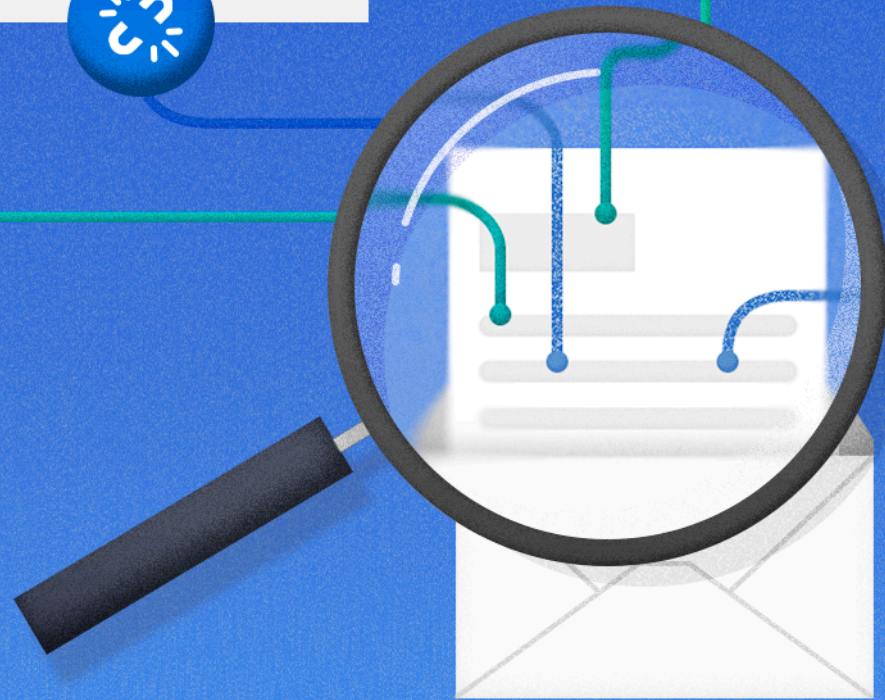
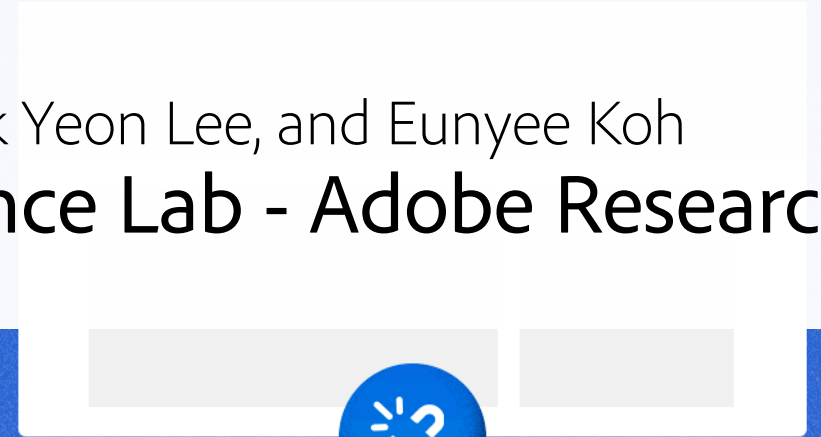
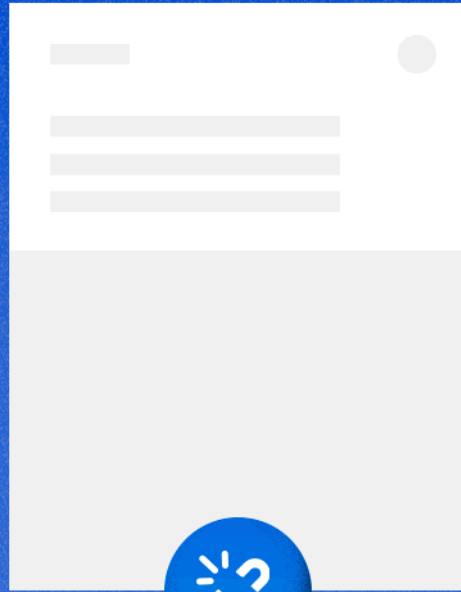
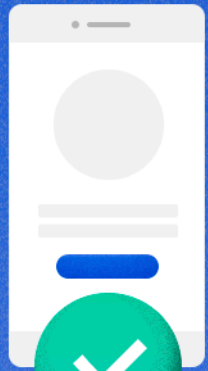


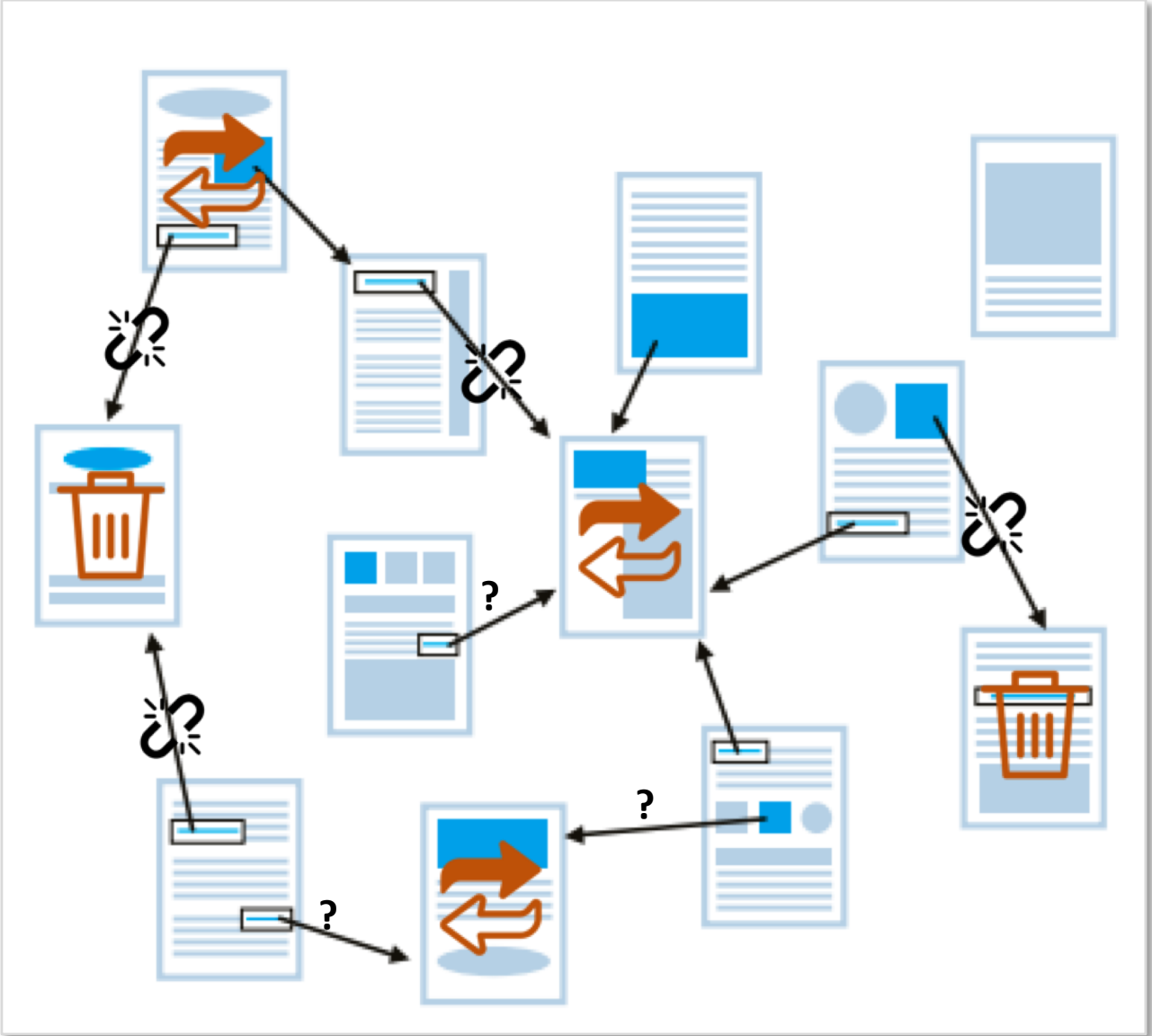


# Detecting link and landing page misalignment in marketing emails

*Nedim Lipka, Tak Yeon Lee, and Eunyee Koh*  
Big Data Experience Lab - Adobe Research









~20% of Web pages changed every time they were visited

~40% of Web pages changed within a week

Cho and Garcia-Molina, 2000



0.25-0.5% of documents on the Internet disappear  
every week

Fetterly et al, 2003

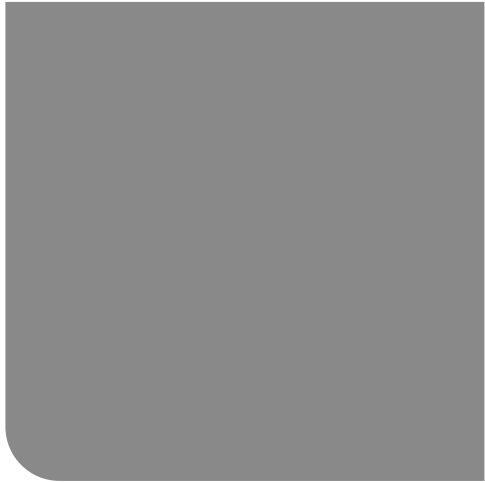


# *Changing the World*

Through Digital Experiences







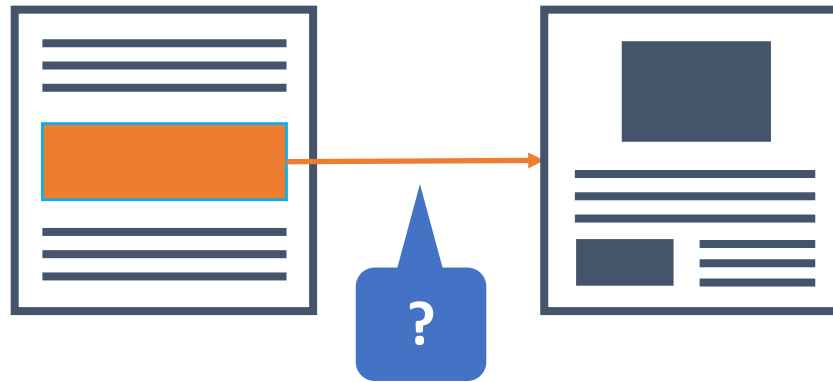
## Task

---

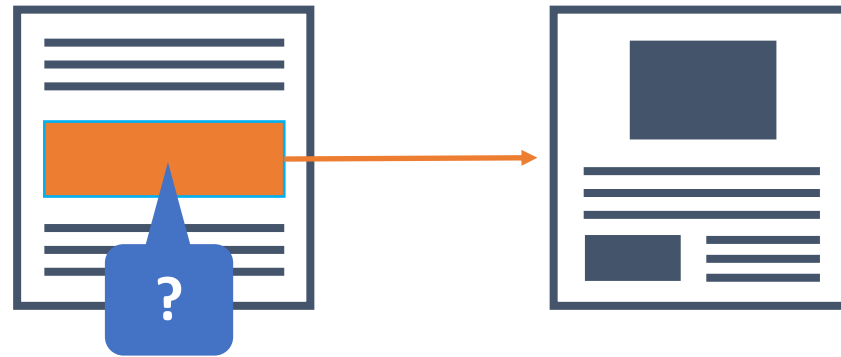
Given a link within a marketing email and the corresponding landing page, detect misalignment.



# Use case: Link Monitoring System



# Use case: Smart Authoring Tool





# Summary

**Data:** 160 marketing emails, 4266 links, labeled by MTurk crowd workers

**Model Comparison:** Unsupervised (cosine similarity) vs. supervised (learned similarity with neural networks)

**Evaluation:** Best F measure: 0.75

# Outline

- I. Link Issue Taxonomy
- II. Data
- III. Model
- IV. Evaluation
- V. Recent and Future Work

# Outline

- I. Link Issue Taxonomy
- II. Data
- III. Model
- IV. Evaluation
- V. Recent and Future Work

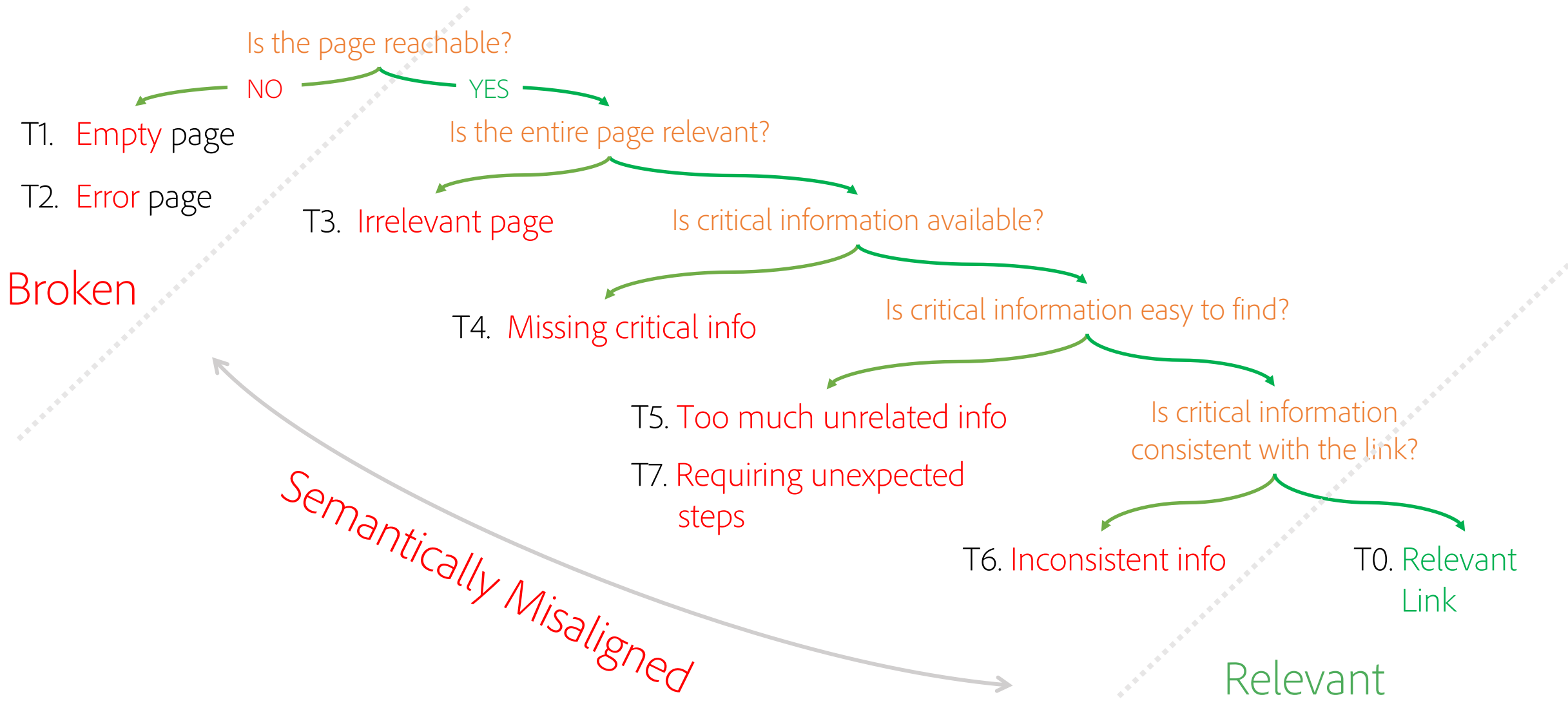


# A taxonomy of link issues in marketing emails

ID	BATCH	ERROR_404	ERROR_CUSTOM	ERROR_EMPTY	MISSING_OFFER	MISMATCHING_OFFER	INCONSISTENT_OFFER	REDIRECT_OTHER_DOMAIN	REDIRECT_OTHER_STEP	LIST_NOT_ITEM	REDIRECT_ROOT	REDIRECT_OTHER_DOMAIN	UNEXPECTED_STEP	NO_ISSUE	Impact	Issue
41	MTurk12										v				2	No, company page
42	MTurk14									v					1	I expected it to take me to the full menu, not just the specials
43	MTurk17										v				1	I expected it to take me to the Netflix page of this artist but it took me to netflix's home page.
44	MTurk20										v				3	I exoected it to take me to the catalog, it just took me to the Homepage.
46	GForm1												v		4	Join us leads to the home page rather than the sign-up page.
47	GForm1										v				4	All the above three texts on top of a light brown bar (bottom of a promotion email) link to the citi
48	MTurk10												v		4	This is the main link from the promotional email. When clicking, what I expected to see was a pa
49	MTurk13												v		2	I was wanting to just look at the glasses, but I had to take a quiz first.
50	MTurk8												v		5	Why would I want to subscribe to an Email list that i just recieved an email from.
51	MTurk18												v		5	because it takes me to a sign up
52	MTurk8												v		5	It took me a page that began a bunch of surveys and 'offers'. I never got to a place to request Fel

Iterative Coding

Why participants thought the link was misaligned



# Outline

- I. Link Issue Taxonomy
- II. **Data**
- III. Model
- IV. Evaluation
- V. Recent and Future Work

# Training data collection

TASK 3 / 5

Can you see the landing page on the right? Compare the link and the landing page carefully. **Are the email link and the landing page relevant to each other?**

Irrelevant

Relevant

Tell us why they are NOT relevant.

- The landing page is empty.
- The page shows an error.
- The page is entirely irrelevant.
- The page is missing critical information.
- The page contains too much unrelated information.
- Important details (e.g. price, discount, date) are inconsistent with the email.
- The page requires unexpected steps (e.g. sign-up, subscription)

Submit and Proceed

- 160 marketing emails
- 4266 links

MARKETING EMAIL

fit original

Subject: Thu, 01 Mar 2018 16:48:07 -0600

Subscribe to BostonGlobe.com

**GEORGE BURNS:** After a few years touring with fellow ?Comedy Get Down? comics D.L. Hughley, Cedric the Entertainer, Eddie Griffin, and Charlie Murphy, comedian **George Lopez** is giving himself some much-needed personal space on his current tour. And boy could he use it. Dude recently **gave** up his beloved game of golf until Trump deals with

LANDING PAGE

fit original

The screenshot shows a landing page for a comedy tour. At the top, there's a banner for 'LOVELESS' with a 'SUBSCRIBE' button. Below that, the main headline reads 'Solo artist George Lopez joins a team for Comedy Get Down tour'. There are social media sharing icons for Facebook, Twitter, and LinkedIn. To the right, there's a section titled 'Top 30 Trending Articles' with a list of article titles and dates. The page layout is clean and professional, typical of a news or entertainment website.

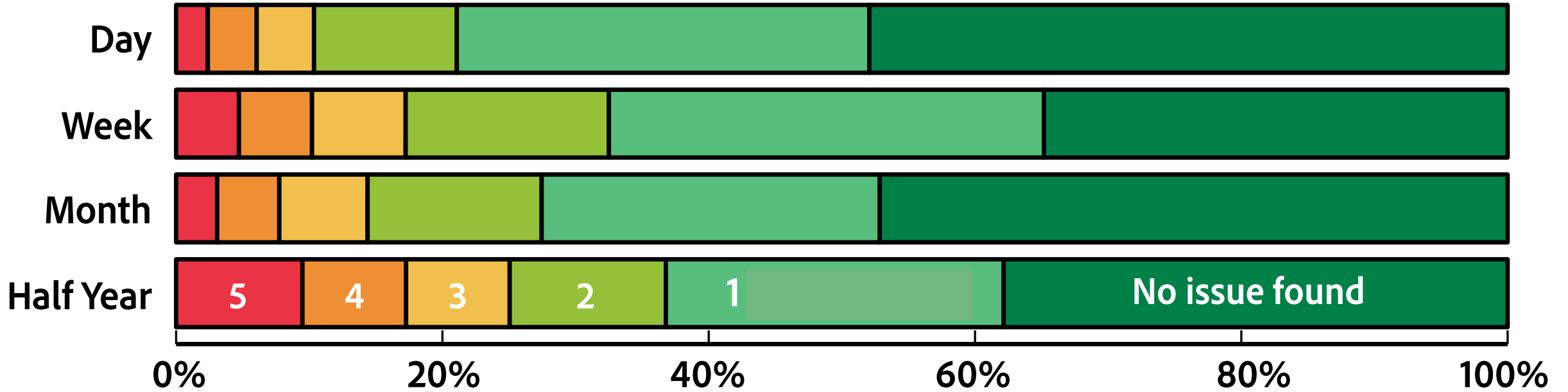


# Challenges

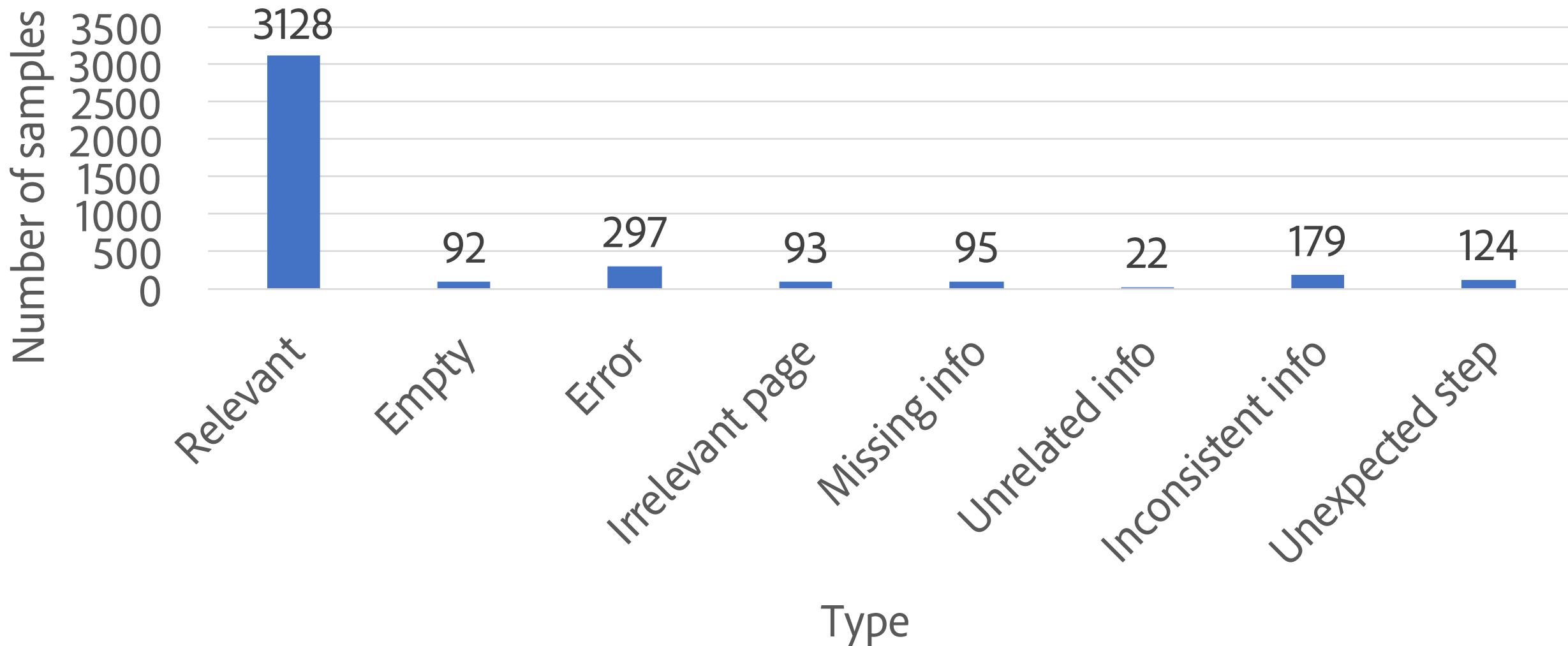
Link relevance is a subjective measure

- Expiration of offers or content
- Conflict of interests

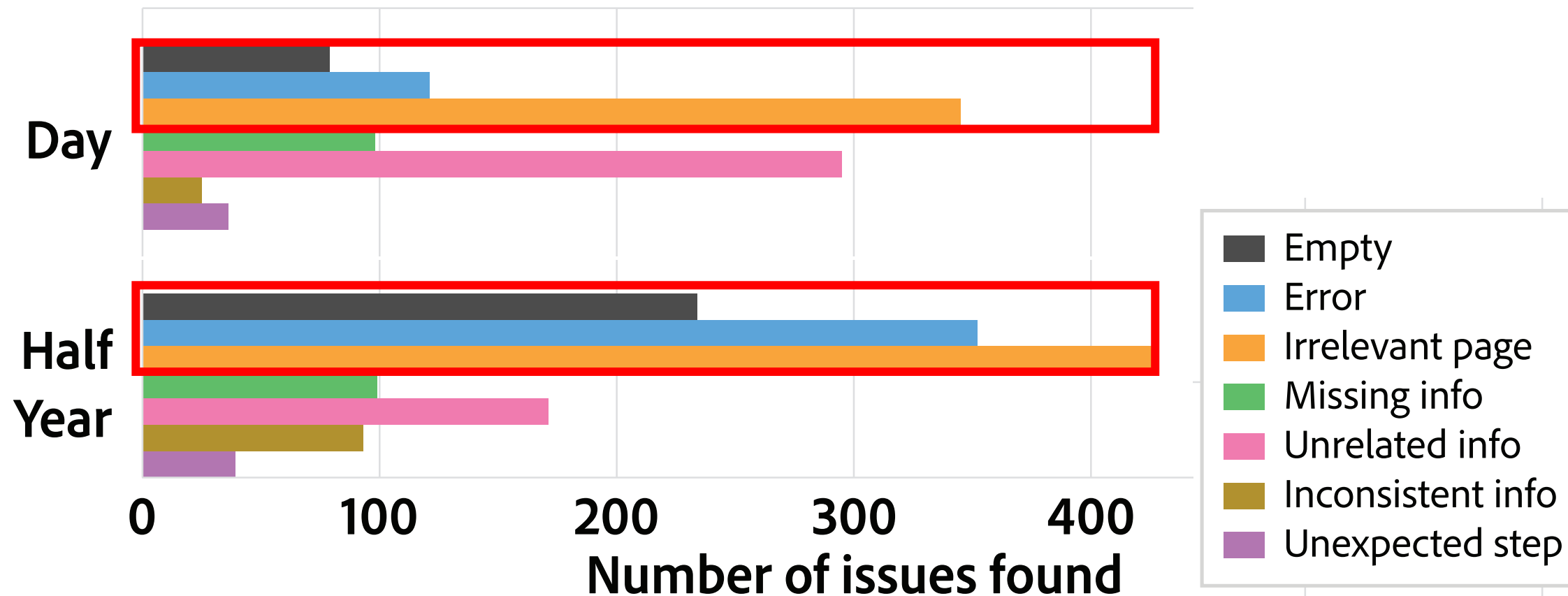
# Perceived Relevance of Links



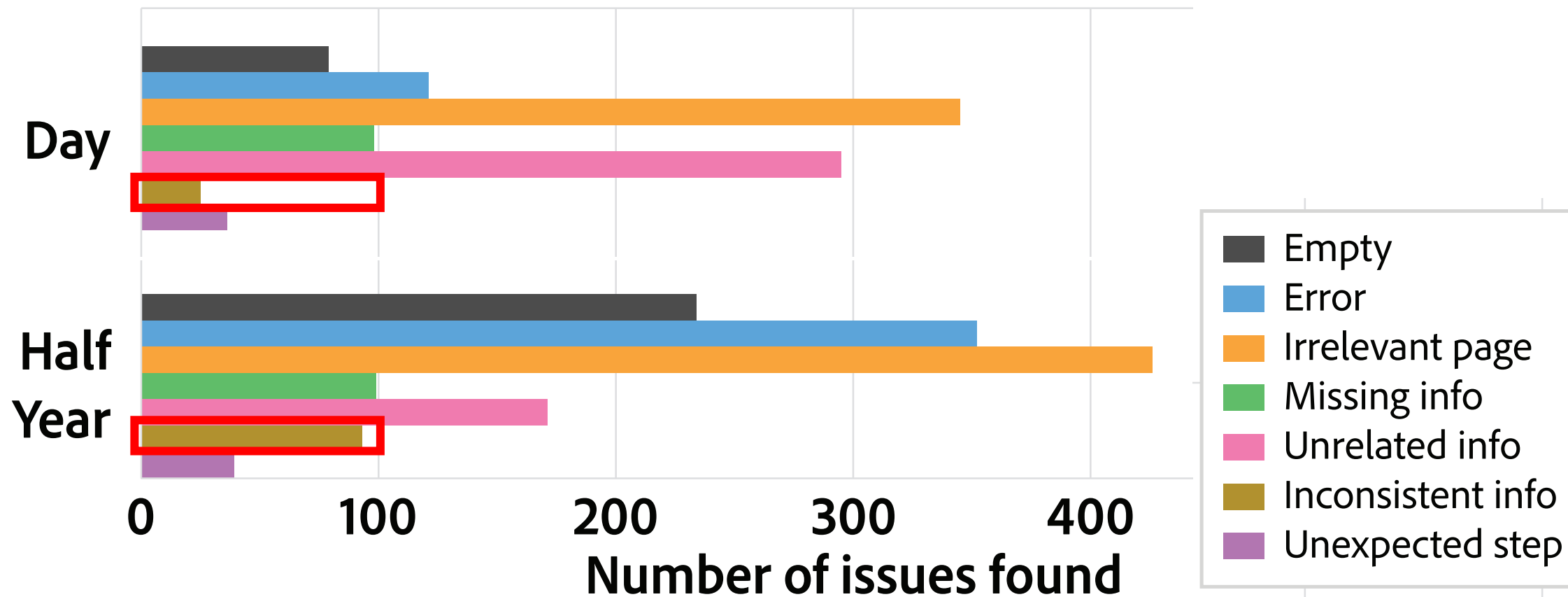
# Class balance



# Distribution of issues over email age



# Distribution of issues over email age



# Outline

- I. Link Issue Taxonomy
- II. Data
- III. Model**
- IV. Evaluation
- V. Recent and Future Work

# Approach

## Finding features

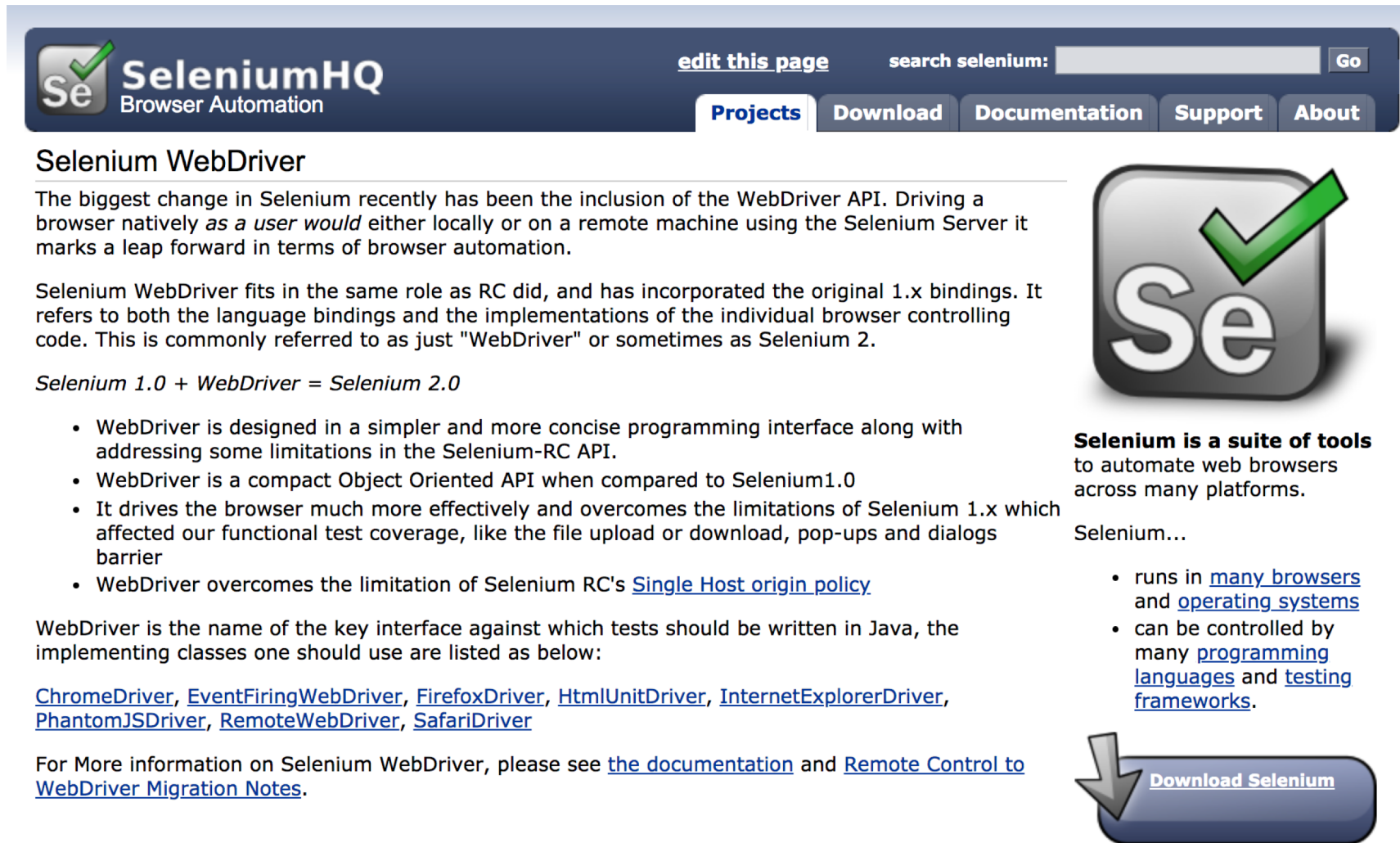
- Image features (OCR and tagging)
- Text features (link, subject line, body, landing page)

## Classification

- Is the landing page relevant? (Binary)
- What is the issue with the landing page? (Multi-class)



# Best practice: Retrieving landing page content



The screenshot shows the SeleniumHQ website. At the top, there is a navigation bar with the SeleniumHQ logo (a green checkmark over the letters 'Se') and the text 'SeleniumHQ Browser Automation'. To the right of the logo are links for 'edit this page', a search bar labeled 'search selenium:' with a 'Go' button, and a menu with buttons for 'Projects', 'Download', 'Documentation', 'Support', and 'About'. The main content area is titled 'Selenium WebDriver' and contains several paragraphs of text and a list of bullet points. On the right side of the page, there is a large Selenium logo with a green checkmark, followed by the text 'Selenium is a suite of tools to automate web browsers across many platforms.' and a 'Download Selenium' button with a downward arrow.

**SeleniumHQ**  
Browser Automation

[edit this page](#) search selenium:

[Projects](#) [Download](#) [Documentation](#) [Support](#) [About](#)

## Selenium WebDriver

The biggest change in Selenium recently has been the inclusion of the WebDriver API. Driving a browser natively *as a user would* either locally or on a remote machine using the Selenium Server it marks a leap forward in terms of browser automation.

Selenium WebDriver fits in the same role as RC did, and has incorporated the original 1.x bindings. It refers to both the language bindings and the implementations of the individual browser controlling code. This is commonly referred to as just "WebDriver" or sometimes as Selenium 2.

*Selenium 1.0 + WebDriver = Selenium 2.0*

- WebDriver is designed in a simpler and more concise programming interface along with addressing some limitations in the Selenium-RC API.
- WebDriver is a compact Object Oriented API when compared to Selenium1.0
- It drives the browser much more effectively and overcomes the limitations of Selenium 1.x which affected our functional test coverage, like the file upload or download, pop-ups and dialogs barrier
- WebDriver overcomes the limitation of Selenium RC's [Single Host origin policy](#)

WebDriver is the name of the key interface against which tests should be written in Java, the implementing classes one should use are listed as below:

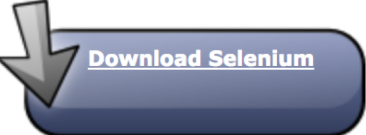
[ChromeDriver](#), [EventFiringWebDriver](#), [FirefoxDriver](#), [HtmlUnitDriver](#), [InternetExplorerDriver](#), [PhantomJSdriver](#), [RemoteWebDriver](#), [SafariDriver](#)

For More information on Selenium WebDriver, please see [the documentation](#) and [Remote Control to WebDriver Migration Notes](#).


**Selenium** is a suite of tools to automate web browsers across many platforms.

Selenium...

- runs in [many browsers](#) and [operating systems](#)
- can be controlled by many [programming languages](#) and [testing frameworks](#).

 [Download Selenium](#)

# Best practice: Retrieving landing page content

Features Pricing Docs More ▾ Log in

## PhantomJS as a Service

PhantomJS Cloud is a crash proof and problem free cloud of Headless Browser instances. Simple to use, no-install required.

### What is PhantomJS Cloud?

A web-browser hosted as a service, using PhantomJS online (Chrome) to render your javascript heavy webpages as HTML, PDF, JPEG, PNG, or JSON.

A high performance Cloud of PhantomJS instances capable of processing more than 1 million requests per hour.

A [Simple and well documented API](#) allowing you to request and obtain your pages.

A Headless Browser Software-as-a-Service (SaaS) that's free for light use, and cheap for heavy use.

### What can you do with it?

Generate ScreenShots of WebSites or HTML you upload. [Example](#)

Extract content from AJAX websites.  
*This example extracts plain-text from Etsy's Pinterest page* [Example](#)

Execute arbitrary JavaScript that modifies the page.  
*This example highlights the word "English"* [Example](#)

Generate and render PDF reports from HTML. [Example](#)

... And More! [Read the Docs](#)

*See the "[Advanced Scenario Samples](#)" section for Page Automation, AutoLogin, etc.*

**Sign up now!**  
Get 500 Pages/Day for free

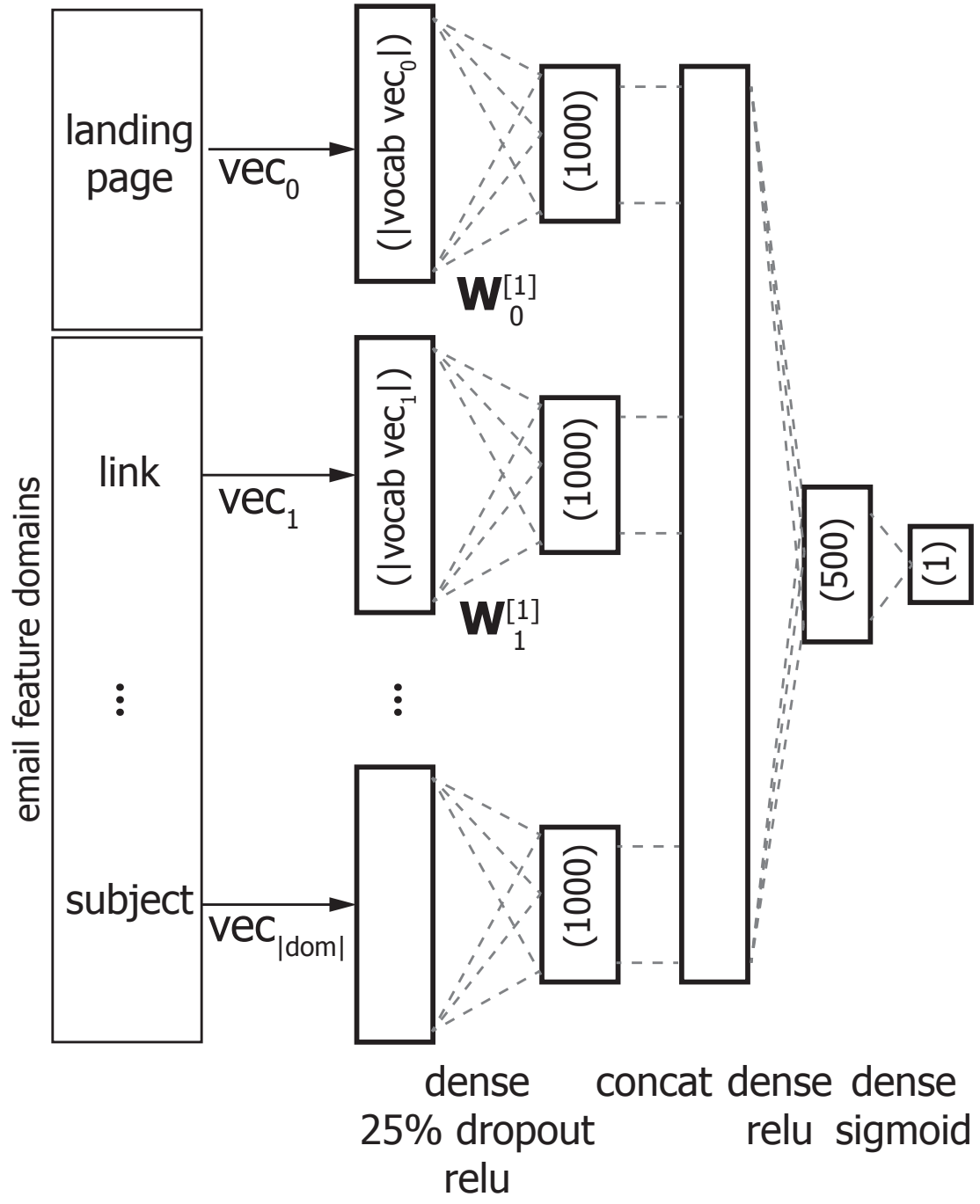
# Best practice: Retrieving landing page content

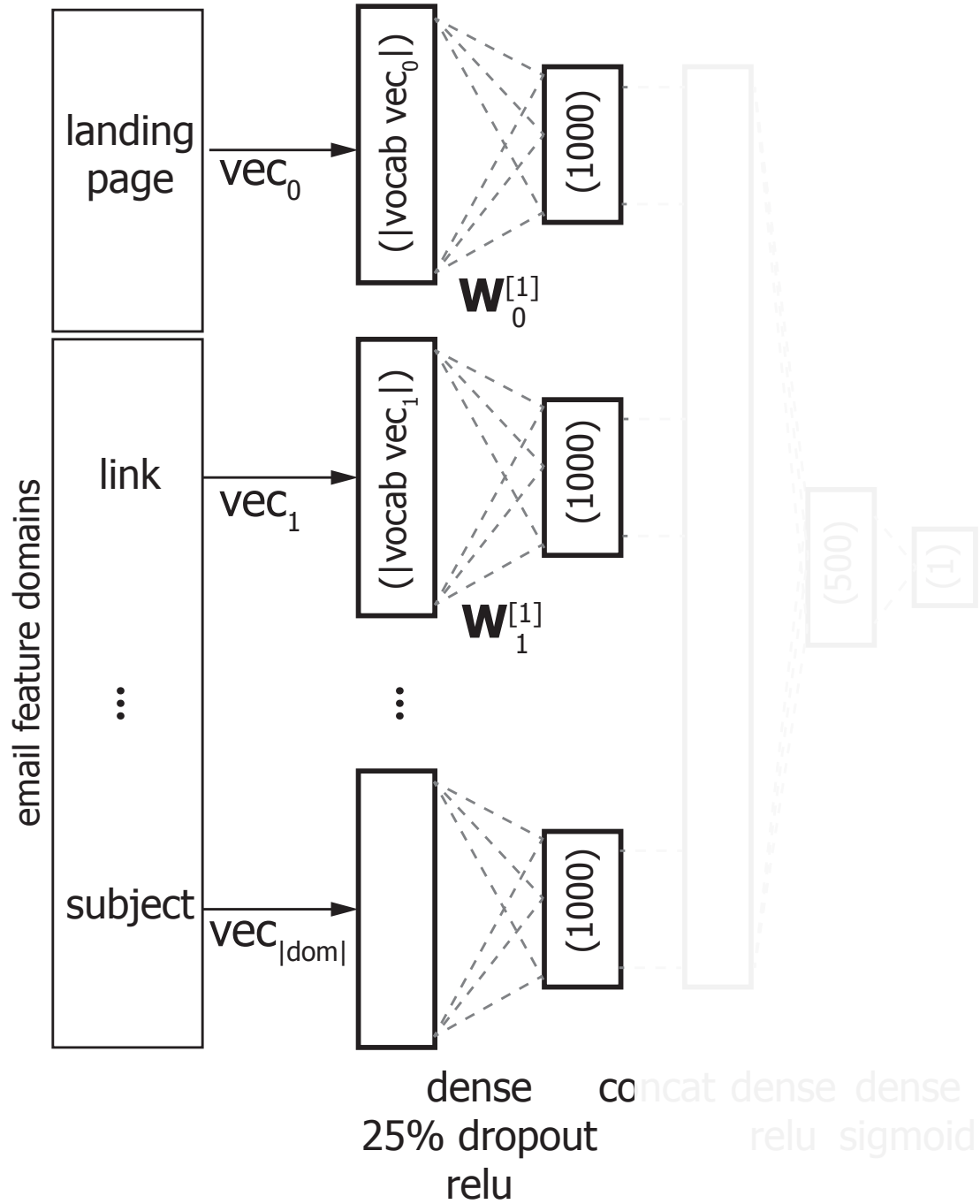
**AppleScript + Firefox CLI**  
for retrieving content and taking screenshots  
of dynamic web pages

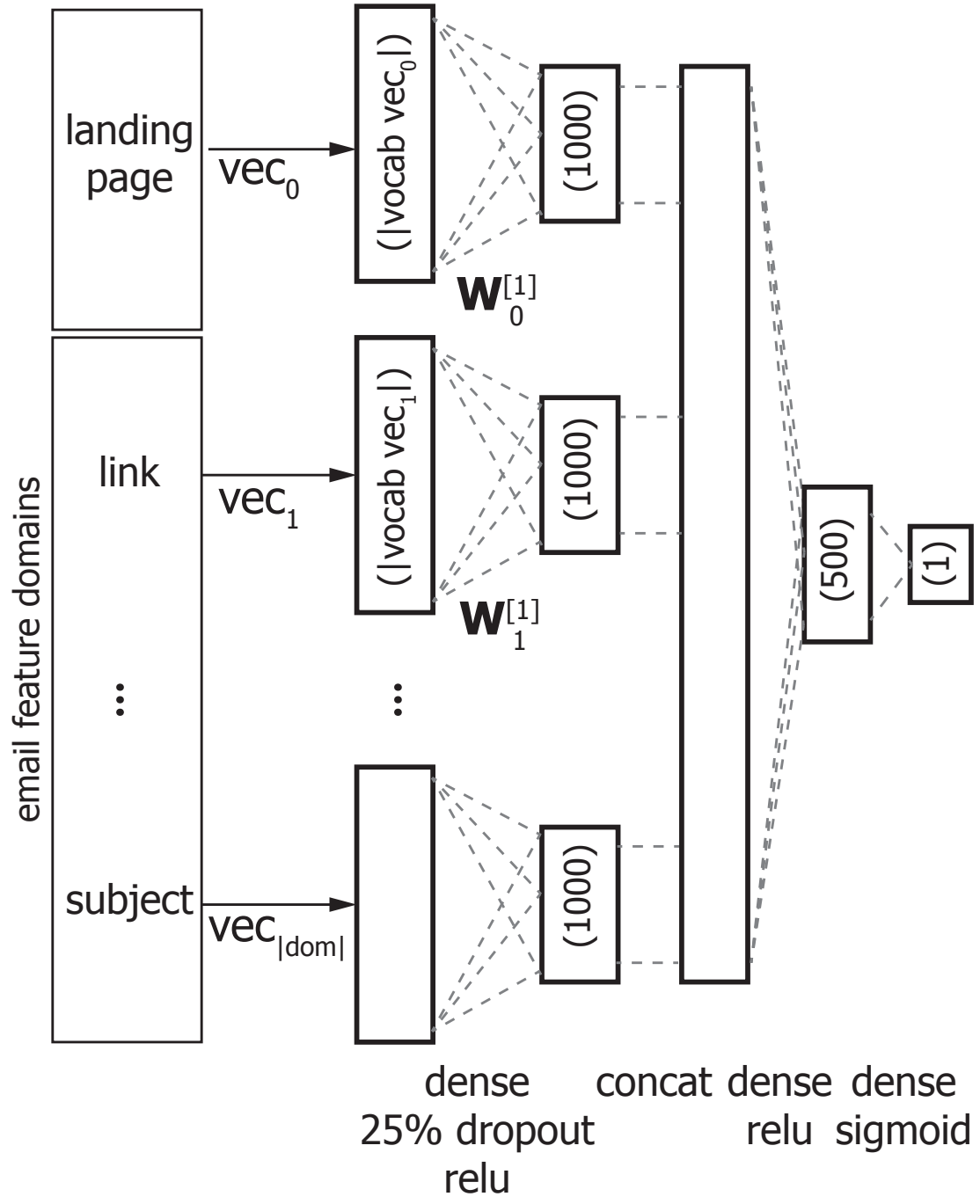
x3 playback

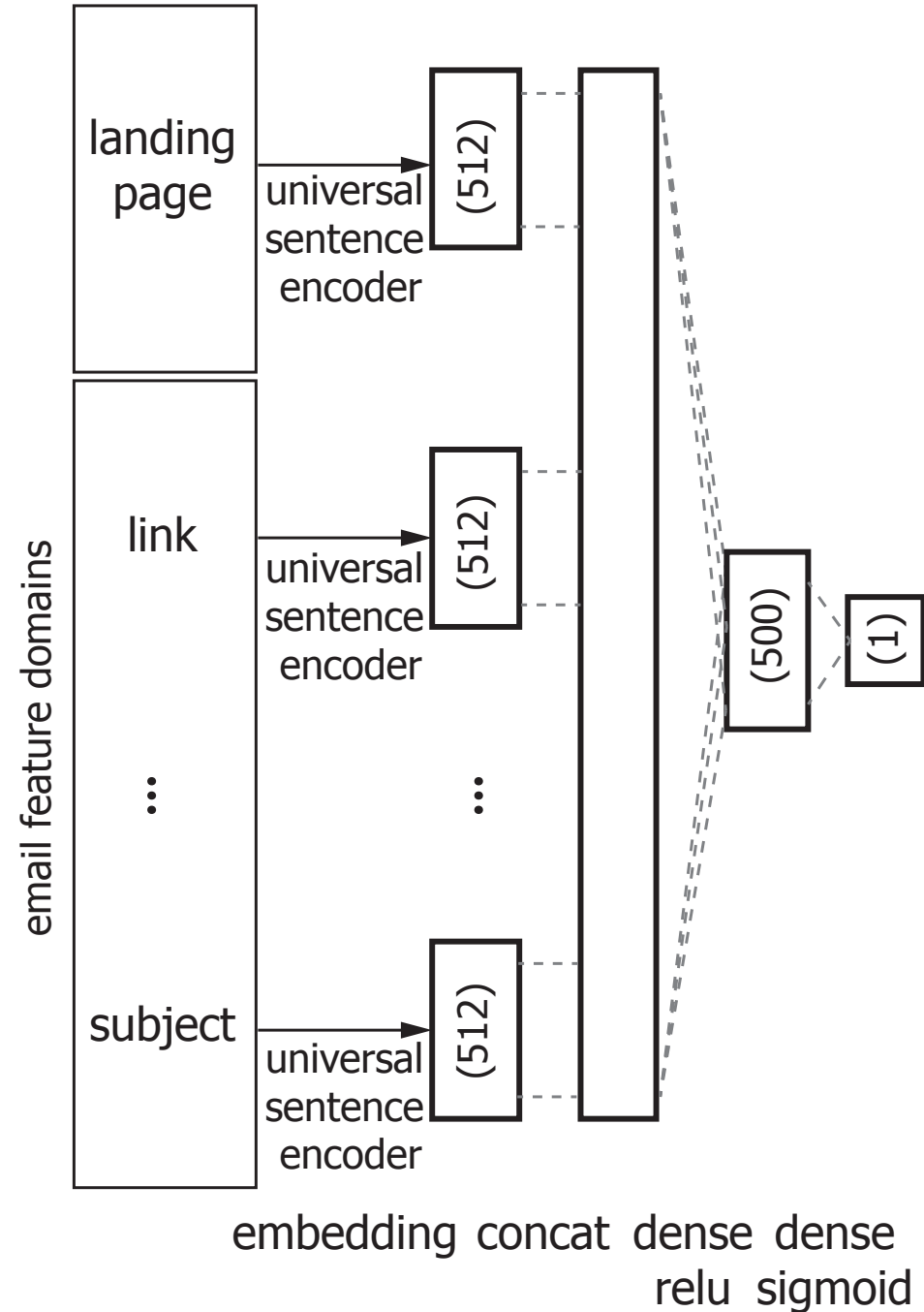
# The Model Zoo



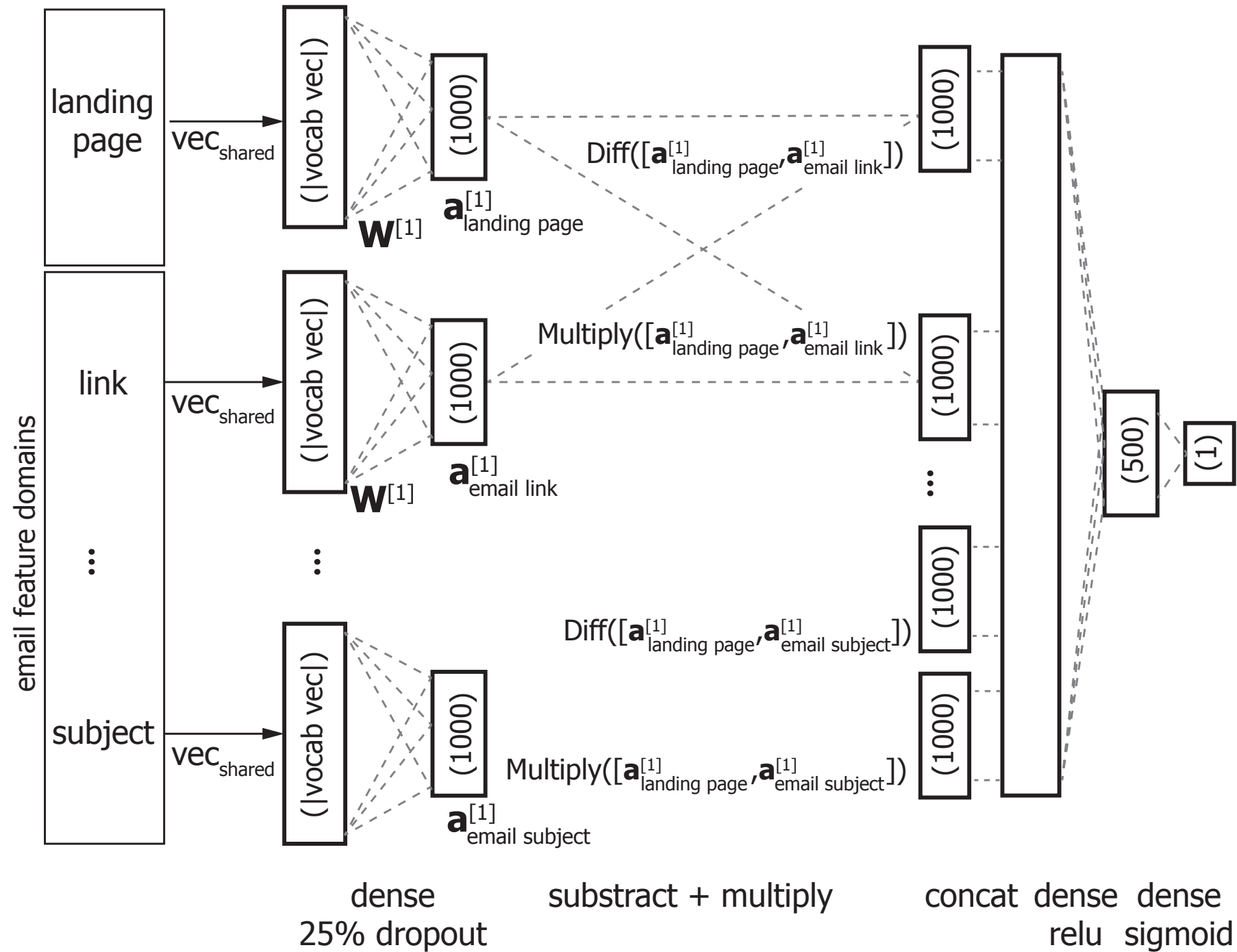


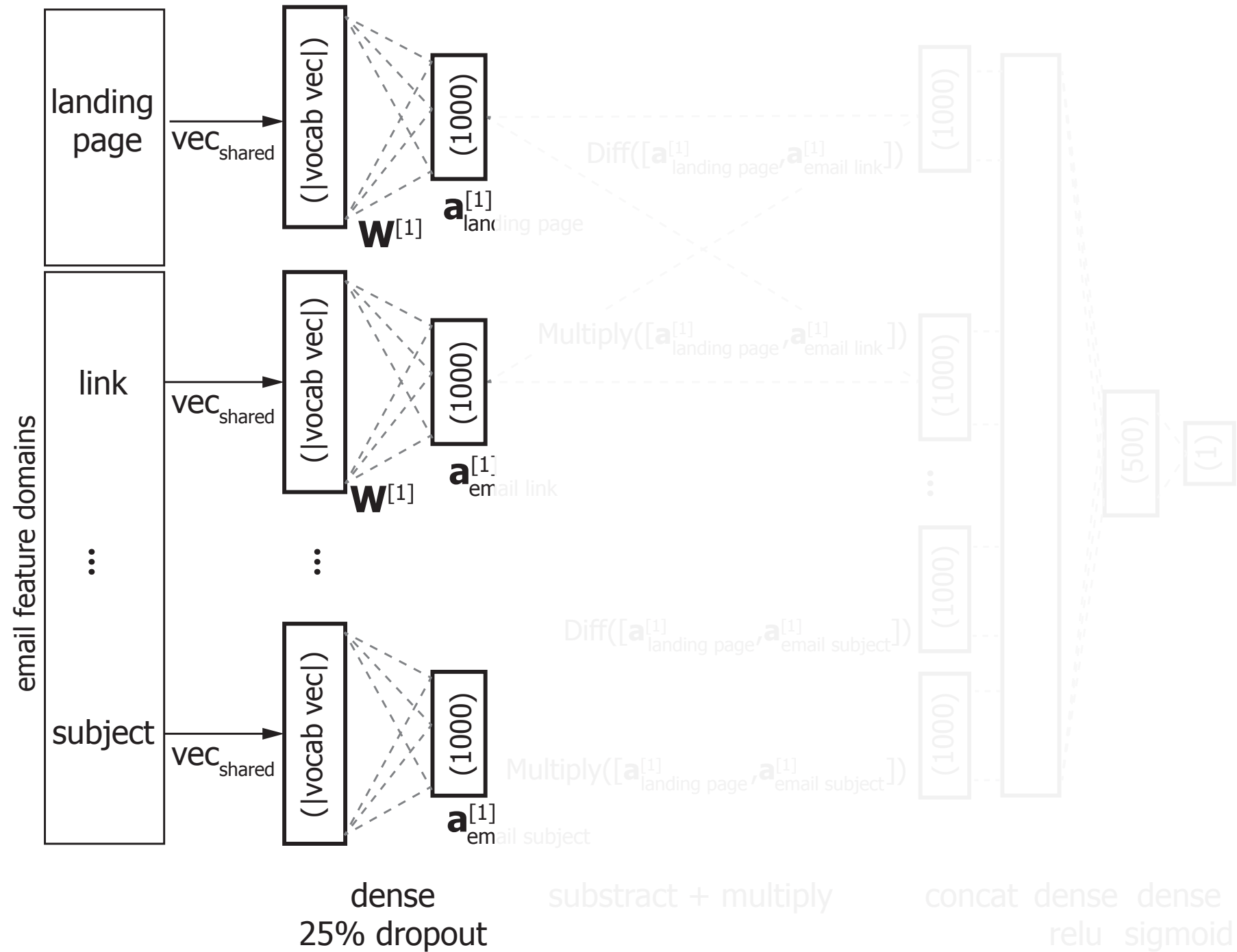


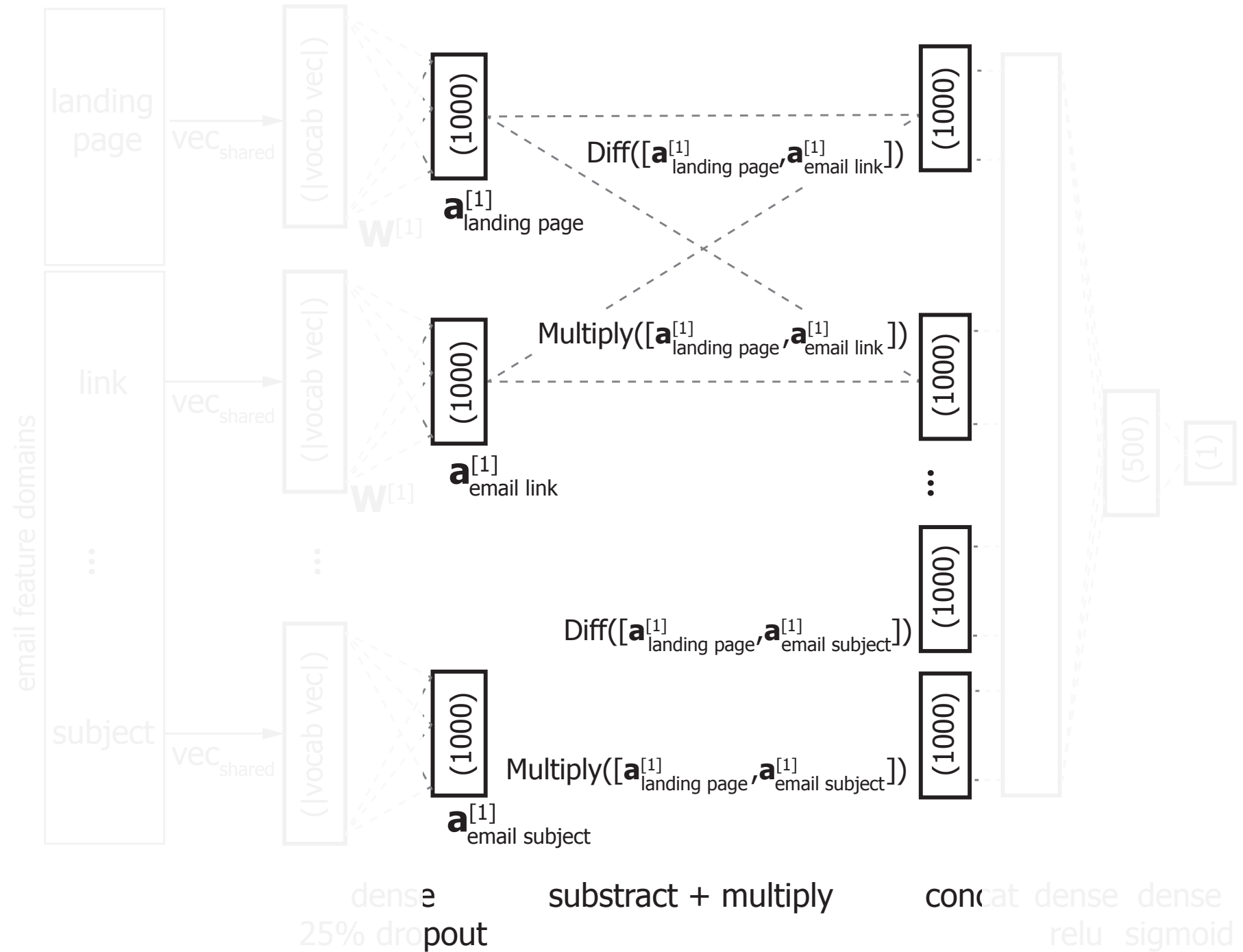












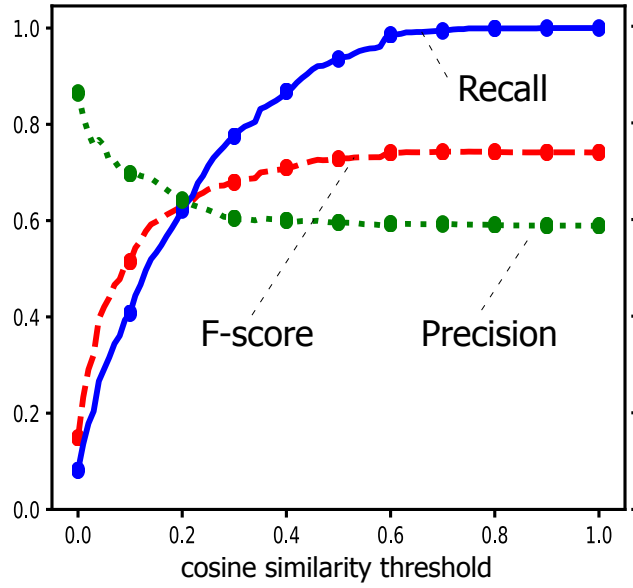
# Outline

- I. Link Issue Taxonomy
- II. Data
- III. Model
- IV. Evaluation**
- V. Recent and Future Work

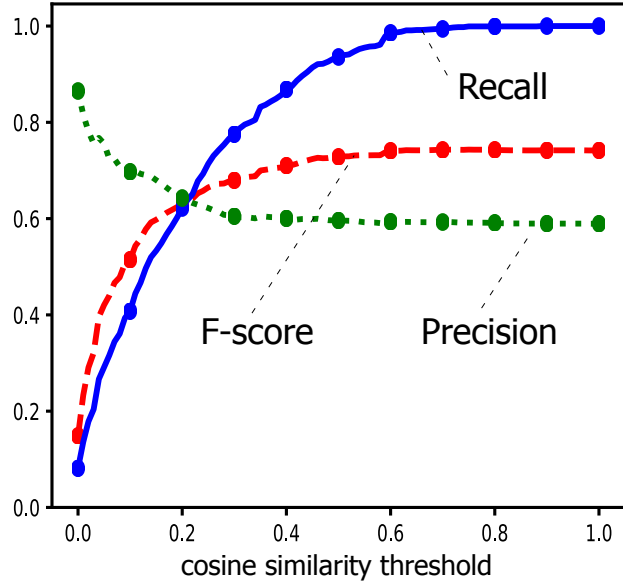


# Unsupervised Baselines

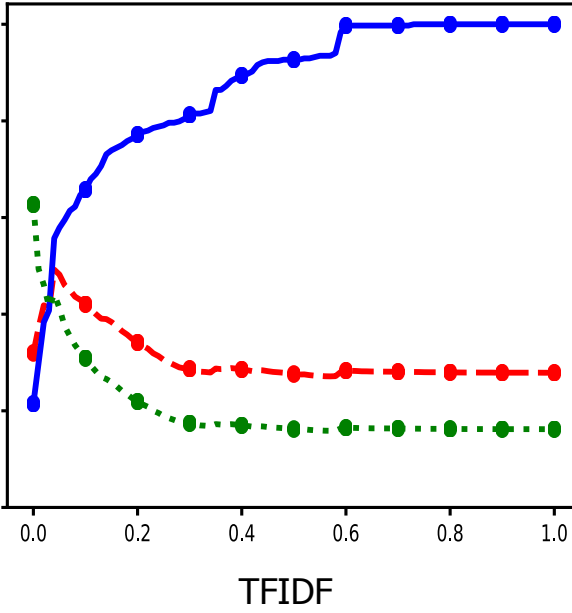
**Target class "misaligned"  
based on 1 vote**



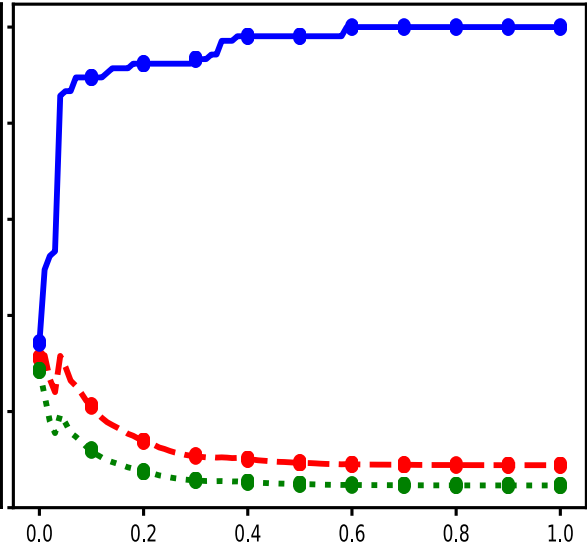
Target class "misalinged"  
based on 1 vote



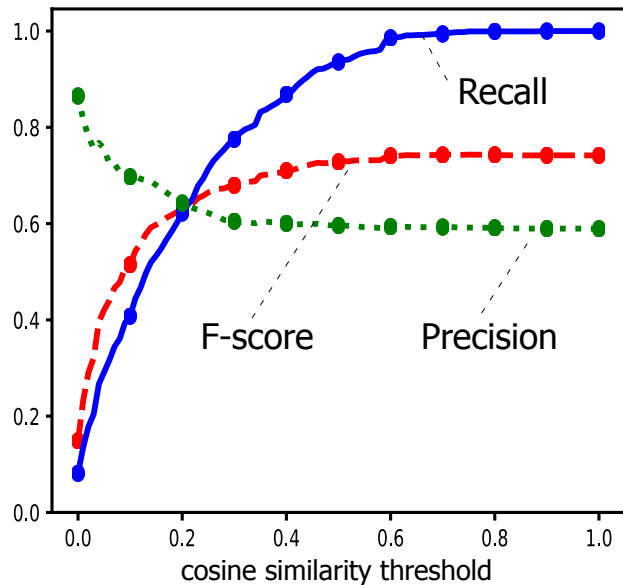
Target class "misalinged"  
based on 3 votes



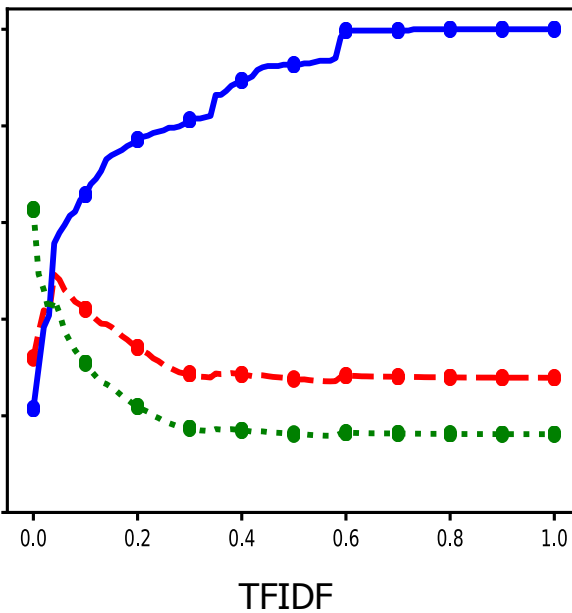
Target class "misalinged"  
based on 5 votes



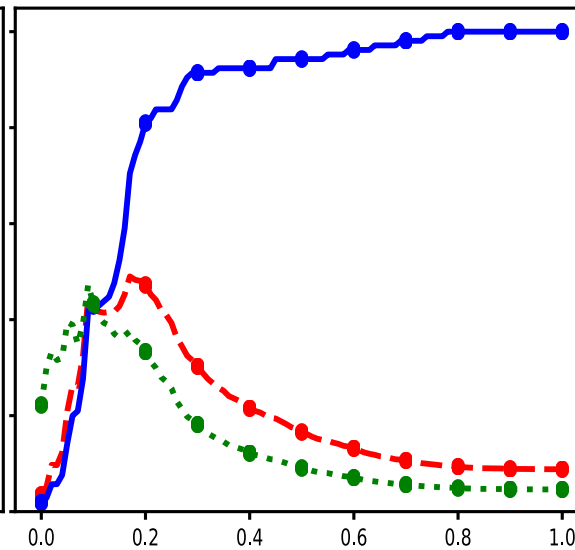
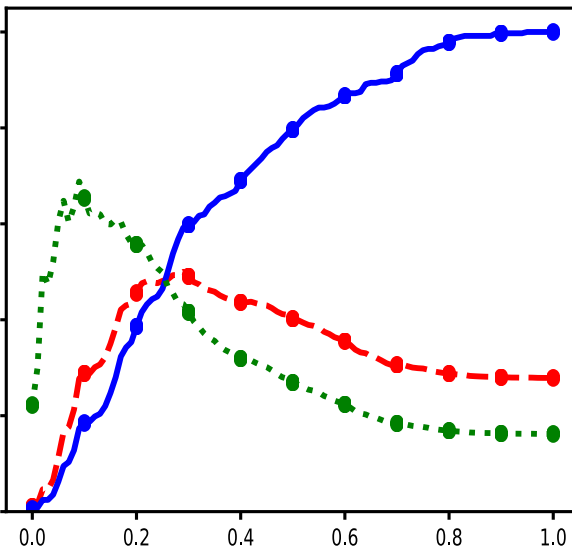
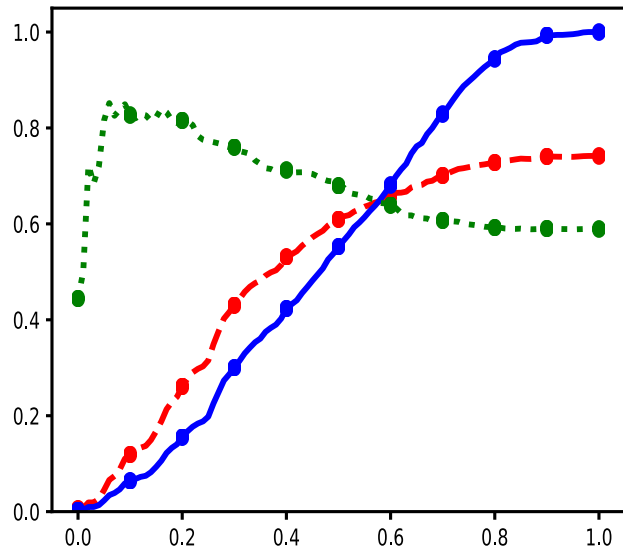
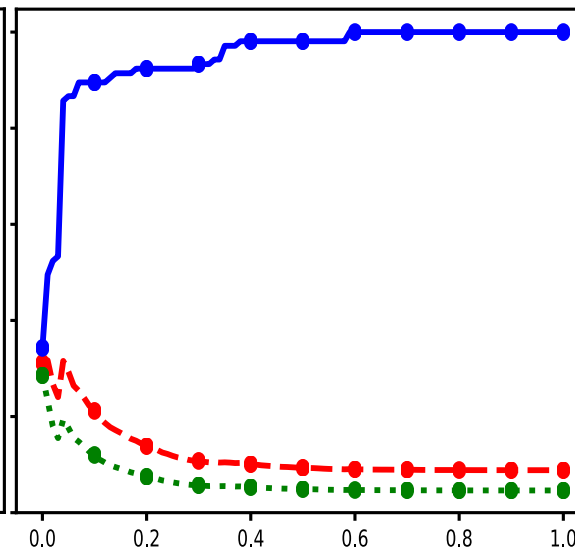
**Target class "misaligned"  
based on 1 vote**



**Target class "misaligned"  
based on 3 votes**



**Target class "misaligned"  
based on 5 votes**



**Universal Sentence Encoding**

Animals from the Zoo



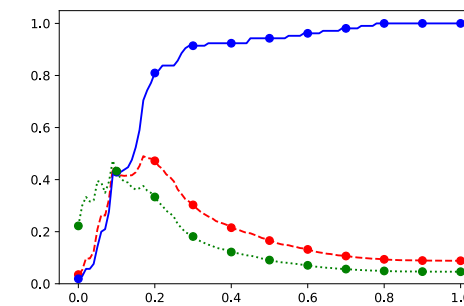
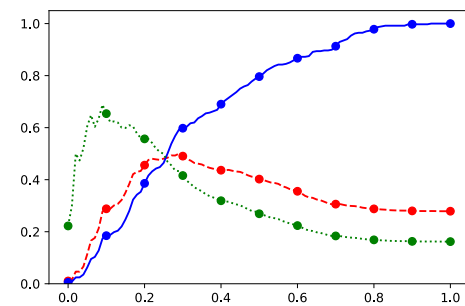
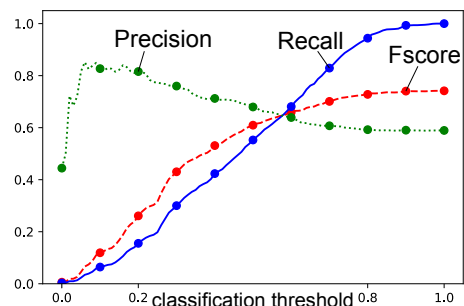


Model

More than 1 Vote  
F P R

More than 3 Votes  
F P R  
all – the page has at least one issue

More than 5 Votes  
F P R



<b>0.75</b>	0.77	0.73
0.72	0.71	0.73
0.74	0.73	0.75

0.58	0.61	0.55
<b>0.63</b>	0.63	0.63
0.51	0.55	0.48

0.66	0.75	0.58
<b>0.68</b>	0.81	0.58
0.66	0.73	0.60



# Outline

- I. Link Issue Taxonomy
- II. Data
- III. Model
- IV. Evaluation
- V. Recent and Future Work

Thank you!

Contact: [lipka@adobe.com](mailto:lipka@adobe.com)