Peer-graded Assignment: DTSA 5798 Supervised Text Classification for Marketing Analytics Final Project

Submit by Apr 11, 12:29 PM IST

It looks like this is your first peer-graded assignment. Learn more

Submit your assignment soon
Even though your assignment is due on Apr 11, 12:29 PM IST, try to submit it 1 or 2 days early if you can. Submitting early gives you a better chance of getting the peer reviews you need in time.

Instructions | My submission | Discussions

Imagine you're working at a media buying company, Chrishare. They have a new client, Theragun. Your challenge now is to build a deep learning algorithm that predicts the probability that a news story is about health and wellness. You'll be using the k-train, which is a wrapper for Tensorflow, Keras, and Huggingface Transformers.

Review criteria

less

Your submission will be graded by 3 of your peers. You'll be graded on:

1. You building a replicable model that can be rerun via python without errors
2. Your ability to report and interpret validation set metrics
3. The quality and professionalism of a 5-minute video overviewing what you did
4. How detailed you were at tuning parameters, exploring preprocessing techniques and using transformers/embeddings
5. How predictive it is compared to your classmates

Step-By-Step Assignment Instructions

less

Imagine you're working at a media buying company, Chrishare. They have a new client, Theragun.

Theragun knows that consumers who value health and wellness are more likely to consider, and ultimately buy their product. So, they'd like to find health and wellness news around the web to advertise on. Their goal with their media campaign is to identify as many news articles that mention health and wellness as possible.

This is called contextual advertising: finding the URLs that match the context in which you'd like your ad to be shown. If you want to learn more about contextual advertising, I humbly recommend my non-credit Digital Advertising Strategy specialization on Coursera!

Your challenge now is to build a deep learning algorithm that predicts the probability that a news story is about health and wellness. You'll be using the k-train, which is a wrapper for Tensorflow, Keras, and Huggingface Transformers.

–

The Data

Files:

http://128.138.93.164/news_category_trainingdata.json

–

Background:

Contextual advertising is an important aspect of digital advertising. Companies like Oracle Data Cloud have solutions that allow advertisers to advertise on web pages that match specific types of content. We're starting a company right here at CU Boulder that does contextual advertising better than the big guys do. Why? Because instead of using unsupervised machine learning, which we've learned is less than perfect, we're using supervised machine learning, specifically deep learning.

So we're going to build AI that detects whether a web page mentions health and wellness news content.

Rishabh Misra, a Kaggle Expert and machine learning engineer in San Francisco, California, open-sourced a data set that can be used to roll our own contextual advertising service in a few lines of code! Okay, so to get this to actually work, in real-time at scale, would take quite a bit of work. But, at the core of the best contextual advertising solutions in the world, there's a deep learning algorithm labeling articles. Let's build our own here with this project.

–

Description:

This dataset contains around 200k news headlines from the year 2012 to 2018 obtained from HuffPost. The model trained on this dataset could be used to identify tags for untracked news articles or to identify the type of language used in different news articles.

Each news headline has a corresponding category. Categories and corresponding article counts are as follows:

- **POLITICS: 32739**
- **WELLNESS: 17827**
- **ENTERTAINMENT: 16058**
- **TRAVEL: 9887**
- **STYLE & BEAUTY: 9649**
- **PARENTING: 8677**
- **HEALTHY LIVING: 6694**
- **QUEER VOICES: 6314**
- **FOOD & DRINK: 6226**
- **BUSINESS: 5937**
- **COMEDY: 5175**
- **SPORTS: 4884**
- **BLACK VOICES: 4528**
- **HOME & LIVING: 4195**
- **PARENTS: 3955**
- **THE WORLDPOST: 3664**
- **WEDDINGS: 3651**
- **WOMEN: 3490**
- **IMPACT: 3459**
- **DIVORCE: 3426**
- **CRIME: 3405**
- **MEDIA: 2815**
- **WEIRD NEWS: 2670**
- **GREEN: 2622**
- **WORLDPOST: 2579**
- **RELIGION: 2556**
- **STYLE: 2254**
- **SCIENCE: 2178**
- **WORLD NEWS: 2177**
- **TASTE: 2096**
- **TECH: 2082**

- **MONEY: 1707**
- **ARTS: 1509**
- **FIFTY: 1401**
- **GOOD NEWS: 1398**
- **ARTS & CULTURE: 1339**
- **ENVIRONMENT: 1323**
- **COLLEGE: 1144**
- **LATINO VOICES: 1129**
- **CULTURE & ARTS: 1030**
- **EDUCATION: 1004**

–

Dr. Vargo's Benchmarks

When I run my deep learning workflow, I get the following evaluation metrics:

	precision	recall	f1-score	support
0	0.88	0.84	0.86	669
1	0.85	0.89	0.87	670
accuracy			0.86	1339
macro avg	0.86	0.86	0.86	1339
weighted avg	0.86	0.86	0.86	1339

–

5-Minute Presentation

You must record a 5-minute video presentation and upload it to Coursera. Spend a lot of time making your presentation something that you would feel comfortable sharing with a potential employer. **If your presentation is longer than 5-minutes, your peer grader will stop watching at the 5-minute mark, and whatever you did not cover you will not receive credit for in the rubric. Do not go over 5 minutes. If you're long, condense and film again.**