SEAS 6401 Final Project

Predicting Excitement for DonorsChoose.org

Introduction

DONORS CHOOSE

Find a classroom to support | About us | Help

Sign in

At school or at home... wherever learning happens, you can make a difference.

With teachers and students facing the toughest learning challenges in a generation, you can help provide the supplies to keep kids learning.

See teachers' projects

⚠ Learn how we're working to support teachers and students during Covid-19.

Search topics, teachers & schools | *near* | city, state, or zip | **Search**

Can we identify exciting projects solely from teacher essays using machine learning?

What is “exciting”?

An exciting project…

- is a fully funded project on DonorsChoose.org
- had at least one teacher-acquired donor
- has a greater-than-average comment percentage among donors
- has at least one “green” donation
- has one or more of:
  - donations from three or more non teacher-acquired donors (three_or_more_non_teacher_referred_donors)
  - one non teacher-acquired donor gave more than $100 (one_non_teacher_referred_donor_giving_100_plus)
  - the project received a donation from a "thoughtful donor"

Source: https://www.kaggle.com/c/kdd-cup-2014-predicting-excitement-at-donors-choose/data

Data

essays.csv

- Essays submitted by the teachers for their projects - 664098 rows

	projectid	essay
0	ffffc4f85b60efc5b52347df489d0238	I am a fourth year fifth grade math teacher. T...
1	ffffac55ee02a49d1abc87ba6fc61135	Can you imagine having to translate everything...
2	ffff97ed93720407d70a2787475932b0	Hi. I teach a wonderful group of 4-5 year old ...
3	ffff7266778f71242675416e600b94e1	My Kindergarten students come from a variety o...
4	ffff418bb42fad24347527ad96100f81	All work and no play makes school a dull place...

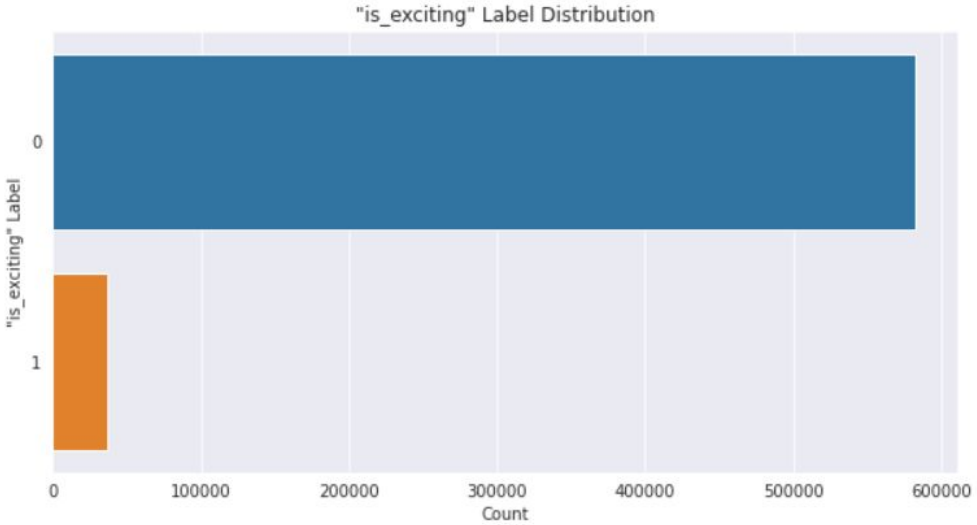
outcomes.csv

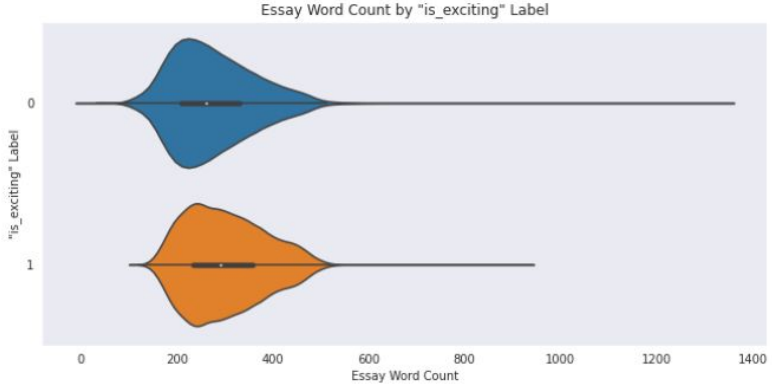
- The label (“is_exciting”) and other project attributes - 619326 rows

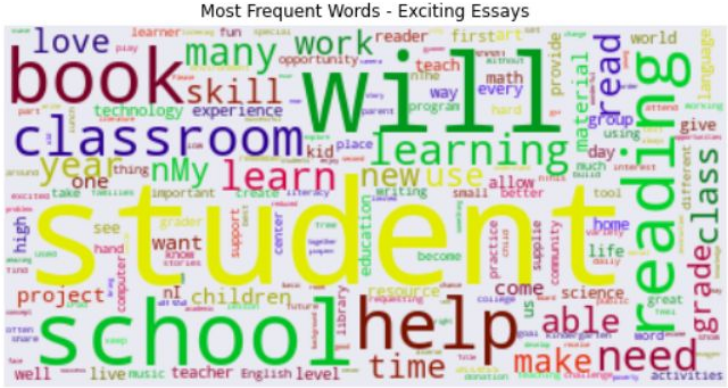
	projectid	is_exciting
0	ffffc4f85b60efc5b52347df489d0238	0
1	ffffac55ee02a49d1abc87ba6fc61135	0
2	ffff97ed93720407d70a2787475932b0	0
3	ffff418bb42fad24347527ad96100f81	0
4	ffff2d9c769c8fb5335e949c615425eb	1

Methodology

Exploratory Data Analysis



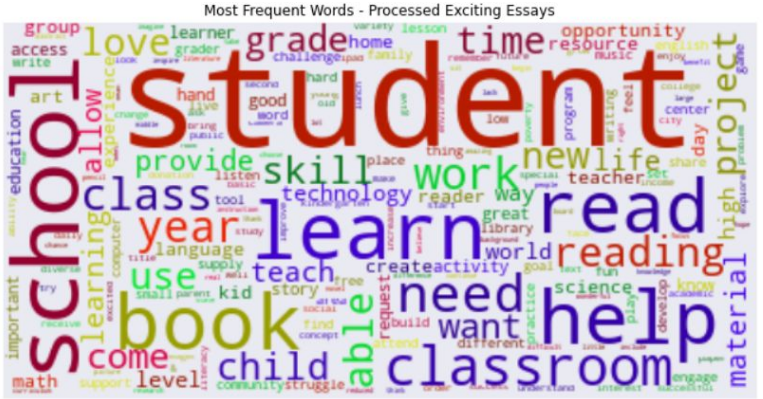




Feature Engineering/Preprocessing

- Main library used for preprocessing: spaCy
- Text processing steps taken:
  - Replace "\r\\n" in text - only step used for transfer learning modeling.
  - Lowercase the text.
  - Remove extra spaces.
  - Tokenize text (spaCy).
  - Remove punctuation.
  - Remove stop words (spaCy).
  - Lemmatize text (spaCy).
  - Remove leftover punctuation.

spaCy



Modeling: Bag-of-Words

Little Bo Peep has lost her sheep, And can't tell where to find them;
Leave them alone, and they'll come home,
Bringing their tails behind them.

The quick, brown fox jumped over the lazy sheep dog.

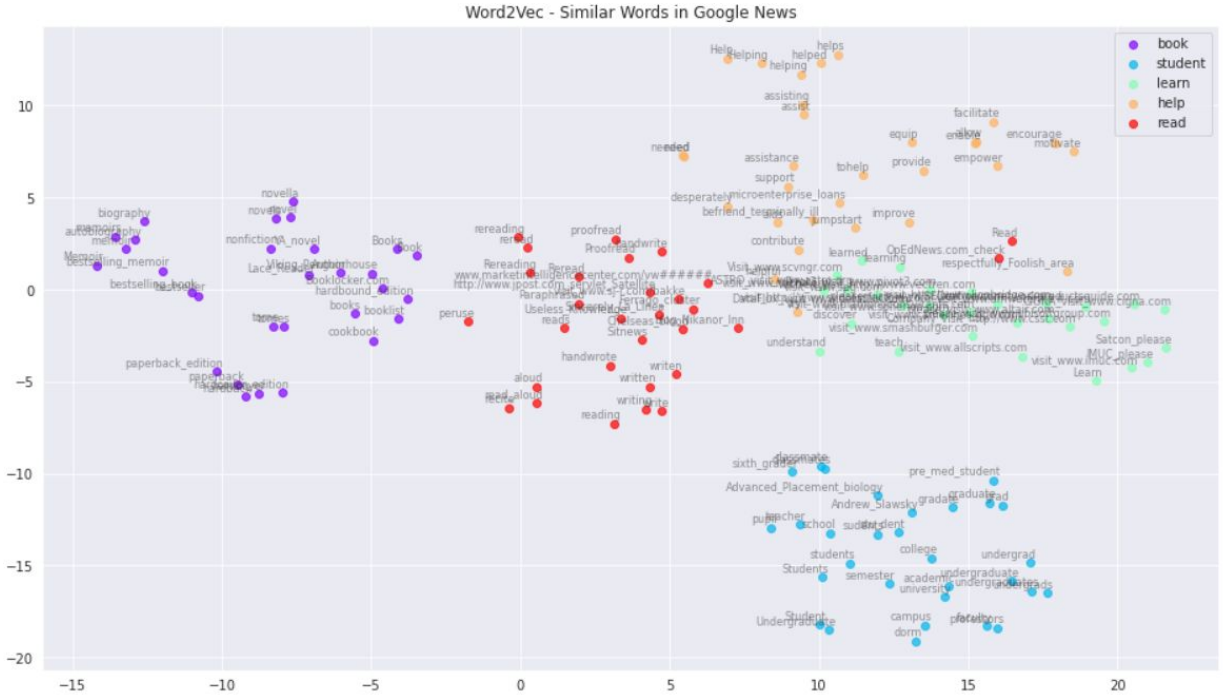
little	sheep	fox	lazy	dog	quick	...
1	1	0	0	0	0	...
0	1	1	1	1	1	...

Modeling: Word Embeddings

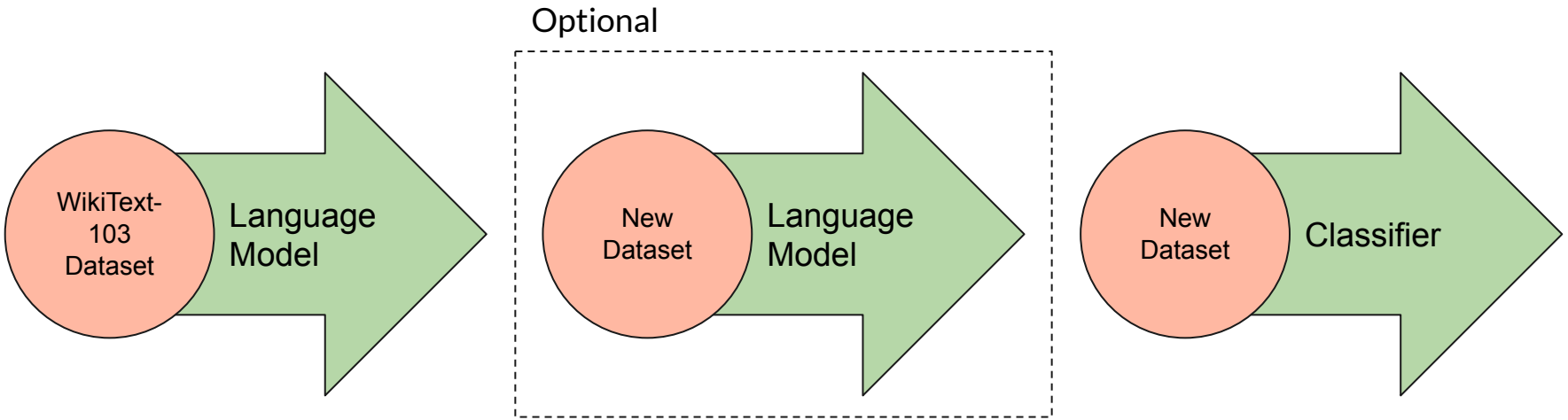
The quick, brown fox jumped over the lazy sheep dog.

the	[10.2, 4.2, 0.9, …]
quick	[-0.9, 5.3, 4.0, …]
brown	[9.1, 23.0, 36.1, …]
…	…

6.1	10.8	13.7	…



Modeling: Transfer Learning



Results & Conclusion

Results

	Description	Accuracy (%)	F1 (%)	Avg. Runtime (min)
Baseline	LR + TF	90	54	9
Bag-of-Words	LR + TF-IDF	86	55	9
	RF + TF-IDF	82	53	7
	Up + LR + TF-IDF	92	53	10
Embeddings	LR	59	44	1
	RF	91	52	7
Transfer Learning	AWD-LSTM + WCE Loss	70	50	36*

Acronyms: LR = Logistic Regression, TF = Term Frequency, RF = Random Forest, Up = Upsampling, AWD-LSTM = ASGD Weight-Dropped LSTM, WCE = Weighted Cross-Entropy

*This is an average **per epoch**. The model was trained for 5 epochs total.

Conclusion

- Best approach: **bag-of-words with logistic regression and TF-IDF vectorization**
- Not enough to use the essays to determine whether it is an “exciting” project or not
- Improvements/Future Work:
    - Change upsampling strategy from random oversampling (SMOTE, ADASYN, Snorkel data augmentation)
    - Try additional models besides logistic regression and random forest
    - Train the transfer learning model for more epochs
    - Experiment with more transfer learning parameters (learning rate, batch size, etc.)
    - Train a language model on the essays (include the middle step from the transfer learning process)

Questions?

Thank you!