



Predicting Bias in English-Language WIKIPEDIA Articles

Luke Dowker



June 2, 2022

Project Supervisor



Luke Dowker
Data Scientist, Flatiron School

Agenda

01

Business
Problem

02

Data &
Methods

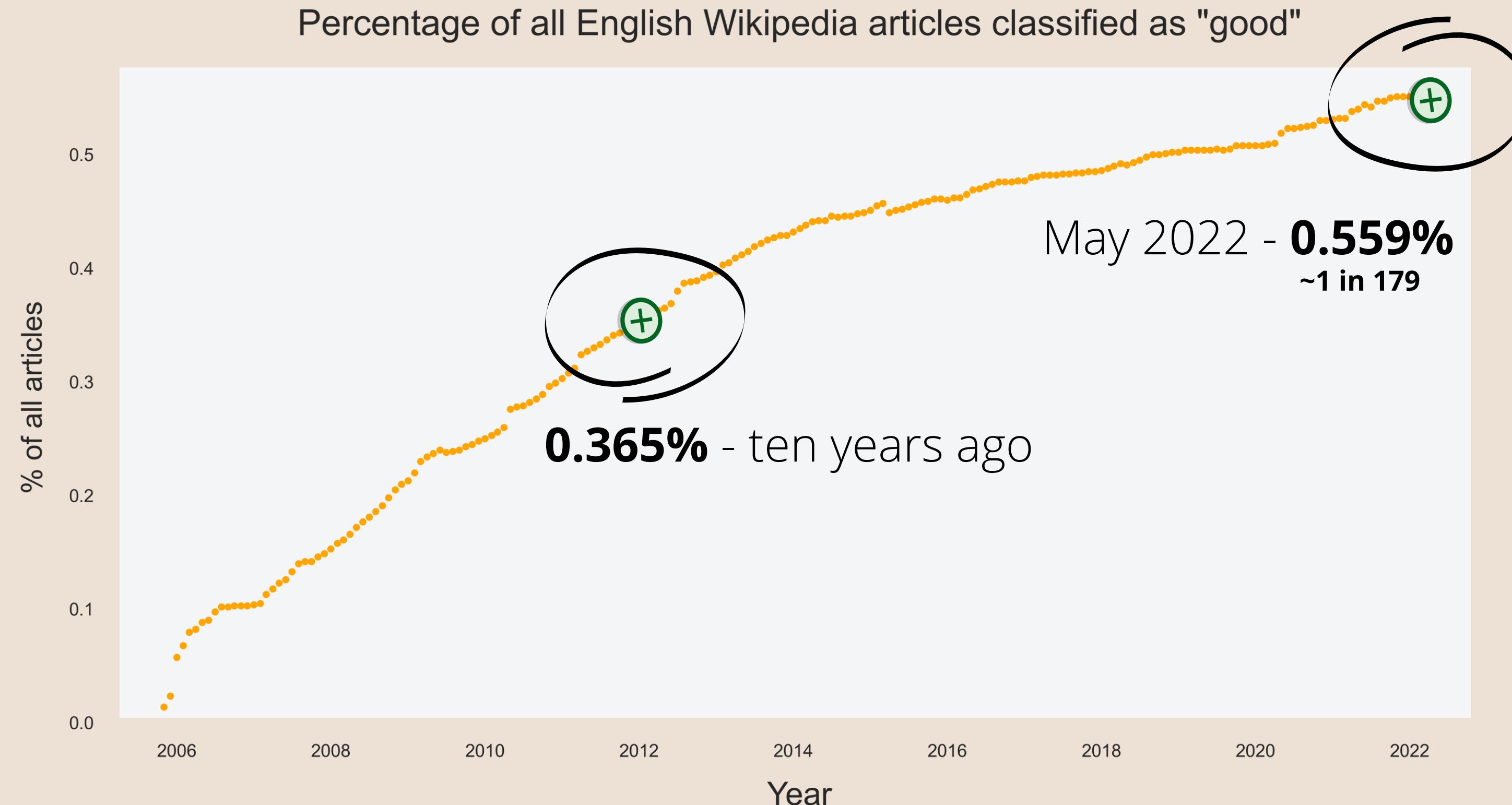
03

Results

04

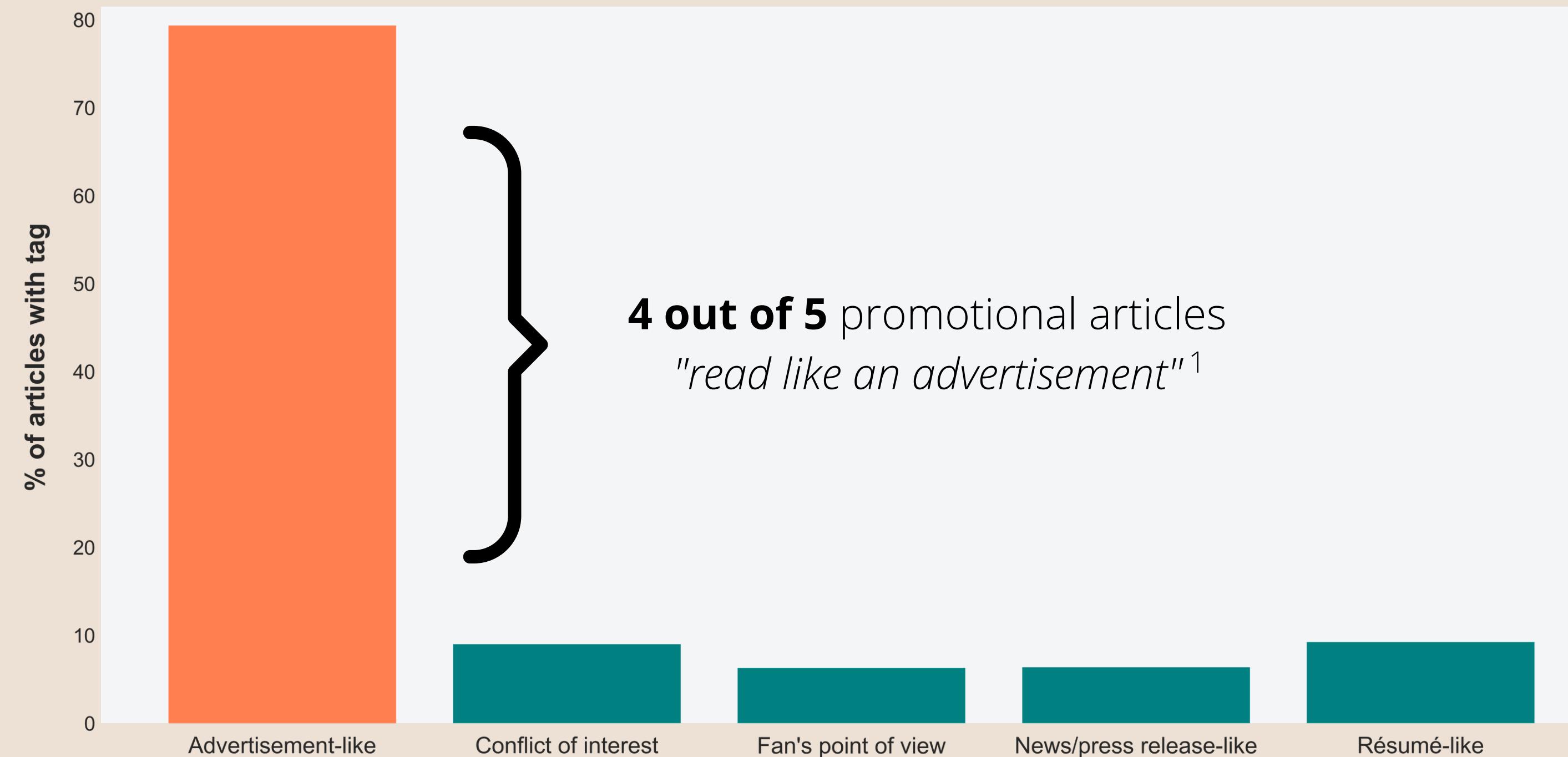
Conclusions
& Next Steps

Article quality: On the up!



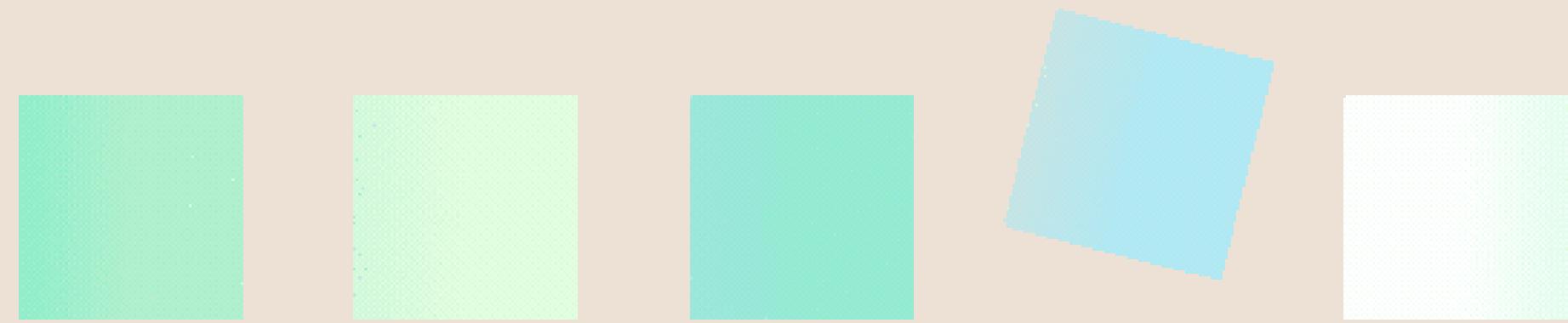
...but not *every* article

Distribution of promotional article subclasses



1. Some articles were marked as belonging to more than one subclass.

Process



Dataset

- Good articles slightly outnumber promotional ones
- Active edits = **evolving data!**
 - ~581 new articles daily²



2. Cited from Wikipedia's internal statistics [page](#).

How it works:

INPUT

raw text, e.g.
Wikipedia article



OUTPUT

prediction: "good"
or "promotional"

Text data: Two approaches



Term count

CountVectorizer

Frequency of terms is the sole metric used for predictions.

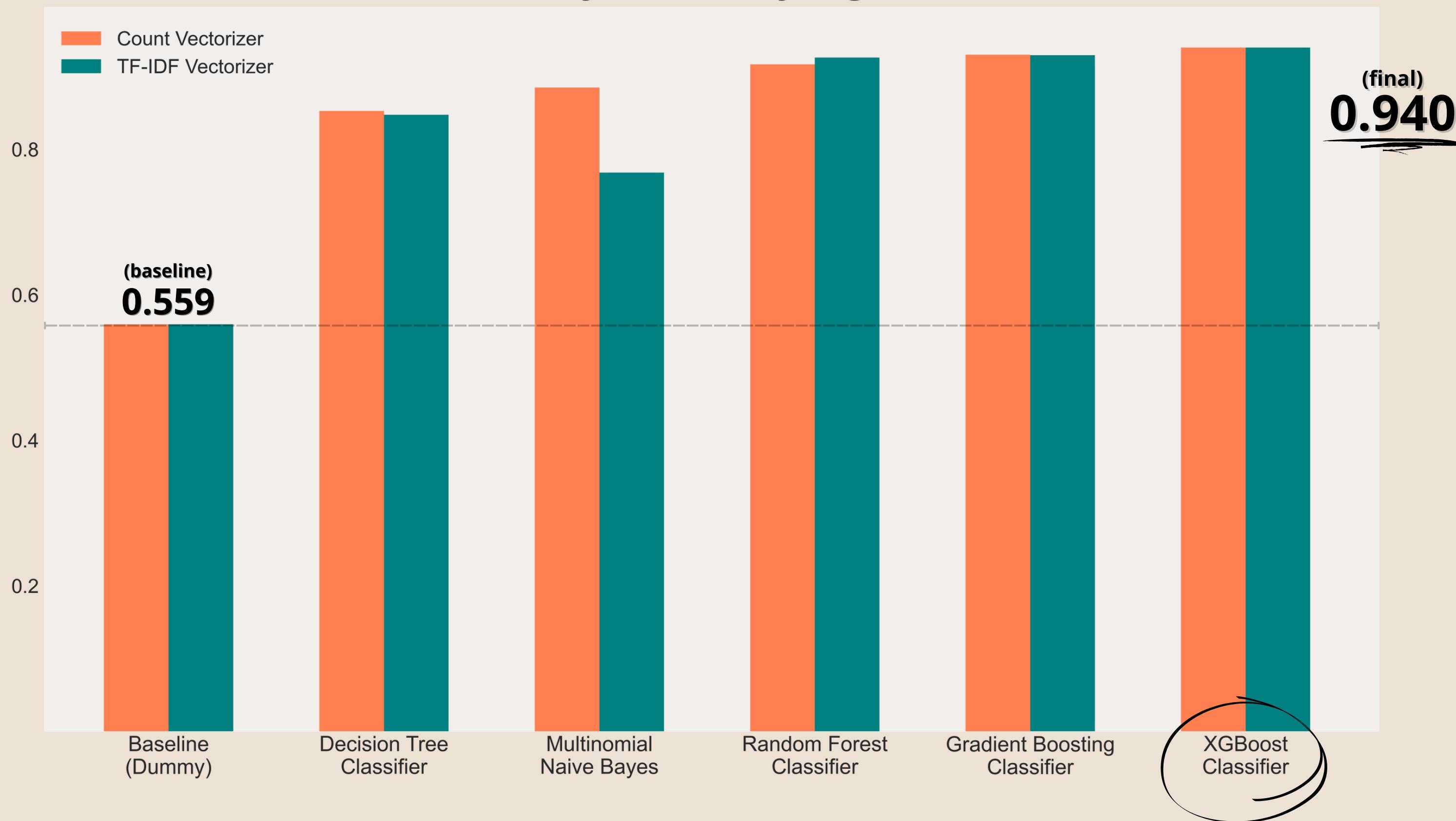


Term importance

TfidfVectorizer

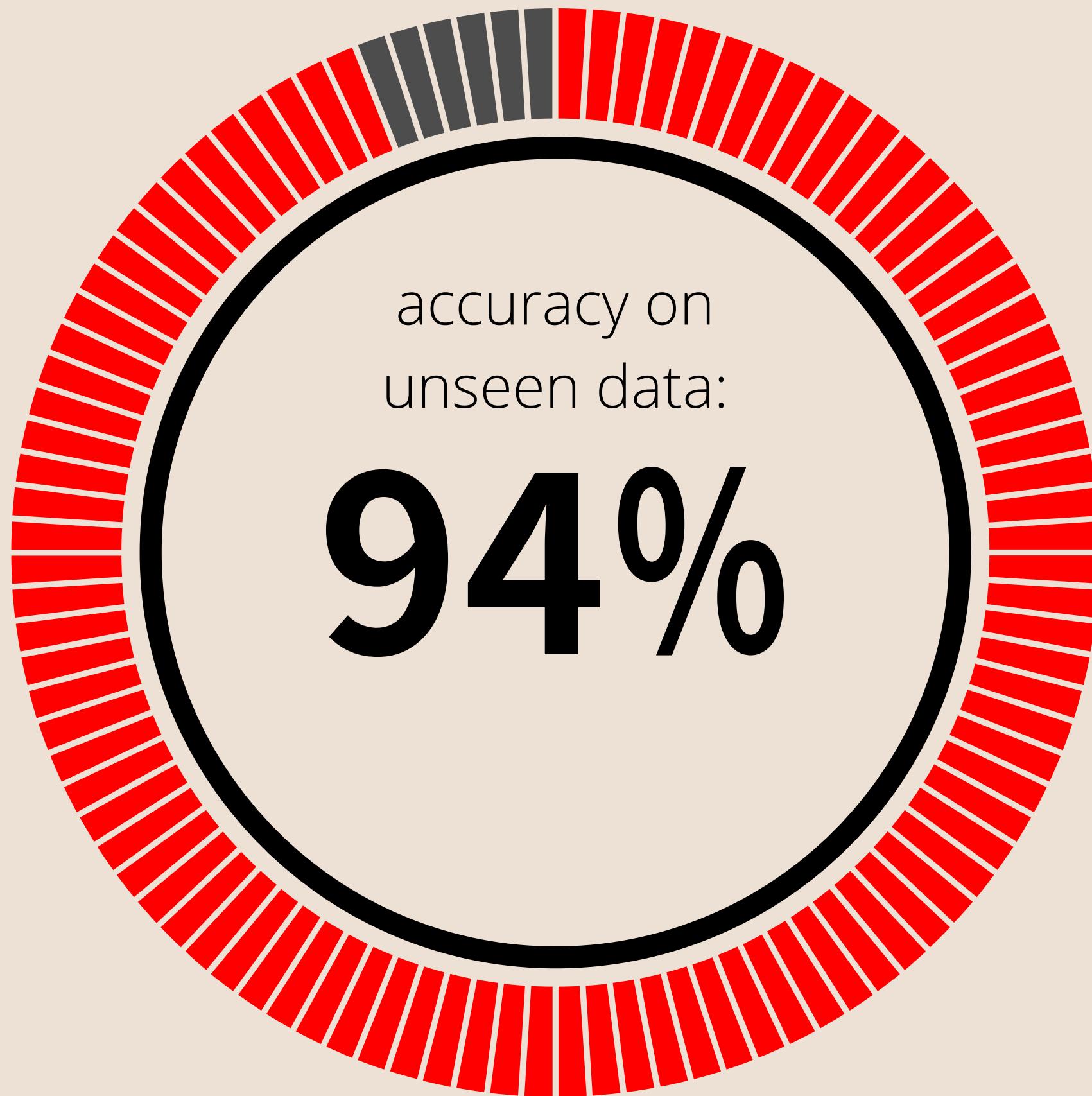
Calculates the **relative importance** of terms to make predictions.

Accuracy scores by algorithm³

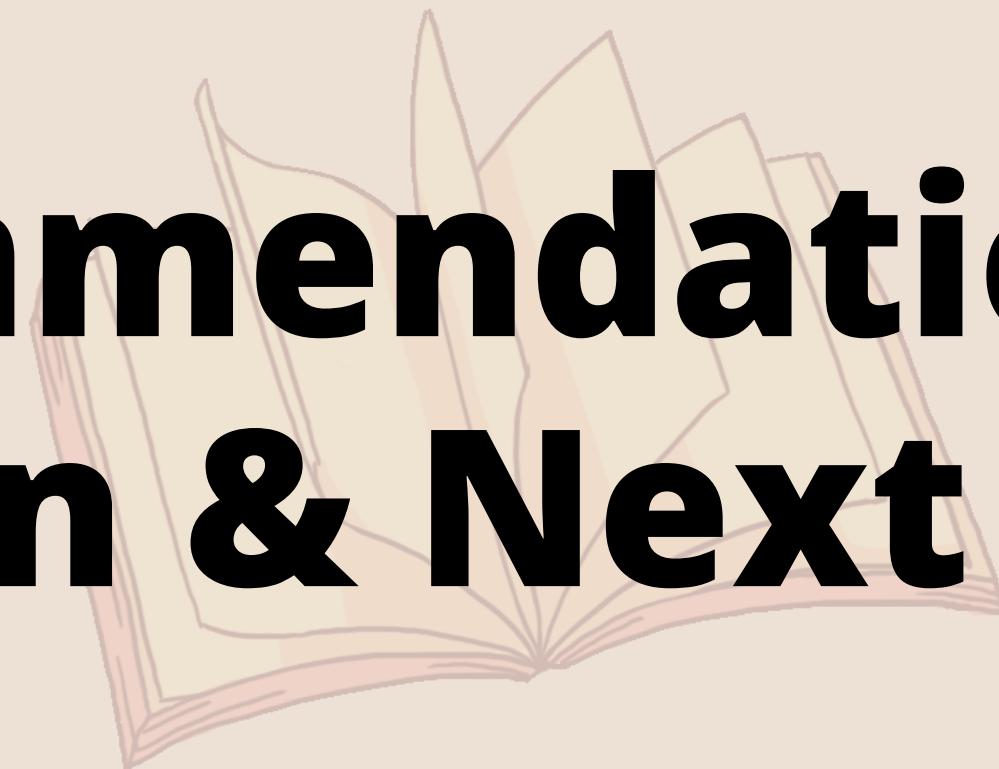


3. Results of 5-fold cross-validation on twice-split data., i.e. a validation test set.

Performance



Recommendations for Action & Next Steps



What now?

- Integrate **Streamlit app** into user interface
- **Pre-classify** newly published content

The screenshot shows a Streamlit application window titled "Is this Wikipedia article up to snuff?". The window has a toolbar with various icons at the top. Below the title, there is a text input field with the placeholder "Find out by pasting the text of your article below!". A sample text block is present with the heading "THIS IS SAMPLE TEXT". The text describes the Papuan mountain pigeon (*Gymnophaps albertisii*) as a species of bird in the pigeon family Columbidae, found in the Bacan Islands, New Guinea, the D'Entrecasteaux Islands, and the Bismarck Archipelago. It details its physical characteristics, including slate-grey upperparts, chestnut-maroon throats and bellies, whitish breasts, and a pale grey terminal tail band. The species is described as frugivorous, feeding on figs and drupes, breeding from October to March in the Schrader Range, and laying a single egg. The text also notes its social nature and flock size. At the bottom of the window, there is a button labeled "Click to run!" and a status message "Looking good!".

Is this Wikipedia article up to snuff?

Find out by pasting the text of your article below!

THIS IS SAMPLE TEXT

The Papuan mountain pigeon (*Gymnophaps albertisii*) is a species of bird in the pigeon family Columbidae. It is found in the Bacan Islands, New Guinea, the D'Entrecasteaux Islands, and the Bismarck Archipelago, where it inhabits primary forest, montane forest, and lowlands. It is a medium-sized species of pigeon, being 33–36 cm (13–14 in) long and weighing 259 g (9.1 oz) on average. Adult males have slate-grey upperparts, chestnut-maroon throats and bellies, whitish breasts, and a pale grey terminal tail band. The lores and orbital region are bright red. Females are similar, but have grayish breasts and grey edges to the throat feathers.

The Papuan mountain pigeon is frugivorous, feeding on figs and drupes. It breeds from October to March in the Schrader Range, but may breed throughout the year across its range. It builds nests out of sticks and twigs in a tree or makes a ground nest in short dry grass, and lays a single egg. The species is very social and is usually seen in flocks of 10–40 birds, although some groups can have as many as 100 individuals. It is listed as Least Concern by the International Union for the Conservation of Nature (IUCN) due to its wide distribution and large population.

Click to run!

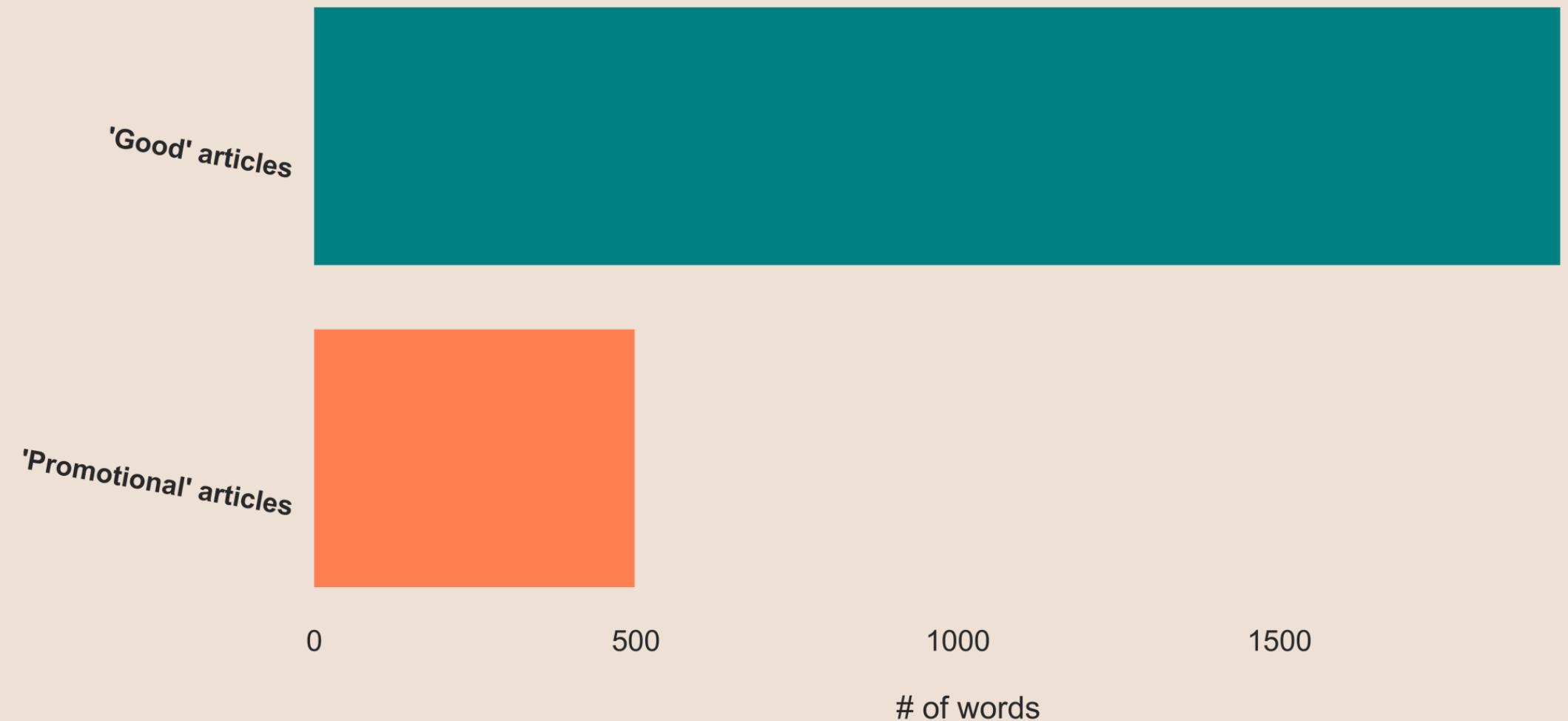
Looking good!

Future development

- Engineer **new features**
- Collaborate with **Fandom** wikis
- Word embedding & **neural networks**
- Create predictions from **URLs** using HTML



Median word count



Thank you!

Contact me:



lhadowker@gmail.com

or on [LinkedIn](#) / [GitHub](#)