



Christianity and Buddhism

| The case of the silly intern

The Problem

After finishing up successful campaigns at the **College Board** and **Zillow**, word of Hoppes LLC's data science prowess has been traveling far and wide. So far and wide that one day you get a call from Canada's largest and longest multi-faith event, the World Religions Conference.

Experiencing GOD in Today's World

Moderator: Phil Ailt | Christianity: Michael Clifton | Islam: Afzal Mirza | Buddhism: B. Sasanasarana | Sikhism: Jasmine Kaur

Baha'i Faith: Jaellayna Palmer | Indigenous: Brenda Jacobs | Judaism: Daniel Maoz | Hinduism: Haripriya Parlvrajika | Humanism: Sasan Patel

Canada's **LARGEST & LONGEST** Multi-Faith Event

www.WorldReligionsConference.com

The graphic features a dark blue background with a star-filled galaxy. In the foreground, there is a black silhouette of a person standing on a small hill. To the right, there is a circular logo for "World Religions Conference" featuring a stylized tree with various religious symbols (including a cross, a crescent moon, and a Star of David) and the Arabic word "Allah". The bottom right corner contains text about the conference being the "LARGEST & LONGEST" and includes a website URL.



The Intern: Johnny Doe

Johnny Doe's job was to work with IT to set up religion specific email addresses and update the marketing materials with those email addresses so the thousands of conference attendees could follow with presenters and organizers. This is especially important as this is the year's flagship event that generates a disproportionate amount of connections, and it is imperative that the religion divisions be able to respond quickly.

Johnny was feeling the pressure, and he was in such a rush that he forgot to update the materials for the Buddhism and Christianity, religions that total a combined 2.7 billion devotees world wide. These events usually generate 1,000s of emails and now they will be jumbled and unorganized in a general inbox. The Christianity and Buddhism divisions are freaking out, and needless to say Johnny was relegated back to coffee duty for the remainder of his summer.



r/Christianity

170k

Subscribers

355

Online

/r/Christianity is a subreddit to discuss Christianity and aspects of Christian life. All are welcome to participate.

SUBSCRIBE

CREATE POST

COMMUNITY DETAILS



r/Buddhism

158k

Subscribers

204

Online

A reddit for all kinds of Buddhist teaching

SUBSCRIBE

CREATE POST



The Solution

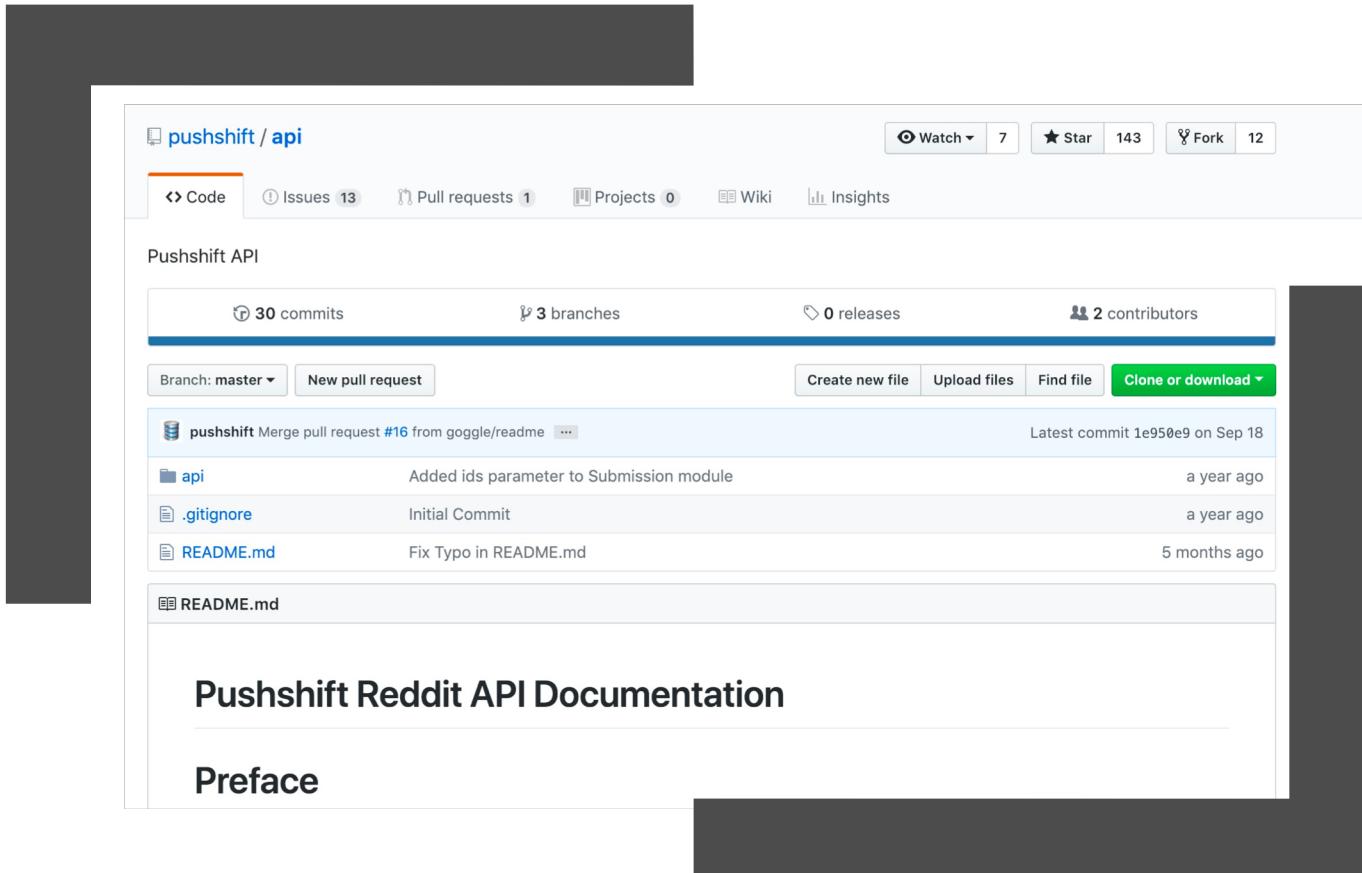
- When they called Hoppe's LLC, they stressed the problem was extremely urgent and that we needed a classification algorithm to be working right away.
- Because you don't have access to their emails, and the request is time sensitive, you have another idea. You check reddit, and it looks like they have about 150-170K subscribers a piece and are fairly active. You explain to the client that this is a robust enough set of posts to train a model on.
- Being a religious organization and knowing Reddit's reputation they were squeamish, but obliged.

The Collection

r/Buddhism

6,578 posts

Over 285 days



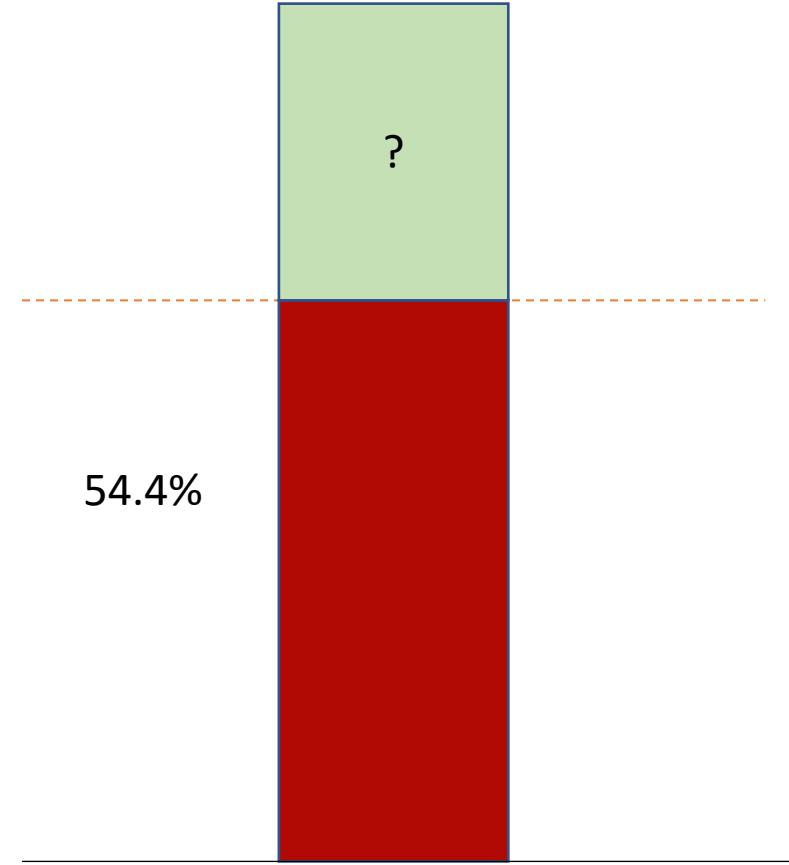
r/Christianity

7,857 posts

Over 165 days

Baseline

The baseline is the expected accuracy we would get if we guessed the most common outcome, Christianity every time.



Top 10 Words

	Count
one	3,549
buddhist	2,904
feel	2,825
life	2,800
thing	2,691
know	2,612
time	2,535
people	2,388
buddha	2,309
mind	2,260

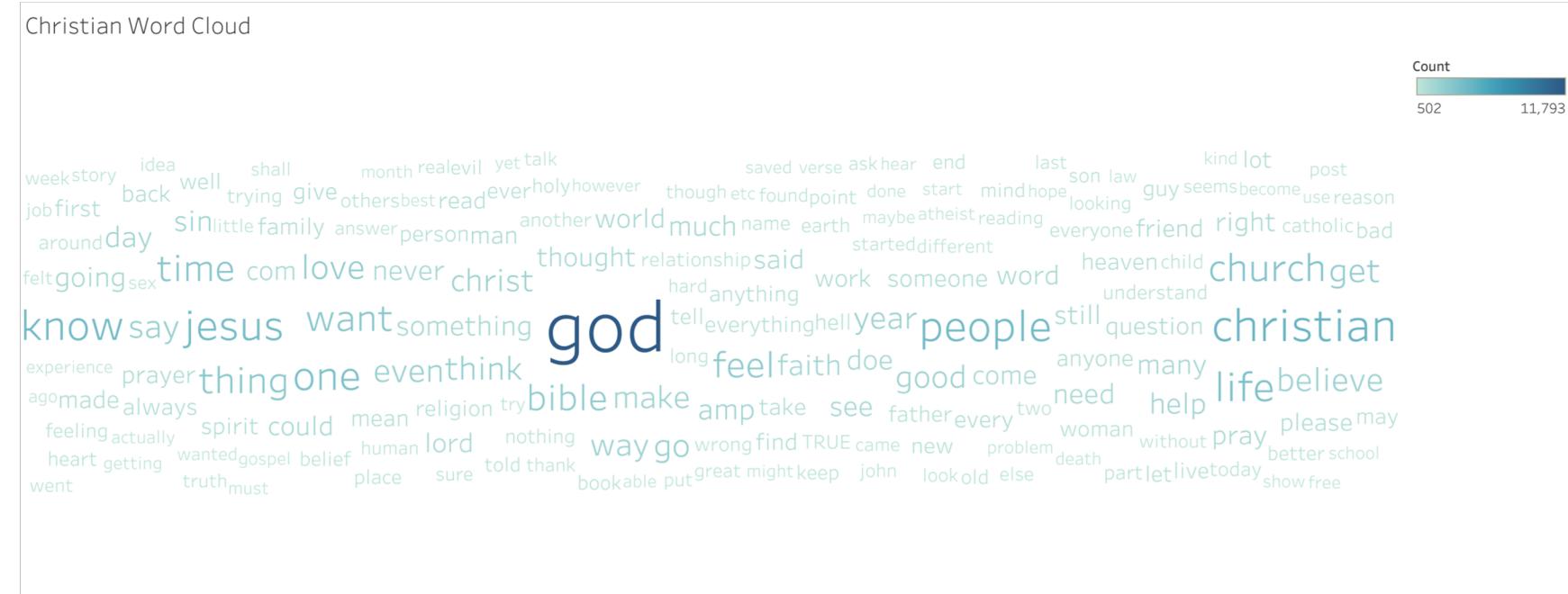
Buddhist Word Cloud



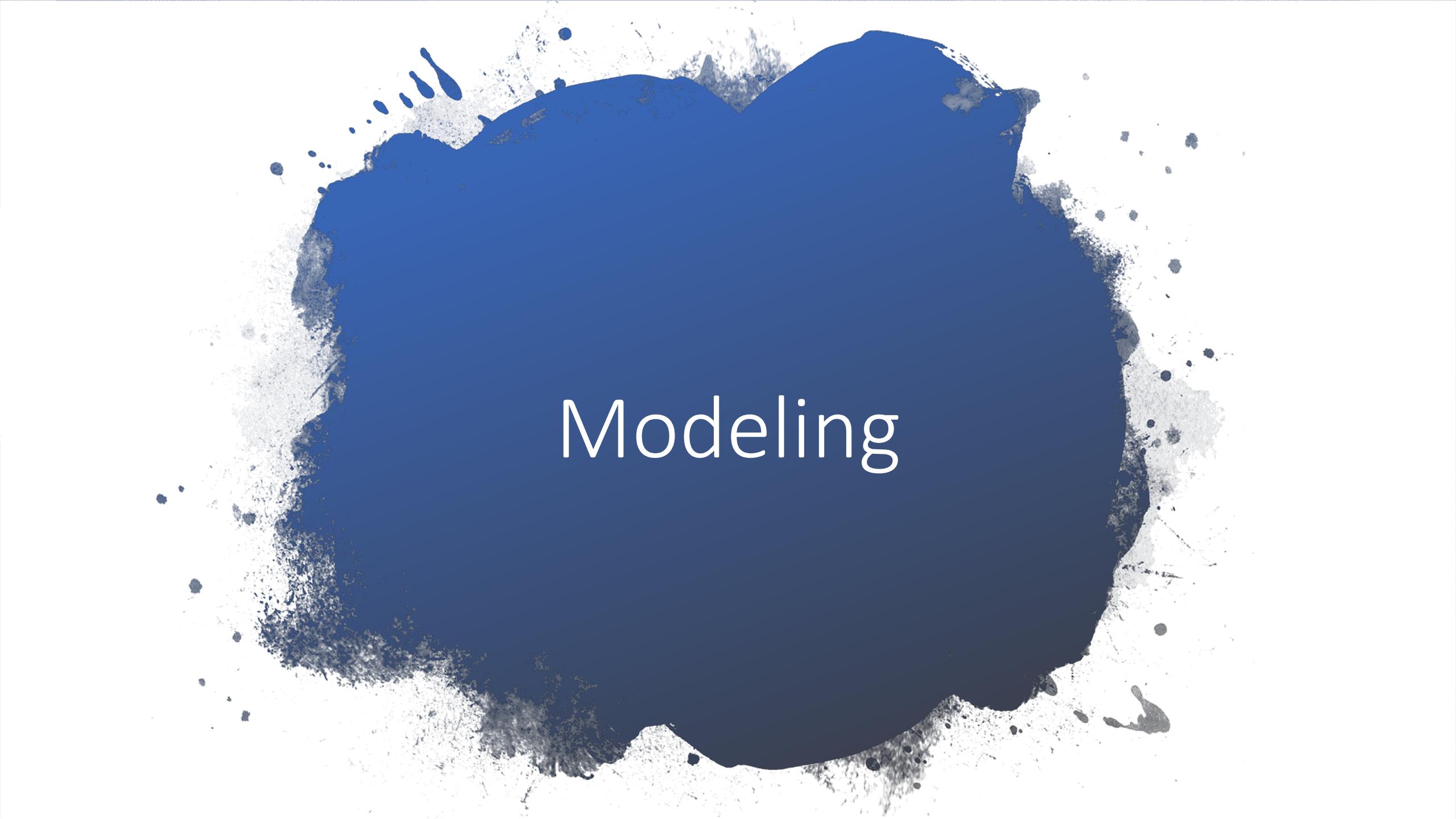
Analysis – Top Words - Buddhism

Top 10 Christian Words

	Count
god	11,793
christian	4,562
people	4,114
know	4,037
jesus	4,004
life	3,541
one	3,535
church	3,302
time	3,072
want	2,933

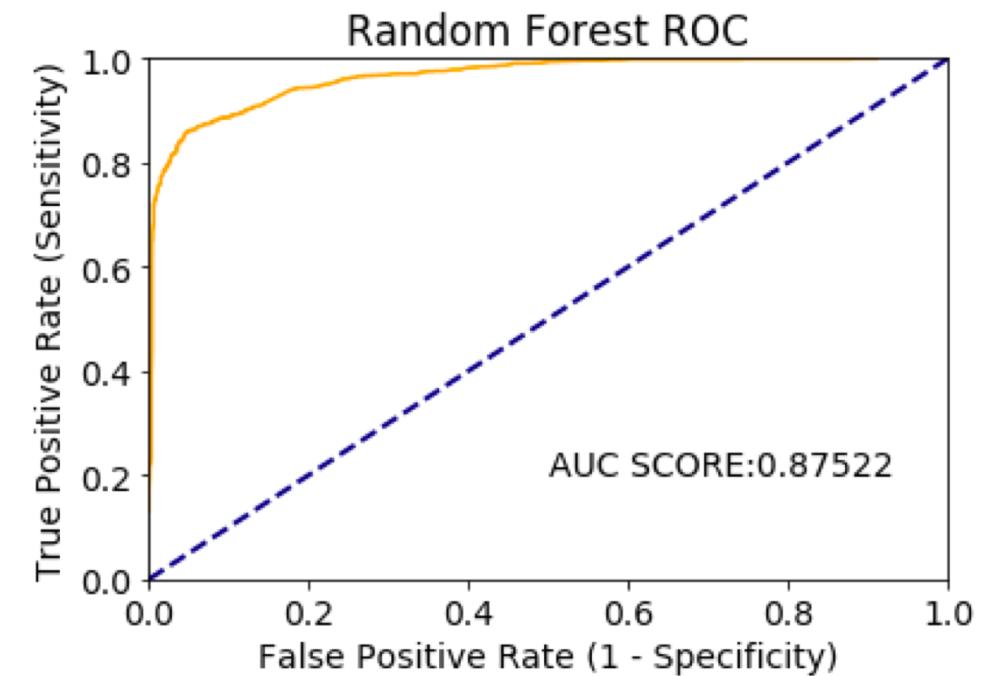
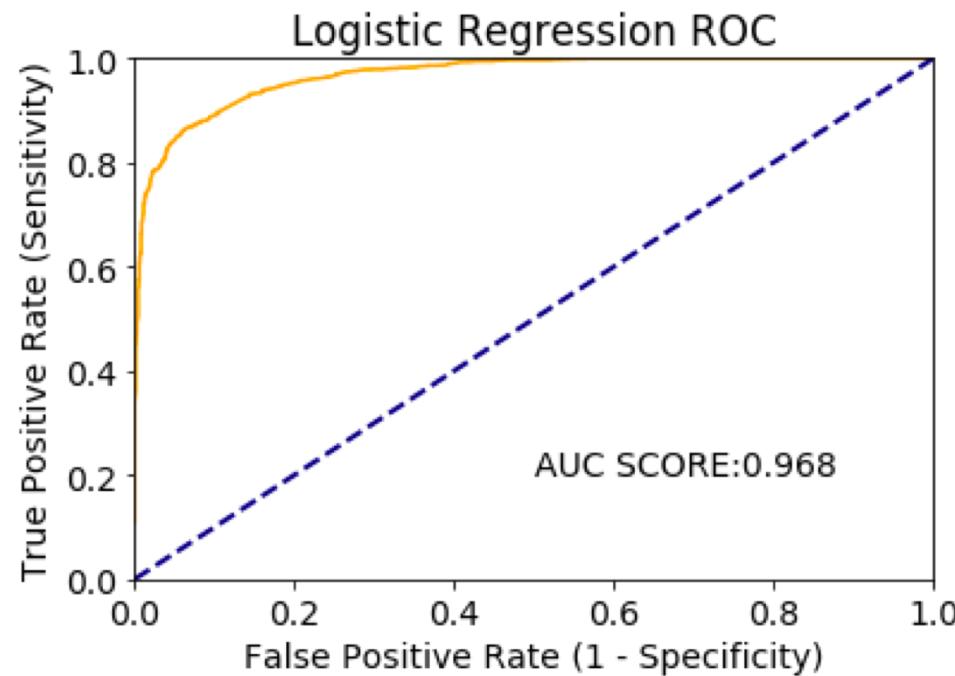


Analysis – Top Words - Christianity



Modeling

Logistic Regression vs Random Forest



Deeper Classification Scores

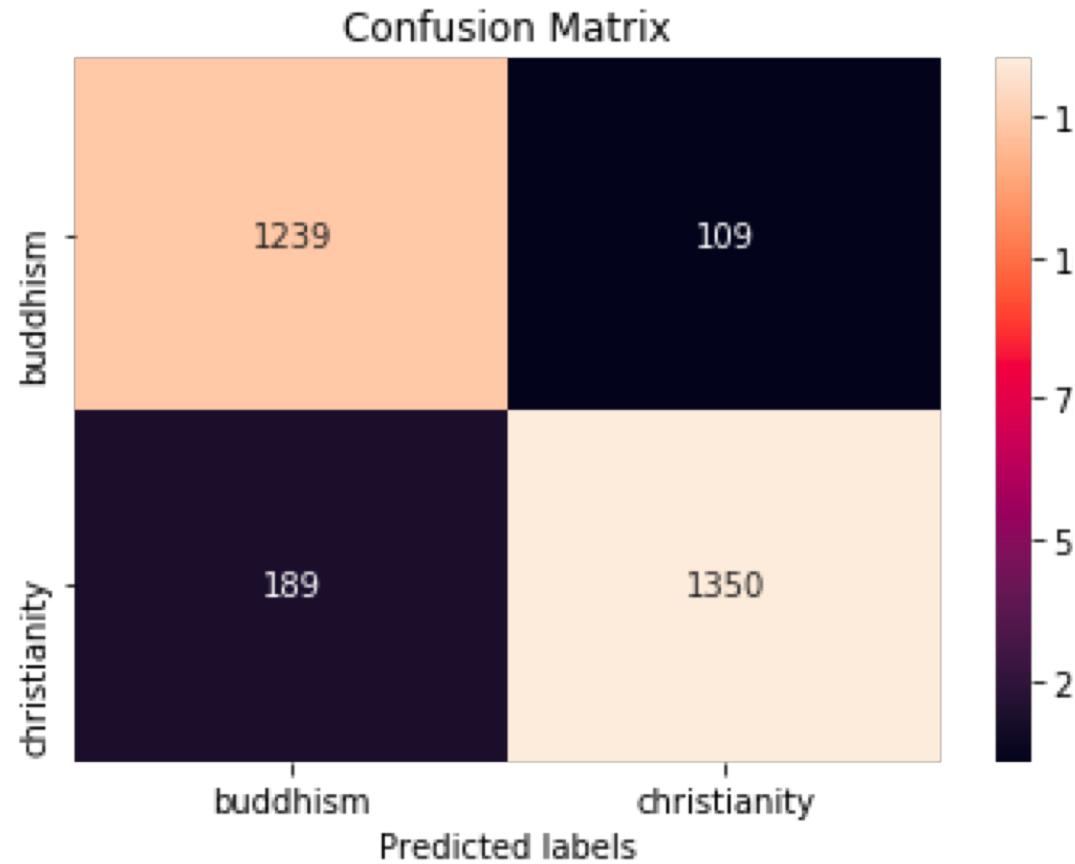
Logistic Regression

Accuracy = 89.7 %

Specificity = 91.9%

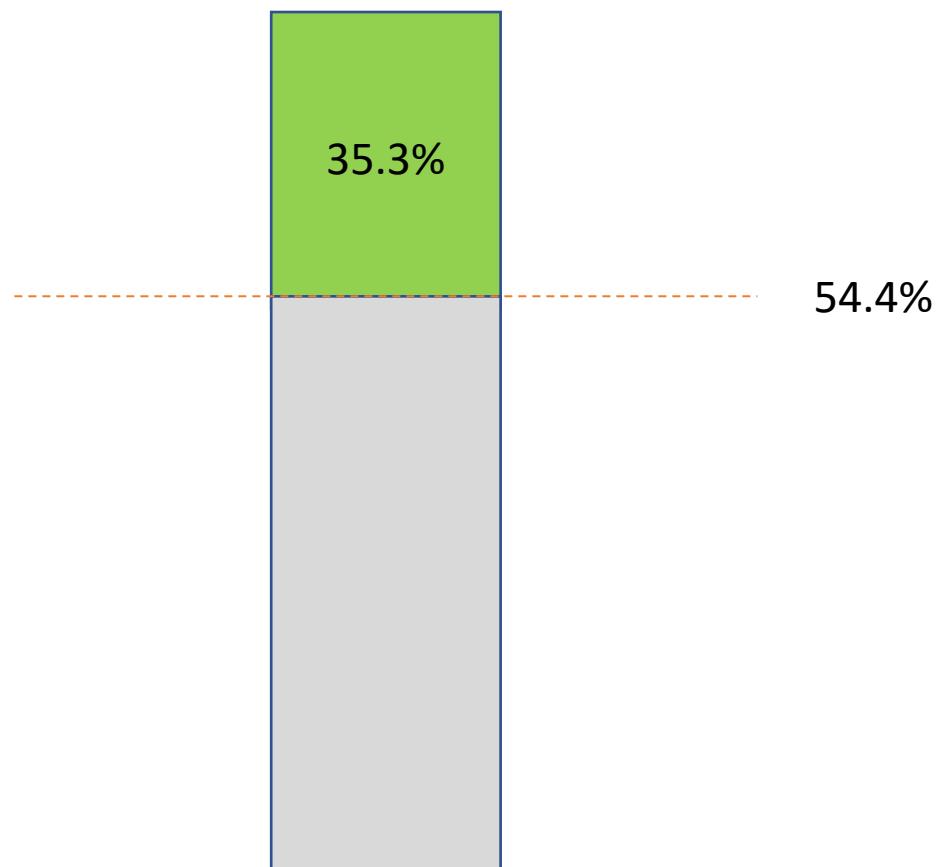
Sensitivity = 87.7%

Misclassification Rate = 10.3%



Final Result

Our **accuracy rate** for the Logistic Regression **stood at 89.7%** on our test data. This was a full **35.3% points above our baseline.**



This is significant, because it means that roughly 9 out of 10 emails we can sort correctly. While 1 out of every ten pieces of email may get sorted into the wrong box, this is a much more manageable total than our original 4.5 out of ten emails if we were to pick all Christianity, and much faster than having a human sort through every one.