TeejLab™

Enhancing the TeejLab Analyzer for Data Service Agreements

Team Members Chao Wang
Jessie Yu
Sylvia Lee
Talha Siddiqui

Mentor Varada Kolhatkar

Date June 17, 2019

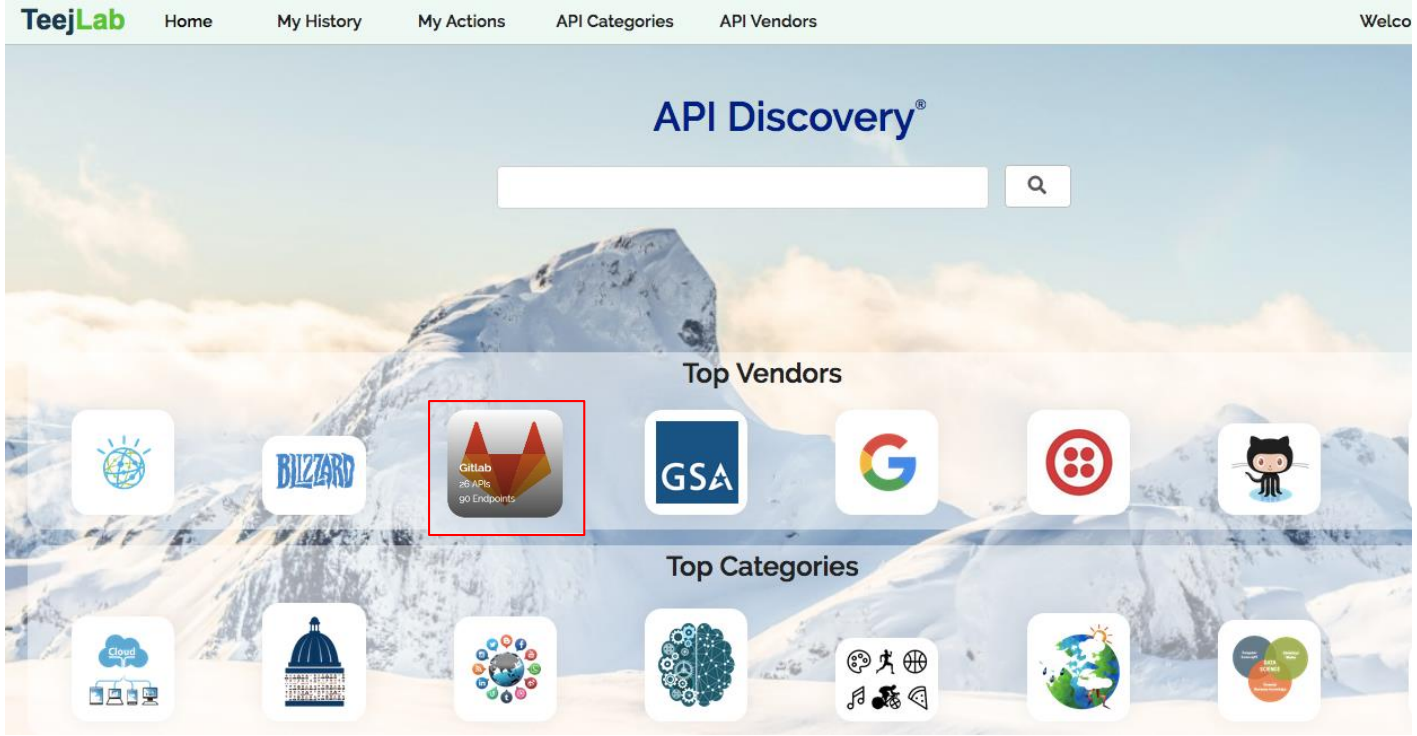
Agenda

TeejLab™

- Background & Problem Review
- Pursed Scientific Objectives
- Our Data Products
- Data Science Techniques
- Conclusions
- Questions

Background & Problem Review

TeejLab™



- Software-as-a-Service Platform
- Designed for developers

Background & Problem Review

TeejLab™

TeejLab Home My History My Actions API Categories API Vendors Welcome, chao

Gitlab

Homepage: https://about.gitlab.com/

26 APIs

Column Visibility | Show 10 Rows | Search API | GO | CANCEL

Analyze Agreement

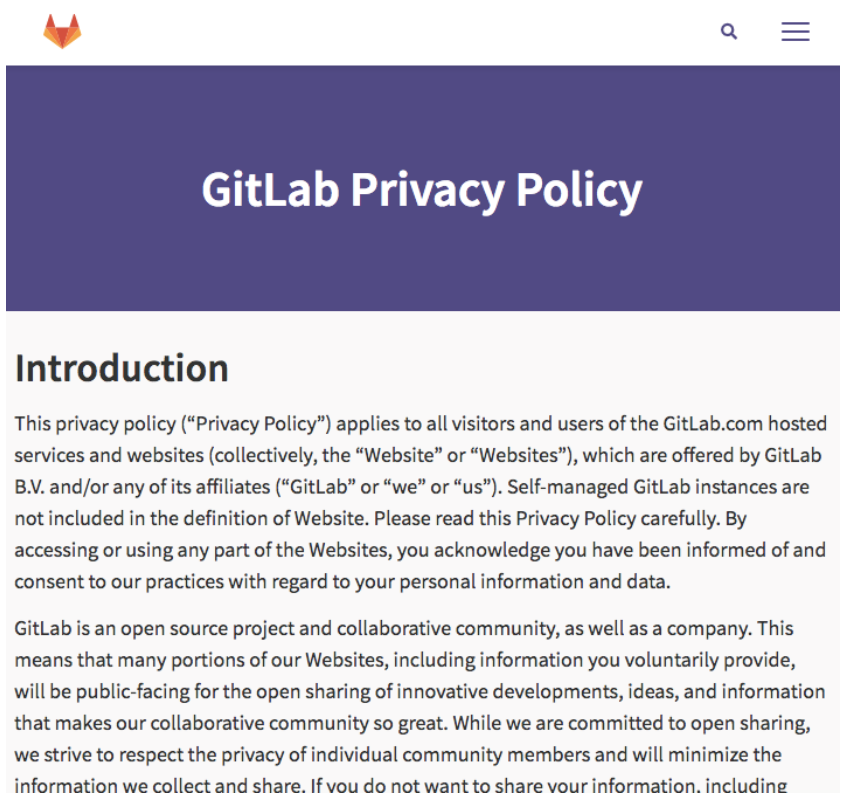
GitLab Privacy Policy

API Name	Version	Category	Homepage	API Agreements	API Keys
GitLab CI YMLs API	4	Software & Services			
Gitlab Award Emoji API	4	Software & Services			
Gitlab Branches API	4	Software & Services			
Gitlab Commits API	4	Software & Services			
Gitlab Environments API	4	Software & Services			
Gitlab Events API	4	Software & Services			
Gitlab Gitignores API	4	Software & Services			

© 2019 TeejLab Terms Privacy

Background & Problem Review

TeejLab™

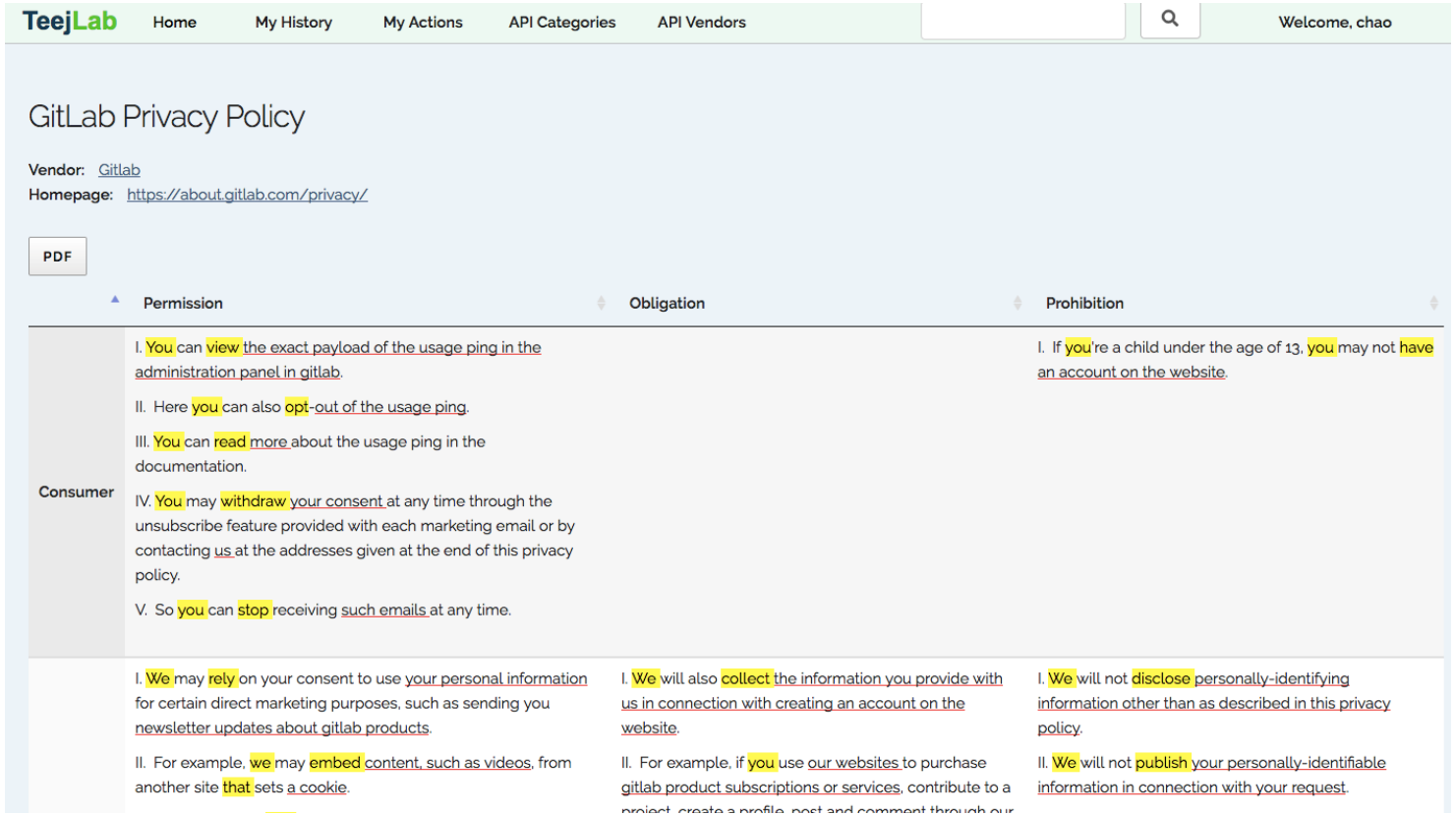


GitLab Privacy Policy

Introduction

This privacy policy ("Privacy Policy") applies to all visitors and users of the GitLab.com hosted services and websites (collectively, the "Website" or "Websites"), which are offered by GitLab B.V. and/or any of its affiliates ("GitLab" or "we" or "us"). Self-managed GitLab instances are not included in the definition of Website. Please read this Privacy Policy carefully. By accessing or using any part of the Websites, you acknowledge you have been informed of and consent to our practices with regard to your personal information and data.

GitLab is an open source project and collaborative community, as well as a company. This means that many portions of our Websites, including information you voluntarily provide, will be public-facing for the open sharing of innovative developments, ideas, and information that makes our collaborative community so great. While we are committed to open sharing, we strive to respect the privacy of individual community members and will minimize the information we collect and share. If you do not want to share your information, including



TeejLab Home My History My Actions API Categories API Vendors Welcome, chao

GitLab Privacy Policy

Vendor: Gitlab
Homepage: https://about.gitlab.com/privacy/

PDF

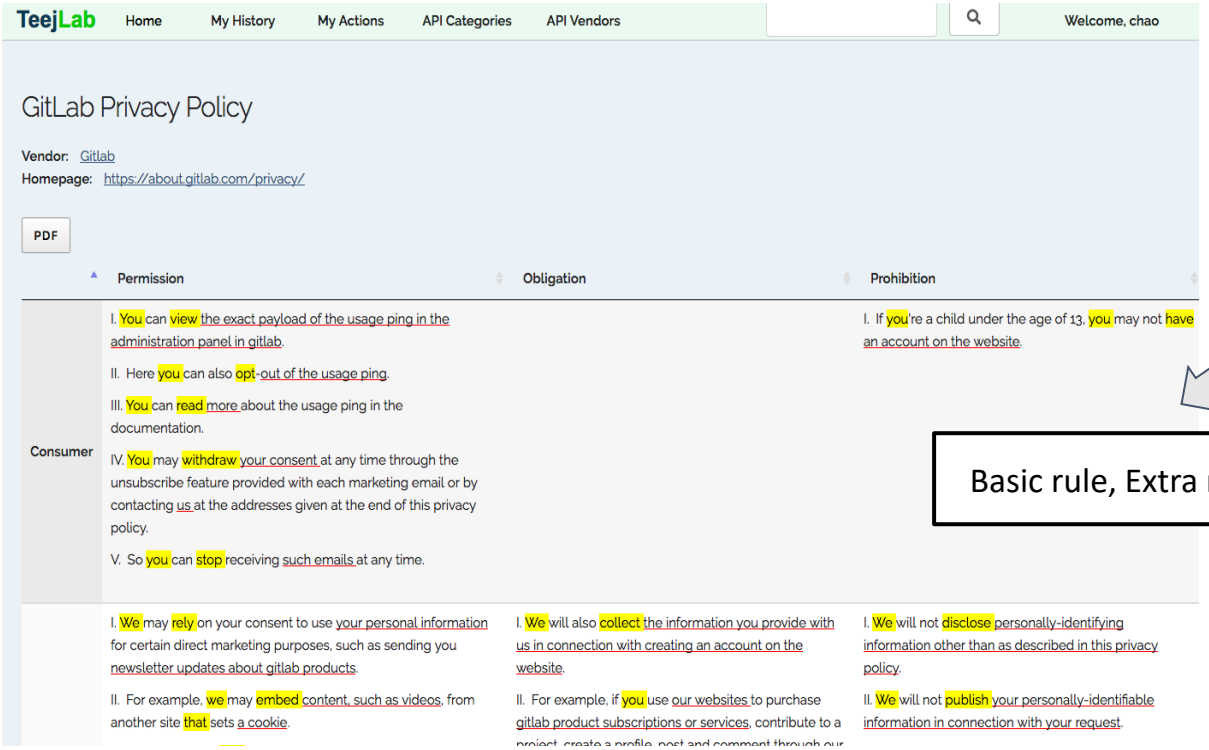
	Permission	Obligation	Prohibition
Consumer	I. You can view the exact payload of the usage ping in the administration panel in gitlab. II. Here you can also opt-out of the usage ping. III. You can read more about the usage ping in the documentation. IV. You may withdraw your consent at any time through the unsubscribe feature provided with each marketing email or by contacting us at the addresses given at the end of this privacy policy. V. So you can stop receiving such emails at any time.		I. If you're a child under the age of 13, you may not have an account on the website.
	I. We may rely on your consent to use your personal information for certain direct marketing purposes, such as sending you newsletter updates about gitlab products. II. For example, we may embed content, such as videos, from another site that sets a cookie.	I. We will also collect the information you provide with us in connection with creating an account on the website. II. For example, if you use our websites to purchase gitlab product subscriptions or services, contribute to a project, create a profile, post and comment through our	I. We will not disclose personally-identifying information other than as described in this privacy policy. II. We will not publish your personally-identifiable information in connection with your request.

GitLab Privacy Policy Web Page

TeejLab Agreement Analyzer Results

Previous Analyzer

TeejLab™



TeejLab Home My History My Actions API Categories API Vendors Welcome, chao

GitLab Privacy Policy

Vendor: Gitlab
Homepage: https://about.gitlab.com/privacy/

PDF

	Permission	Obligation	Prohibition
Consumer	I. You can view the exact payload of the usage ping in the administration panel in gitlab. II. Here you can also opt-out of the usage ping. III. You can read more about the usage ping in the documentation. IV. You may withdraw your consent at any time through the unsubscribe feature provided with each marketing email or by contacting us at the addresses given at the end of this privacy policy. V. So you can stop receiving such emails at any time.		I. If you're a child under the age of 13, you may not have an account on the website.
	I. We may rely on your consent to use your personal information for certain direct marketing purposes, such as sending you newsletter updates about gitlab products. II. For example, we may embed content, such as videos, from another site that sets a cookie.	I. We will also collect the information you provide with us in connection with creating an account on the website. II. For example, if you use our websites to purchase gitlab product subscriptions or services, contribute to a project, create a profile, post and comment through our	I. We will not disclose personally-identifying information other than as described in this privacy policy. II. We will not publish your personally-identifiable information in connection with your request.

URL

Scraping, Preprocessing

Find the most two common parties

Basic rule, Extra rule

Modal words follows party

Modal words in keyword list

If negition appears

Unidentified

Permission

Obligation

Prohibition

JSON

```
modalPermission = ['can', 'could', 'may', 'might', 'allowed']
modalObligation = ['must', 'should', 'shall', 'will', 'responsible for']
```

Major Challenges

- Lack of labelled data
- No established evaluation matrix

This Photo by Unknown Author is licensed under CC BY-SA-NC

Scientific Objectives

TeejLab™

Evaluation framework	Accurate & Robust Multiclass Classification	Improving UI/UX (NLP highlighting)
1st Priority – Recall	Supervised & Unsupervised Learning Rule-Based Model	New Category Statement Highlighting

Result Snapshots | UI/UX

TeejLab™

Before

Permission	Obligation	Prohibition
I. For example, you can sign up for a google account if you want to create and manage content like emails and photos, or see more relevant search results. II. And you can use many google services when you're signed out or without creating an account at all, like searching on google or watching youtube videos. III. You can also choose to browse the web privately using chrome in incognito mode. IV. And across our services, you can adjust your privacy settings to control what we collect and how your information is V. And if you have any questions about this privacy policy, you can contact us.information VI. You can also choose to add a phone number or payment information to your account. VII. Even if you aren't signed in to a google account, you might choose to provide us with information — like an email address to receive updates about our services.we also collect the content		

After

	Permission	Obligation	Prohibition
Customer	• Even if you aren't signed in to a Google Account, you might choose to provide us with information — like an email address to receive updates about our services. • If you don't want this level of search customization, you can search and browse privately or turn off signed-out search personalization. • For example, you may see ads for things like "**Cooking and Recipes**" or "**Air Travel**." We don't use topics or show personalized ads based on sensitive categories like race, religion, sexual orientation, or health.	• We collect information to provide better services to all our users — from figuring out basic stuff like which language you speak, to more complex things like which ads you'll find most useful, the people who matter most to you online, or which YouTube videos you might like. • Things you create or provide to us • When you create a Google Account, you provide us with personal information that includes your name and a password.	• This Privacy Policy doesn't apply to services that have separate privacy policies that do not incorporate this Privacy Policy. • However, some website features or services may not function properly without cookies.
Provider	• When you use our services, you're trusting us with your information. • You can use our services in a variety of ways to manage your privacy. • For example, you can sign up for a **Google** Account if you want to create and manage content like emails and photos, or see more relevant search results. • And you can use many **Google** services when you're signed out or without creating an account at all, like searching on **Google** or watching YouTube videos.	• This includes things like email you write and receive, photos and videos you save, docs and spreadsheets you create, and comments you make on **YouTube** videos. • If you're using an **Android** device with **Google** apps, your device periodically contacts **Google** servers to provide information about your device and connection to our services. • In some circumstances, **Google** also collects information about you from publicly accessible sources. • For example, understanding how people organized their	
Provider Disclaimer	• The information **Google** collects, and how that information is used, depends on how you use our services and how you manage your privacy controls. • When you're not signed in to a Google Account, we store the information we collect with unique identifiers tied to the browser, application, or device you're using. • The information we collect includes unique identifiers, browser type and settings, device type and settings, operating system, mobile network information including carrier name and phone number, and application version number.		

Solution Flow

TeejLab™

Data Acquisition

- Labelled data
- Web scraping

Feature Processing

- HTML metadata parsing
- Text vectorization

Classification

- Hybrid multiclass classification pipeline
- Model tuning & Feature selection

UI/UX Presentation

- Layouts redesign
- New highlighting rules

Data Acquisition

Feature Processing

Classification

UI/UX Presentation

Labelled Data | Lawyers

TeejLab™

1,787 statements in 5 weeks

Three Lawyers in Canada and United States

Top 5 Vendors – Facebook, Google, Stripe, Blizzard, AWS

Labelled Data | Crowdsourcing

TeejLab™

Data Acquisition

Feature Processing

Classification

UI/UX Presentation

What Is The Most Suitable Category For These Sentences?

Statement:

This Agreement is not assignable, transferable, or sublicensable by you except with Tumblr's prior written consent.

Which class best fits the statement? (required)

- ○ Permission
- ◉ Prohibition
- ○ Obligation
- ○ Disclaimer and Good-to-know
- ○ Irrelevant

Which party does this Prohibition concern? (required)

- ○ User
- ○ Vendor

ℹ Vendor is the service provider typically denoted by 1st person pronouns (e.g. We, Us, Our). User is the service consumer typically denoted by 2nd person pronouns (e.g. You, Your). In the absense of a stated party, User is usually the concerned party.

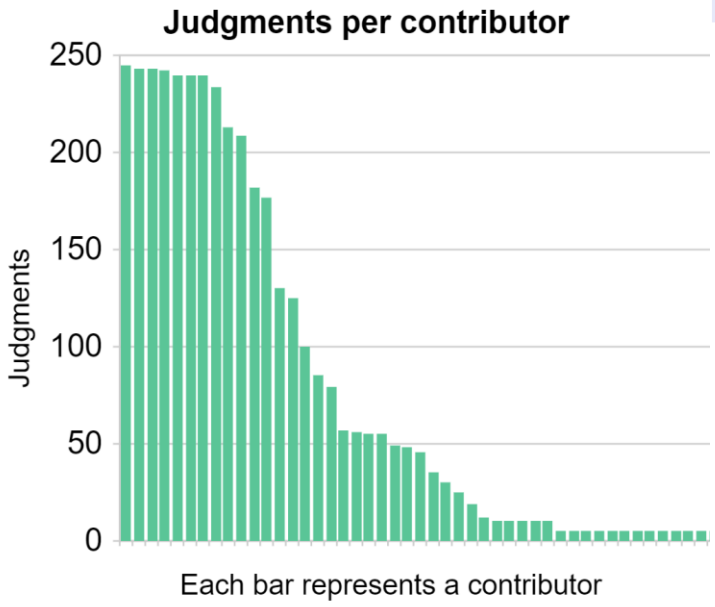
Any comments:

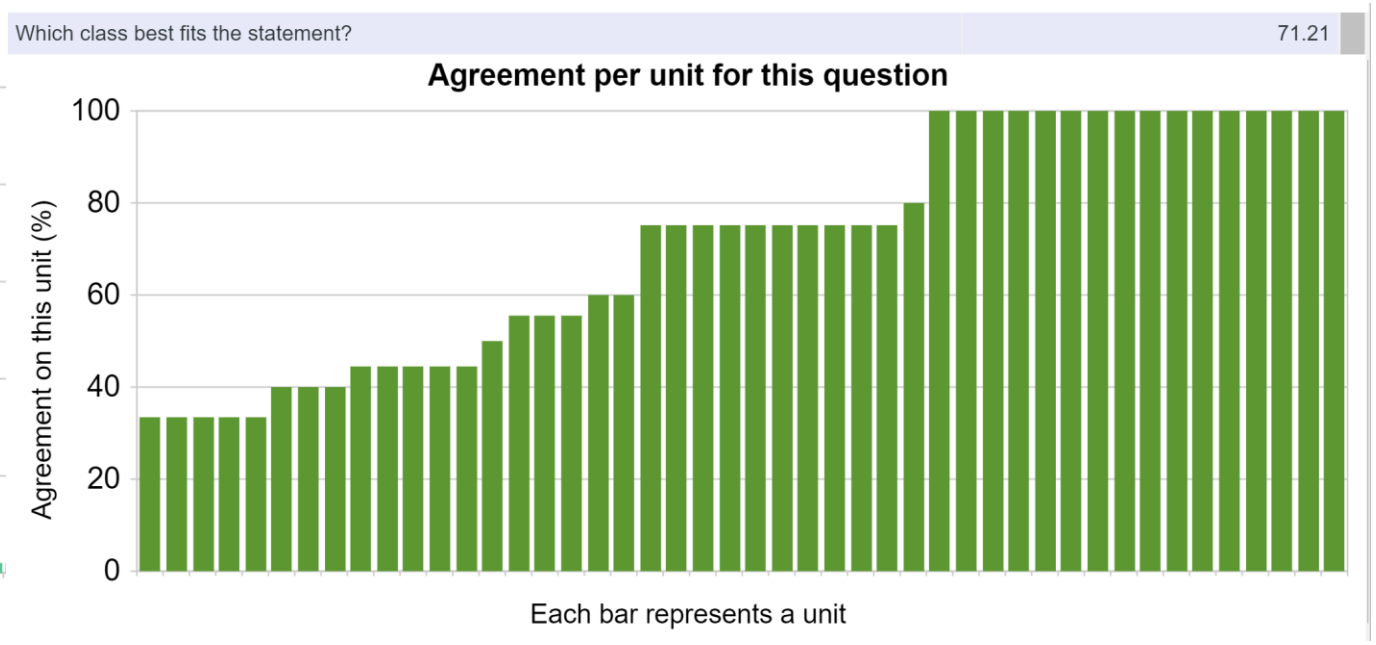
Labelled Data | Crowdsourcing

TeejLab™

Figure Eight annotation platform

100%	$87	52	500	2,767	92
Complete	Cost	Active Test Questions	Rows	Trusted Judgments	Untrusted Judgments





Labelled Data | Crowdsourcing

TeejLab™

Settings:

- 2¢ per judgement
- 3 to 8 judgements per statement (as soon as 70% agreement reached)
- 70% min accuracy from test questions
- Contributors from English-speaking countries

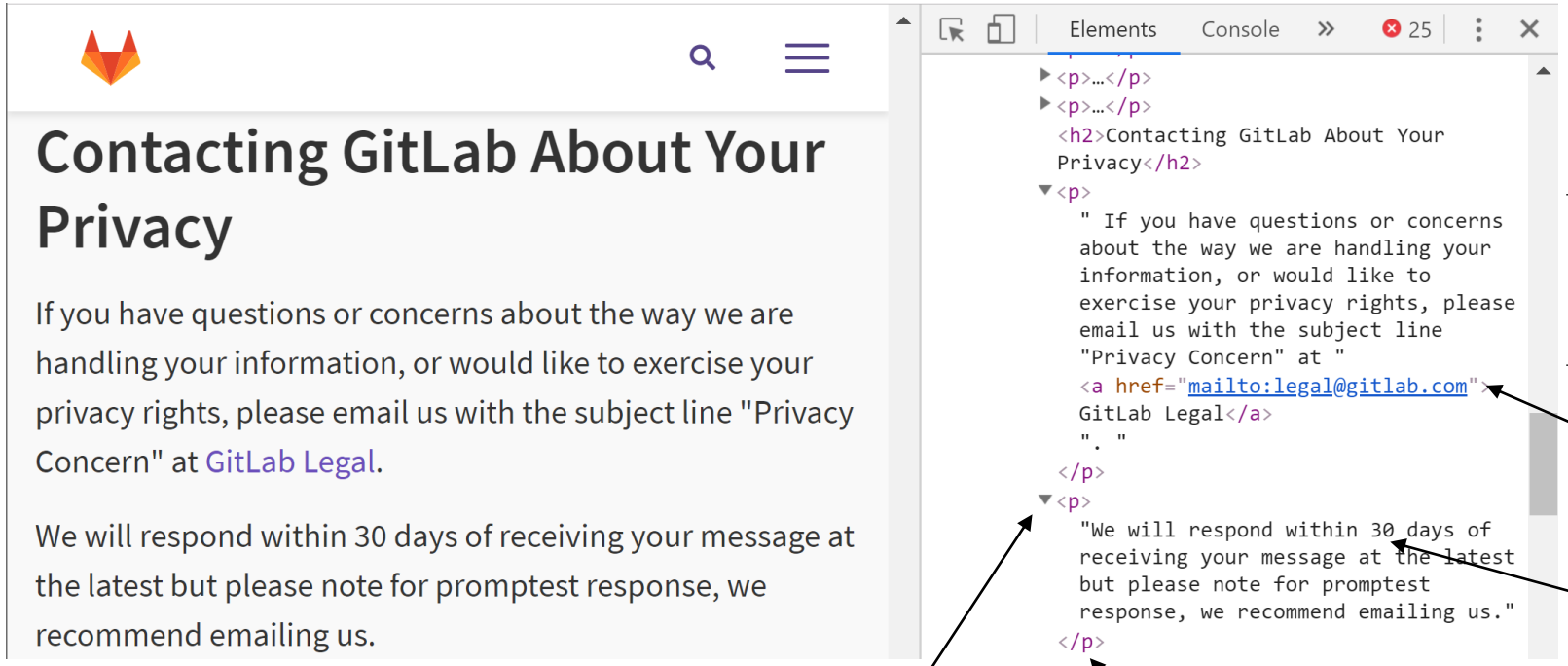
$231.84

20 hours

1,368
statements

Web Scraping

TeejLab™



Word count, character count, average length of word etc.

In-page or external Hyperlinks

Digits

Position

HTML Tags

Data Acquisition

Feature Processing

Classification

UI/UX Presentation

HTML Metadata Parser

TeejLab™

Extract 27 features of data

- HTML tags, character & word count, punctuation & digit count, position in the document etc.

tag_ind	tag	is_p	is_h1	is_h2	is_h3	is_h4	...	hyperlink_txt_cc	hyperlink_txt_cr	hyperlink_url_cc	hyperlink_txt_wc_avg	hyperlink_txt_wr_avg
0	h2	False	False	True	False	False	...	0	0.000000	0	0.0	0.000000
1	p	True	False	False	False	False	...	23	0.920000	8	4.0	1.000000
1	p	True	False	False	False	False	...	19	0.904762	8	3.0	1.000000
1	p	True	False	False	False	False	...	12	0.857143	5	2.0	1.000000

Text Vectorization

TeejLab™

N-gram Vectorizer (Uni-, Bi- & Trigram)

TF-IDF Vectorizer

- Convert a collection of text documents to a matrix of TF-IDF features, measuring the importance of words

POS Vectorizer

- Append POS tags to each token before converting text to token counts

BERT

- State-of-the-art: Pre-trained deep bidirectional embedding transformer
- Server-client portal for reduced time and performance requirements

Model Tuning

TeejLab™

Models

- Random Forest
- KNN
- SVM
- Logistic Regression
- Feed Forward Neural Network
- XGBoost

Evaluation:

- Select the best model out of 6 with the highest F-beta score
- F-beta score is the weighted harmonic mean of precision and recall

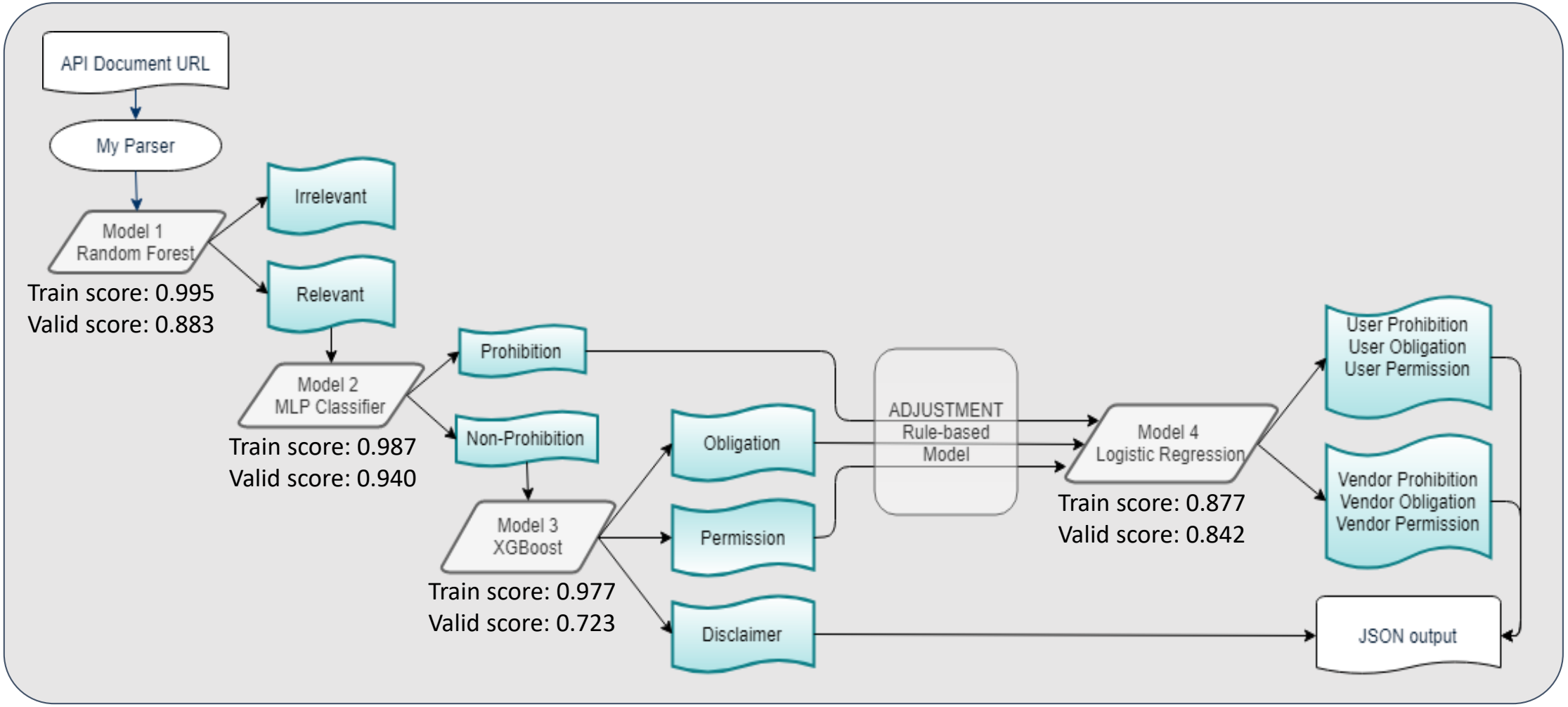
Feature Selection

TeejLab™

- Score Evaluation
  - Select the best text vectorization out of five with the highest F-beta Score
- Forward Selection
  - In each forward step, add the features of data that gives the single best F-beta to the model among 27 features

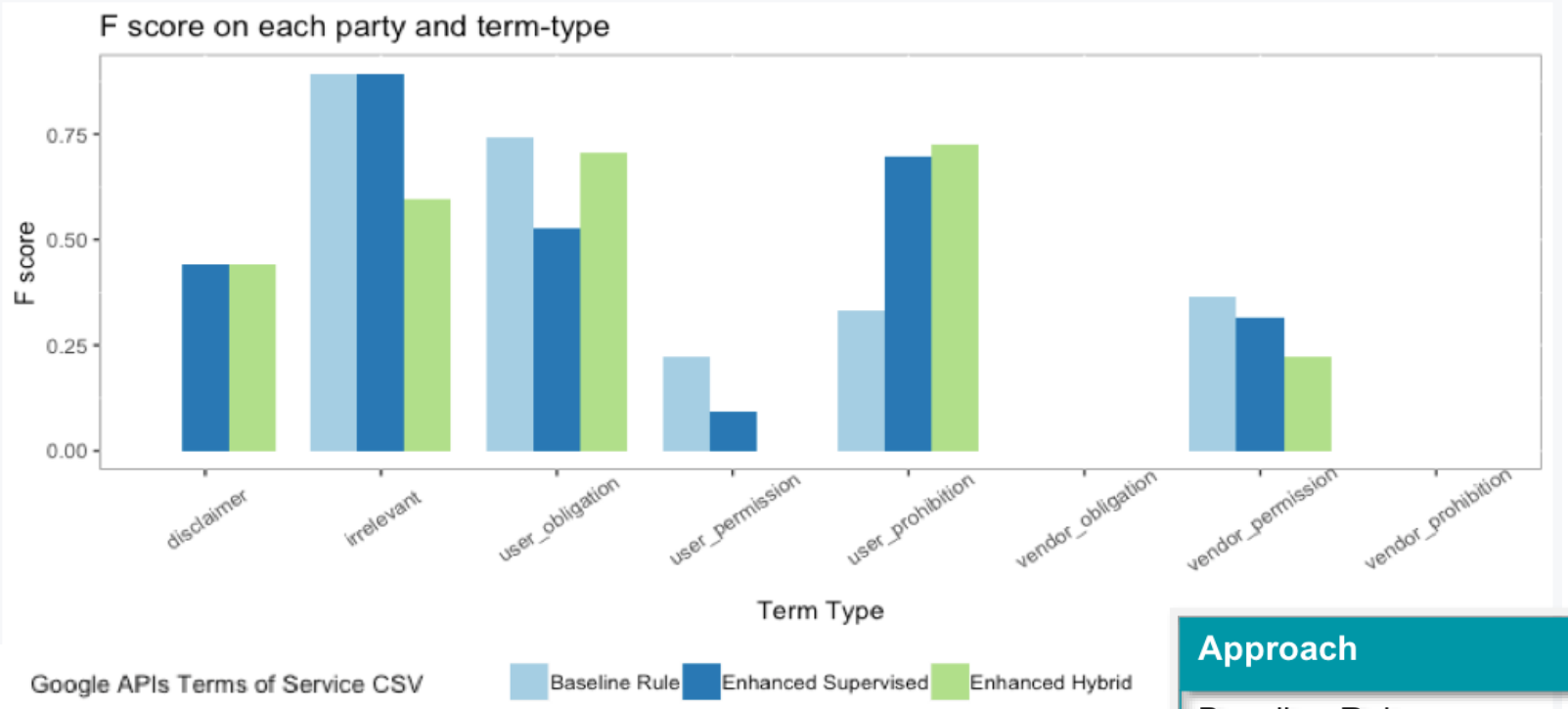
Classification Pipeline

TeejLab™



Model Evaluation

TeejLab™



Approach	Precision	Recall	F1-score
Baseline Rule	0.461	0.494	0.407
Enhanced Supervised	0.639	0.614	0.608
Enhanced Hybrid	0.657	0.639	0.640

Model Speed

TeejLab™

Live URL → JSON Output

Step / Time (s)	Previous Rule-Based	Enhanced Hybrid
Scrap & preprocess	0.59	0.29
Classification	1.66	2.61

On Apple Website Terms of Use

Data Acquisition

Feature Processing

Classification

UI/UX Presentation

UI/UX Design

TeejLab™

Don't use the API to stream directly from a mobile phone or tablet camera to Facebook.

Stripe provides Data to third-party service providers, including Financial Services Providers and their affiliates, as well as Stripe's global affiliates, to allow us to provide Services to you and other users.

Intertrust is permitted to, and Customer hereby grants to Intertrust the right to, access and use Customer Content in connection with providing the Platform.

If you're using iOS to run your app, use an iOS approved payment method.

Don't use the **API** to stream directly from a mobile phone or tablet camera to **Facebook**.

Stripe provides **Data** to **third**-party service providers, including **Financial Services Providers** and their affiliates, as well as **Stripe**'s global affiliates, to allow us to provide Services to you and other users.

Intertrust is permitted to, and **Customer** hereby grants to Intertrust the right to, access and use **Customer Content** in connection with providing the **Platform**.

If you're using iOS to run your app, use an iOS approved payment method.

To aid reading and understanding:

- Proper nouns are **bolded** to highlight addressed objects
- Root verbs are underlined to emphasize actions involved

Implemented with Spacy:

- Name-entity recognition
- Dependency parsing

Attempted Techniques

TeejLab™

Unsupervised approaches

- LDA, K-means, DBSCAN

Supervised Approaches

- SIF Embedding, Bidirectional Recurrent Neural Network

Rule-based Approaches

- Spacy (dependency parsing), Chunking

Conclusions / Take-Aways

TeejLab™

1. Adopt Hybrid Hierarchical Model to achieve best performance.
2. Think about User Cases
3. Indirect solutions to some Capstone partner identified problems due to the change from rule-based to hybrid model
4. Data Product/ Techniques Decision Criteria

TeejLab™

Questions

This Photo by Unknown Author is licensed under CC BY-NC-ND