so I split the train and test data in different folds based on the statistical method called cross-validation to estimate my performance classifier. There are two methods used for cross-validation: either the non-exhaustive method that does not compute all methods of webpage splitting or the exhaustive method, which computes all ways of splitting our original webpages that I used. My approach is a K - 2 folds cross-validation technique that implements the holdout method by randomly shuffling before splitting the train and test webpages [3].

5.3 Evaluation Setup

k cross-validation procedure has a single parameter called k that refers to the number of groups that the webpages be split into. For each web page function, the k - 2 cross-validation method helped me verify that my train and test web pages did not negatively impact the predictions. The confusion matrix for the prediction showed relative consistency in classifying classes for each webpage function, including the case when I binarized the dataset by merging 'mainly' and 'also' into the same label. Figure 5.3 shows that even though we collapse 'mainly' and 'also' into one label, the prediction of eCommerce webpages is quite similar to our multi-label prediction as shown in Figure 6.4e (eCommerce). The confusion matrix from cross-validating prediction outcomes by splitting the train and test data and comparing the performance of our test webpages we get an average figure as shown in Figure 5.2.

The F score is used to measure a model's accuracy in binary classification problems [Sasaki, 2007]. It is calculated from the precision and recall of the prediction. Precision and recall are performance metrics applied to data retrieved from a collection. Precisely, the F score is the harmonic mean of precision and recall. In the prediction output, the highest value is 1.0 indicating a perfect score of both precision and recall, while the lowest is 0. For each fold of the cross-validation, I computed the f score, and the average f score in the random forest model is 0.75.

The standard deviation is the average amount of variability in our dataset. It tells on average how each value deviates from the mean. The standard deviation and the mean together tell us where most of the values in our distribution

[3] `https://www.mygreatlearning.com/blog/cross-validation/`

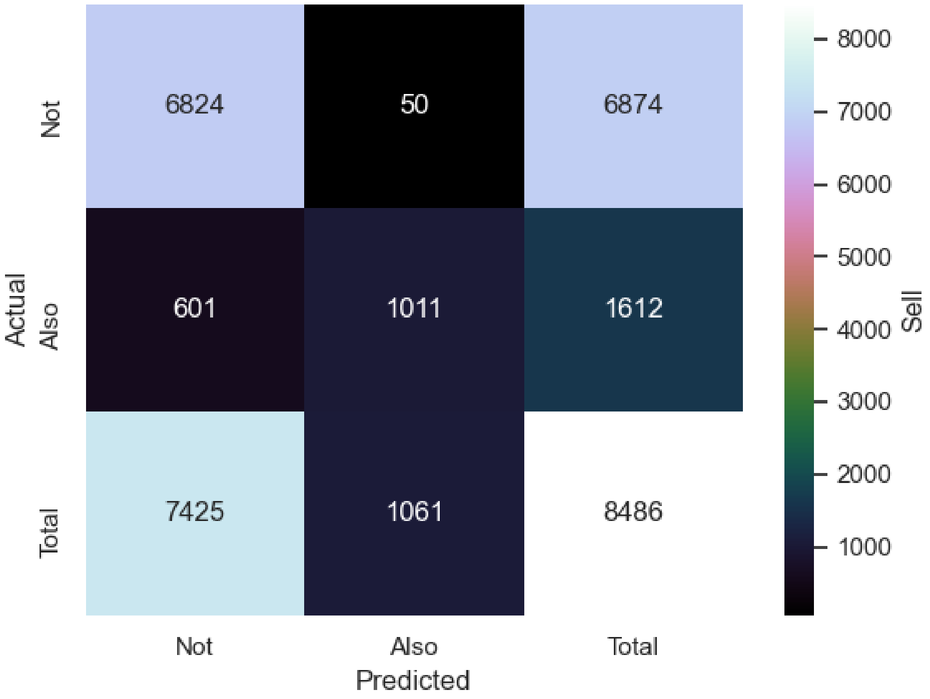


Figure 5.1: Binarized classes for eCommerce webpages confusion matrix

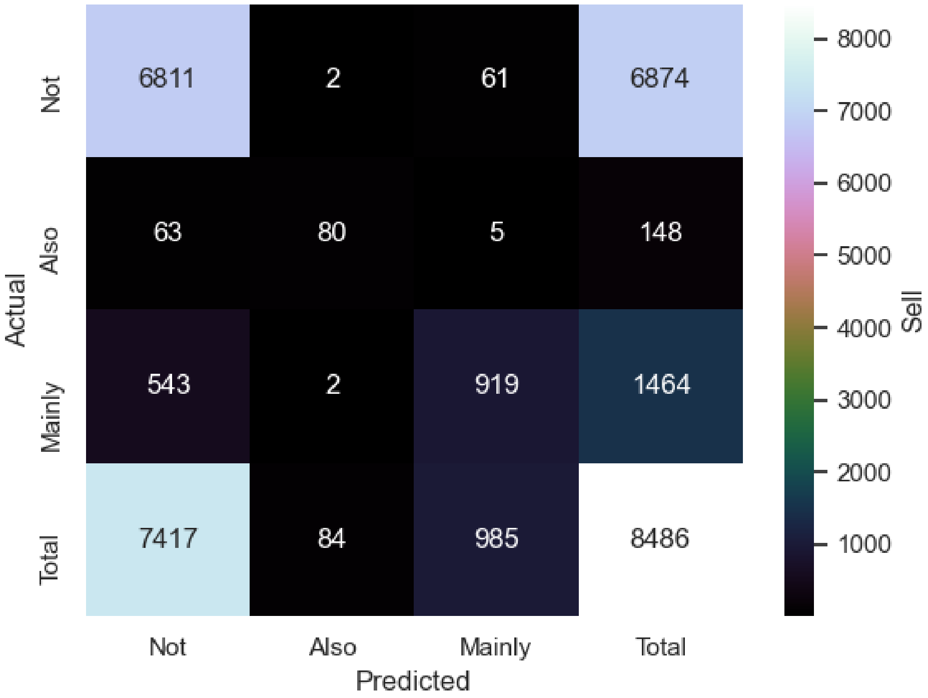


Figure 5.2: Average prediction for eCommerce webpages confusion matrix shows that our classifier is not over-fitted

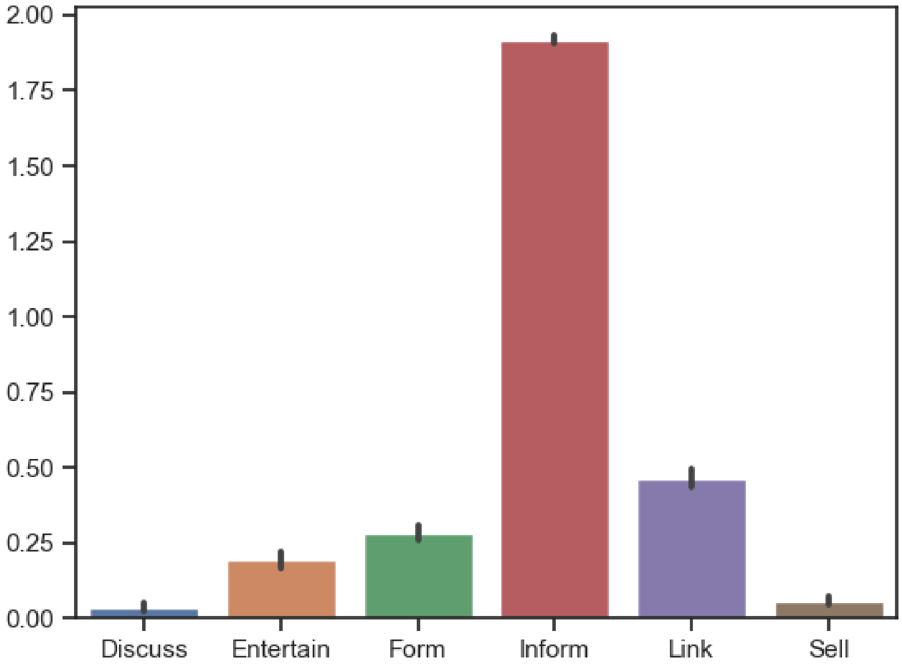


Figure 5.3: This figure shows a low standard deviation indicates that values are clustered close to the mean, while a high standard deviation means that values are generally far from the mean. On the Y-axis, 0 indicates not, and 2 indicates mainly; thus, the higher the bar chart, the more prediction the 'mainly' label.

lie since they follow a normal distribution as is the case with our results [4]. The figure 5.3gives more insights into the performance of our classifier.

[4] https://www.scribbr.com/statistics/standard-deviation

Chapter 6

Results

The chapter 5 discusses our experimental setup, features used in this thesis, random forest classification as well as evaluation setup. In Chapter6, we will present the results of our experiments, overall performance, and findings. Furthermore, we revisit our research question of annotating the largest publicly available dataset using a tool that considers the multi-class and multi-label format of our webpages.

6.1 Ground Truth vs. Prediction

This thesis focused on the classification of webpages in order to automatically predict what their function is. For this purpose, we used the segmentation technique of webpages to annotate our ground truth from the Webis Web Segments 2020 dataset. This dataset was created using a novel approach of crowdsourcing segmentations. Using this dataset, a random forest classifier was built to solve the challenge of automatic webpage identification, which can now be used to predict genres of webpages. The classifier predicted correctly that most of the webpages are Information webpages followed by Link/Suggestion webpages. Entertainment and form/data collection webpages have a similar number of webpages as mainly while discussion webpages are the least predicted. 80% of information webpages are mainly while 18% of entertainment webpages are mainly, 15% each of both Link/Suggestion and eCommerce webpages then 4% of discussion and 5% of Form/Data entry webpages are mainly. Figure 6.2 shows the percentage of predictions of webpages. Additionally, in Figure 6.4 we show all webpage functions and the confusion matrix of their predictions. These predictions indicate the high performance of the classifier.

The result of predictions was compared to the ground truth annotation as shown in Figure 6.2. Given that the dataset is multi-label, i.e., 'mainly,' 'also,' and 'not' and not simply binary such as 'mainly' and 'not,' we had non-binary

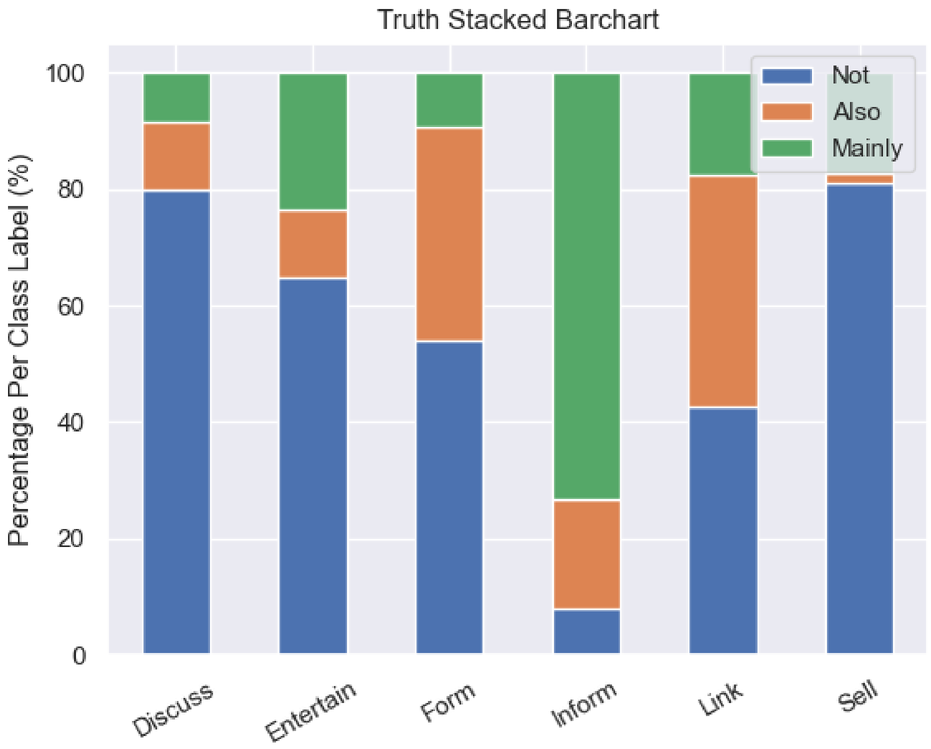


Figure 6.1: The ground truth bar chart shows the distribution of the different webpage functions and the percentage of mainly also and not based on our annotation data

labels for each webpage. This multi-label feature posed additional complexity to the classification task. A less complex configuration is binarizing the labels to either 'mainly' or 'not'; however, we would lose the information that 'also' carries.

Subsequently, the results of the multi-class multi-label classifier shown in Figure 6.2 was compared with the results of a binarized classifier shown in 6.3, and I discovered that the merging of 'mainly' and 'also' into one class has no impact on the prediction of the model. However, a classifier with multi-label attributes performs better due to the loss of information when binarizing the labels. This binarization was done by making all classes in the ground truth that are 'mainly' and 'also' as true (1) while all classes that are 'not' as false (0). Figure 6.3 shows a stacked bar chart with the prediction results for all webpage functions. Information webpage has the highest percentage of 'mainly' and 'also' with 96% followed by Linking/Suggestion webpages with 62%. Then Forms/Data Collection webpages also have 38% as 'mainly' and 'also.' On the other hand, Selling webpages has the lowest percentage of 15% and Discussion webpages have 16%. In contrast, Entertainment webpages have 29% as 'mainly' and 'not.'

6.2 Baseline Comparison

Firstly, the result was compared with a baseline Zero Rule algorithm, a benchmark procedure for classification algorithms. The output of this algorithm is the most frequently occurring webpage function in our webpages. For example, if 55% of webpages are Informational, the Zero rule would predict all webpages have a 55% chance of being informational and would be correct 55% of the time. Random forest classification performs better than the Zero rule baseline because of its use of decision trees for prediction [msg, 2020].

Secondly, the result was compared with the decision stump algorithm, a decision tree consisting of a single fork node with two outcomes. When the model is run against a model that includes several predictor variables, the decision stump node chooses the single predictor variable that most accurately enables the choice between the two alternatives and ignores everything else. On the other hand, the One Rule is a simple yet accurate classification algorithm that generates one Rule for each predictor in the data, then selects the Rule with the slightest total error as its "one rule." Also, our model performs better than the decision stump algorithm.

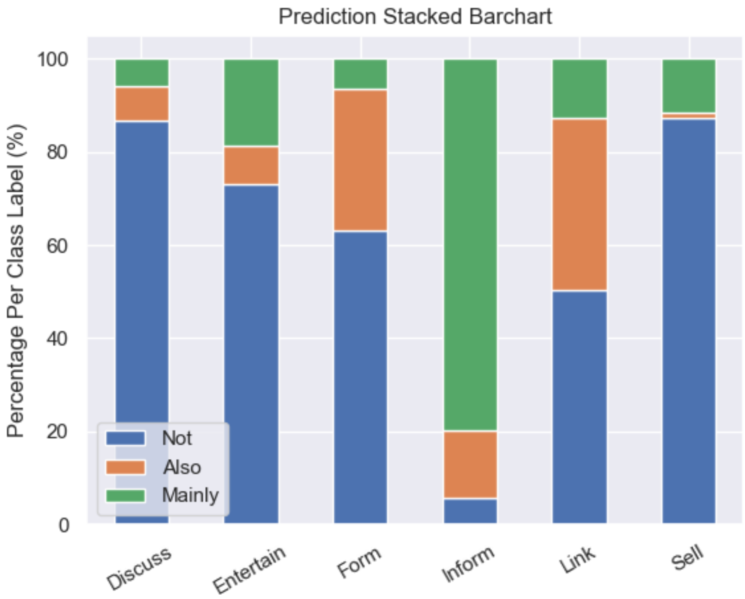


Figure 6.2: The prediction bar chart shows the distribution of the different webpage functions and the percentage of mainly also and not

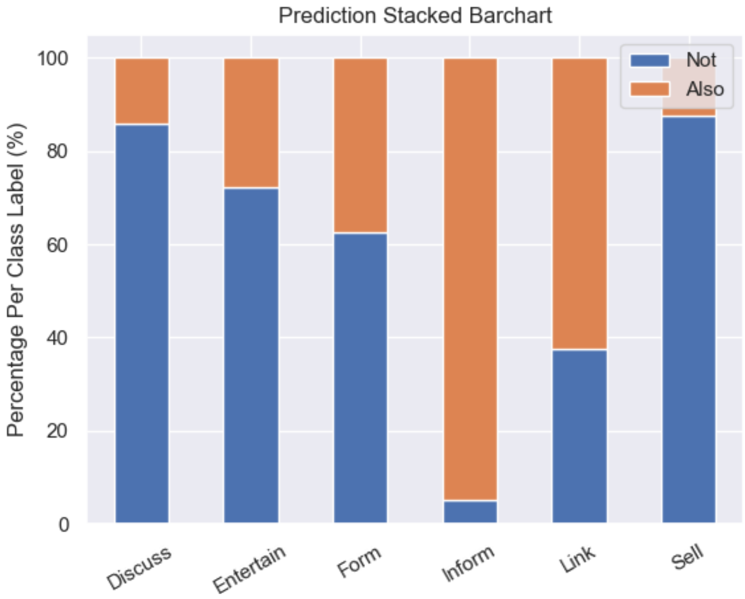
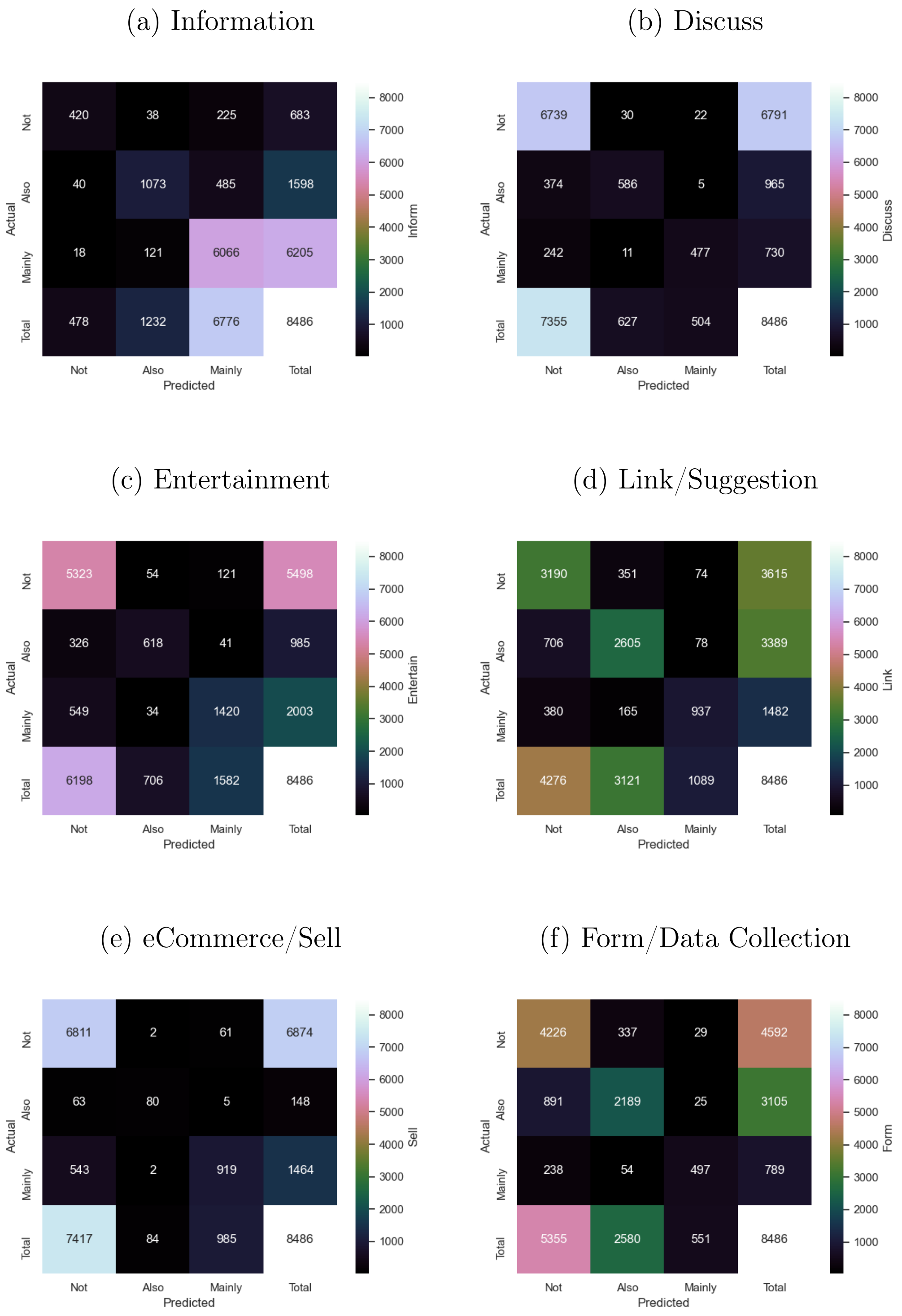


Figure 6.3: Binarized dataset prediction bar-chart shows that the prediction is similar



(a) Information

(b) Discuss

(c) Entertainment

(d) Link/Suggestion

(e) eCommerce/Sell

(f) Form/Data Collection

Figure 6.4: Webpage confusion matrices for all functions. The functions include, Information, Discussion, Entertainment, Link/Suggestion, eCommerce/Sell and Form/Data Collection

Chapter 7

Conclusion

In chapter 3 I presented the six different functions (classes) a webpage can perform namely Discussion, Information, Linking/Suggestion, eCommerce, Form/Data Collection, and Entertainment webpage functions. Then I discussed the annotation tool used for collecting annotations based on these functions in 4. Furthermore, I analyzed the predictions, and we know that annotating the largest publicly available segmented webpages was successful. In chapter 5, I presented the steps taken to achieve automatic webpage function detection and the evaluation of my classifier. Then in chapter 6, I present the results of the thesis and its comparison to the ground truth. I also compared the prediction with the ground truth in one section and a random baseline in another.

In conclusion, we have seen that genre classification works by using a segmentation tool to annotate data using Amazon Mechanical Turk. Our classifier performs well, as shown in our predictions in the figure6.4.

In summary, we built a classifier to detect the function of the webpages automatically.

In this thesis, the prediction results were outstanding, but improvements can be made to improve the quality of our prediction, which are discussed in the following paragraphs.

It improved the features of this classifier by adding the Document Object Model (DOM) information for each word extracted from webpages. The DOM is the data representation of objects that comprise the structure of a webpage and its words. This feature will improve prediction results because the DOM represents the tree structure content of a webpage. DOM information would be an exciting feature to explore.

Secondly, the use of segment position as another feature will improve predictions of webpages. The segment position is captured during annotation. This feature is essential because Webpages with similar functions will have similar layouts, while webpages with different functions will have different layouts.

Bibliography

Daisy Abbott and Yunhyong Kim. Genre classification. 2008. URL https://www.dcc.ac.uk/guidance/briefing-papers/introduction-curation/genre-classification. 2.2

Hugues Boisvert. Benchmarking web site functions. URL https://www.emerald.com/insight/content/doi/10.1108/14635770610644664/full/html?casa_token=xOy9H9NpAngAAAAA:aWPYR22CsWUMnemx6cxDbNznucsKPu4xUSWLoC3w-fOXyLZuG9FCdml5X1EMdX6odhqqEYwlqrEkW5K0wZUkwbvpyD-u5O92_TQqSHv2GS76U04HXQ. 3

Website builder. Learn about online business, a. URL https://zyro.com/learn/ecommerce-website/. 3

Website builder. Learn about online business, b. URL https://zyro.com/learn/ecommerce-website/. 3

Natashia Carter. Web forms. URL https://askinglot.com/what-is-the-difference-between-web-forms-and-web-pages-in-asp-net. 3

Kaylee F Couvillion. Increased cognitive load during acquisition of a continuous task eliminates the learning effects of self-controlled knowledge of results, a. URL https://pubmed.ncbi.nlm.nih.gov/31648607/. 4.2.3

Kaylee F Couvillion. Increased cognitive load during acquisition of a continuous task eliminates the learning effects of self-controlled knowledge of results, b. URL https://webis.de/downloads/theses/papers/meyer_2020.pdf. 1, 4.1

Johan Dewe. Assembling a balanced corpus from the internet. URL https://aclanthology.org/W98-1611.pdf. 3

MonsterPost Editorial. What is an informational website and how is it built? URL https://www.templatemonster.com/blog/informational-website/. 3

Aidan Finn, Nicholas Kushmerick, and Barry Smyth. Genre classification and domain transfer for information filtering. In Fabio Crestani, Mark Girolami, and Cornelis Joost van Rijsbergen, editors, *Advances in Information Retrieval*, pages 353–362, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg. ISBN 978-3-540-45886-9. 2.1.1

Dirk Hovye. Learning whom to trust, 2013. URL `https://aclanthology.org/N13-1132.pdf`. 4.2.2

A. Kennedy and M. Shepherd. Automatic identification of home pages on the web. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, pages 99c–99c, 2005a. doi: 10.1109/HICSS.2005.114. 2.2, 2.4, 2.4

A. Kennedy and M. Shepherd. Automatic identification of home pages on the web. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, pages 99c–99c, 2005b. doi: 10.1109/HICSS.2005.114. 2.1.1

Marie Williams Kevin Crowston. Reproduced and emergent genres of communication on the world wide web. *The Information Society*, 16(3):201–215, 2000. doi: 10.1080/01972240050133652. URL `https://doi.org/10.1080/01972240050133652`. 2.4

Johannes Kiesel, Florian Kneist, Milad Alshomary, Benno Stein, Matthias Hagen, and Martin Potthast. Reproducible Web Corpora: Interactive Archiving with Automatic Quality Assessment. *Journal of Data and Information Quality (JDIQ)*, 10(4):17:1–17:25, October 2018. doi: 10.1145/3239574. URL `https://dl.acm.org/doi/10.1145/3239574`. 4.1

Johannes Kiesel, Florian Kneist, Lars Meyer, Kristof Komlossy, Benno Stein, and Martin Potthast. Web Page Segmentation Revisited: Evaluation Framework and Dataset. In Mathieu d'Aquin, Stefan Dietze, Claudia Hauff, Edward Curry, and Philippe Cudré-Mauroux, editors, *29th ACM International Conference on Information and Knowledge Management (CIKM 2020)*, pages 3047–3054. ACM, October 2020. doi: 10.1145/3340531.3412782. URL `https://dl.acm.org/doi/10.1145/3340531.3412782?cid=99659134098`. 4.1

Barbara H. Kwasnik. Identifying document genre to improve web search effectiveness. the bulletin of the american society for information science and technology. URL `https://surface.syr.edu/cgi/viewcontent.cgi?article=1133&context=istpub`. 2

Barbara H Kwasnik, Kevin Crowston, Mike Nilan, and Dmitri Roussinov. Identifying document genre to improve web search effectiveness. the bulletin of the american society for information science and technology. *The American Society for Information Science and Technology*, page 23, 2000. 2.1

Bedford/St. Martin's. Genre classification. URL `https://wac.colostate.edu/resources/wac/intro/genre/`. 2.1.1, 2.2

msg. One rule algorithm, 2020. URL `https://machinelearningcatalogue.com/algorithm/alg_one-rule.html`. 6.2

G. Rehm. Towards automatic web genre identification: a corpus-based approach in the domain of academia by example of the academic's personal homepage. In *Proceedings of the 35th Annual Hawaii International Conference on System Sciences*, pages 1143–1152, 2002. doi: 10.1109/HICSS.2002.994036. 2.1, 2.4

D. Roussinov, K. Crowston, M. Nilan, B. Kwasnik, Jin Cai, and Xiaoyong Liu. Genre based navigation on the web. In *Proceedings of the 34th Annual Hawaii International Conference on System Sciences*, pages 10 pp.–, 2001. doi: 10.1109/HICSS.2001.926478. 2.4

Andrés Sanoja and Stéphane Gançarski. Block-o-matic: A web page segmentation framework. In *2014 International Conference on Multimedia Computing and Systems (ICMCS)*, pages 595–600, 2014. doi: 10.1109/ICMCS.2014.6911249. 1

Santini. State-of-the-art on automatic genre identification. URL `https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.5.7680&rep=rep1&type=pdf`. 2

Marina Santini. Characterizing genres of web pages: Genre hybridism and individualization. In *2007 40th Annual Hawaii International Conference on System Sciences (HICSS'07)*, pages 71–71, 2007. doi: 10.1109/HICSS.2007.124. 2.2

Marina Santini. Zero, single, or multi? genre of web pages through the users' perspective. *Information Processing Management*, 44(2):702–737, 2008. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2007.05.011. URL `https://www.sciencedirect.com/science/article/pii/S0306457307001185`. Evaluating Exploratory Search Systems Digital Libraries in the Context of Users' Broader Activities. 3

Yutaka Sasaki. The truth of the f-measure, 2007. URL https://www.toyota-ti.ac.jp/Lab/Denshi/COIN/people/yutaka.sasaki/F-measure-YS-26Oct07.pdf. 5.3

M. Shepherd and C. Watters. The evolution of cybergenres. In *Proceedings of the Thirty-First Hawaii International Conference on System Sciences*, volume 2, pages 97–109 vol.2, 1998. doi: 10.1109/HICSS.1998.651688. 2.4

Wikipedia. classification. URL https://en.wikipedia.org/wiki/Random_forest. 1, 5.2, 2

Wikipedia. Html, 2020. URL https://en.wikipedia.org/wiki/HTML. 1

Wikipedia contributors. Internet genre — Wikipedia, the free encyclopedia, 2020. URL https://en.wikipedia.org/w/index.php?title=Internet_genre&oldid=937478059. [Online; accessed 3-July-2021]. 2, 3

Wikipedia contributors. Internet forum — Wikipedia, the free encyclopedia, 2021a. URL https://en.wikipedia.org/w/index.php?title=Internet_forum&oldid=1025810261. [Online; accessed 3-July-2021]. 3

Wikipedia contributors. Search engine optimization — Wikipedia, the free encyclopedia, 2021b. URL https://en.wikipedia.org/w/index.php?title=Search_engine_optimizationoldid=1031052589. [Online; accessed 2-July-2021]. 2.1.1

Wikipedia contributors. Targeted advertising — Wikipedia, the free encyclopedia, 2021c. URL https://en.wikipedia.org/w/index.php?title=Targeted_advertisingoldid=1029807320. [Online; accessed 1-July-2021]. 1, 2.1, 2.1.1

Zenodo. segmentation. URL https://zenodo.org/record/3988124#.X7IiOVNKhQK. 1, 2.3