Kiwi: Python Tool for Text Processing and Classification

Neelima Pulagam, Sai Marasani, Brian Sass
Ford Motor Company

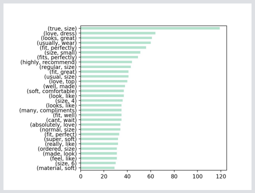
Introduction

Text analysis is a ubiquitous and important task as most unstructured data is textual. Extracting and classifying text can be extremely useful and help automate manual, time consuming tasks. On the other hand, machine learning can be a difficult field to breach. We propose a user-friendly text classification tool that will organize and categorize unstructured text data. This tool will allow users within the field to avoid creating boilerplate code for basic natural language processing (NLP) tasks and users new to machine learning and NLP plug and play with various models and methods. Our main goal is to make natural language processing accessible and easy.

Uploading Files and Data Visualization

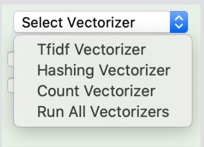
The user begins by choosing a CSV file of text data that has already been cleaned. This includes removing null values and special characters. The chosen file will be displayed in the data upload frame.

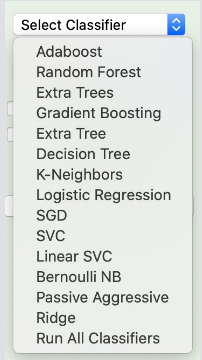
Once the user has selected the CSV file with text data, the corresponding input text column and output label column, they can visualize the data. They can get a visual of the class distribution as a pie chart, word cloud and histograms for n-grams.



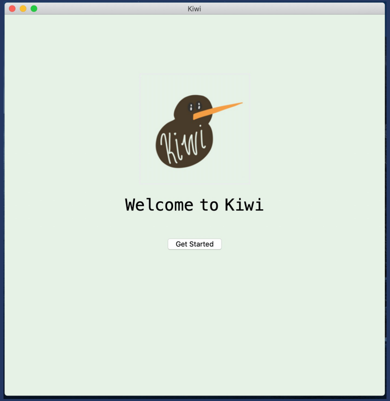
Machine Learning Models

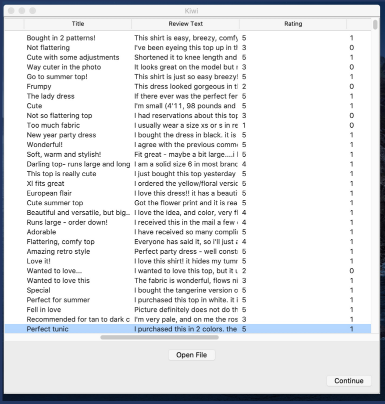
Once the input text and output columns have been chosen, the user can then run and test a few different types of models. To process the data, the user can choose between three different vectorizers.



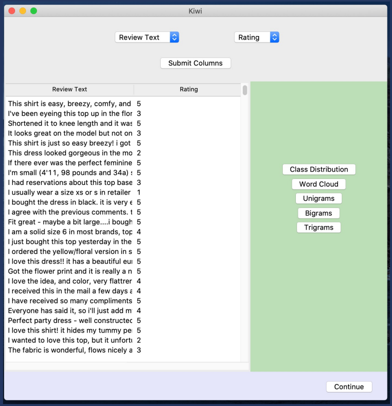


An NLP tool for beginners and experts alike…

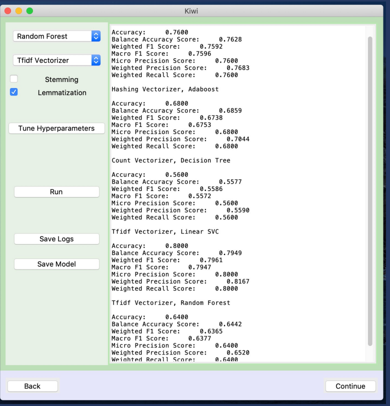




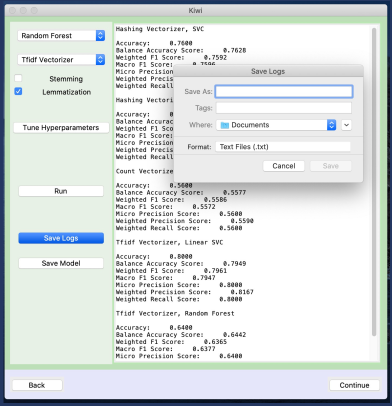
	Title	Review Text	Rating	
	Bought in 2 patterns!	This shirt is easy, breezy, comfy	5	1
	Not flattering	I've been eyeing this top up in tl	3	0
	Cute with some adjustments	Shortened it to knee length and	5	1
	Way cuter in the photo	It looks great on the model but r	3	0
	Go to summer top!	This shirt is just so easy breezy!	5	1
	Frumpy	This dress looked gorgeous in tl	2	0
	The lady dress	If there ever was the perfect fer	5	1
	Cute	I'm small (4'11, 98 pounds and	5	1
	Not so flattering top	I had reservations about this top	3	0
	Too much fabric	I usually wear a size xs or s in re	1	0
	New year party dress	I bought the dress in black. it is	5	1
	Wonderful!	I agree with the previous comm	5	1
	Soft, warm and stylish!	Fit great - maybe a bit large....i l	5	1
	Darling top- runs large and long	I am a solid size 6 in most branc	4	1
	This top is really cute	I just bought this top yesterday	5	1
	Xl fits great	I ordered the yellow/floral versic	5	1
	European flair	I love this dress!! it has a beauti	5	1
	Cute summer top	Got the flower print and it is rea	5	1
	Beautiful and versatile, but big..	I love the idea, and color, very fl	4	1
	Runs large - order down!	I received this in the mail a few (	4	1
	Adorable	I have received so many complii	5	1
	Flattering, comfy top	Everyone has said it, so i'll just a	4	1
	Amazing retro style	Perfect party dress - well consti	5	1
	Love it!	I love this shirt! it hides my tumr	5	1
	Wanted to love...	I wanted to love this top, but it u	2	0
	Wanted to love this	The fabric is wonderful, flows ni	3	1
	Special	I bought the tangerine version o	5	1
	Perfect for summer	I purchased this top in white. it i	5	1
	Fell in love	Picture definitely does not do th	5	1
	Recommended for tan to dark c	I'm very pale, and on me the ros	3	1
	Perfect tunic	I purchased this in 2 colors. the	5	1



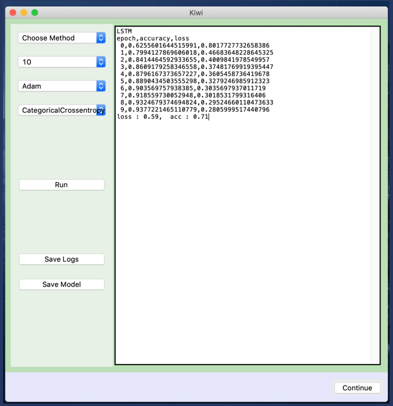
Review Text	Rating
This shirt is easy, breezy, comfy, and	5
I've been eyeing this top up in the flor	3
Shortened it to knee length and it was	5
It looks great on the model but not on	3
This shirt is just so easy breezy! i got	5
This dress looked gorgeous in the mo	2
If there ever was the perfect feminine	5
I'm small (4'11, 98 pounds and 34a) s	5
I had reservations about this top base	3
I usually wear a size xs or s in retailer	1
I bought the dress in black. it is very e	5
I agree with the previous comments. t	5
Fit great - maybe a bit large....i bough	5
I am a solid size 6 in most brands, top	4
I just bought this top yesterday in the	5
I ordered the yellow/floral version in s	5
I love this dress!! it has a beautiful eu	5
Got the flower print and it is really a n	5
I love the idea, and color, very flattrer	4
I received this in the mail a few days a	4
I have received so many compliments	5
Everyone has said it, so i'll just add m	4
Perfect party dress - well constructed	5
I love this shirt! it hides my tummy pe	5
I wanted to love this top, but it unfortu	2
The fabric is wonderful, flows nicely a	3



```
Accuracy:     0.7600
Balance Accuracy Score:     0.7628
Weighted F1 Score:     0.7592
Macro F1 Score:     0.7596
Micro Precision Score:     0.7600
Weighted Precision Score:     0.7683
Weighted Recall Score:     0.7600

Hashing Vectorizer, Adaboost

Accuracy:     0.6800
Balance Accuracy Score:     0.6859
Weighted F1 Score:     0.6738
Macro F1 Score:     0.6753
Micro Precision Score:     0.6800
Weighted Precision Score:     0.7044
Weighted Recall Score:     0.6800

Count Vectorizer, Decision Tree

Accuracy:     0.5600
Balance Accuracy Score:     0.5577
Weighted F1 Score:     0.5586
Macro F1 Score:     0.5572
Micro Precision Score:     0.5600
Weighted Precision Score:     0.5590
Weighted Recall Score:     0.5600

Tfidf Vectorizer, Linear SVC

Accuracy:     0.8000
Balance Accuracy Score:     0.7949
Weighted F1 Score:     0.7961
Macro F1 Score:     0.7947
Micro Precision Score:     0.8000
Weighted Precision Score:     0.8167
Weighted Recall Score:     0.8000

Tfidf Vectorizer, Random Forest

Accuracy:     0.6400
Balance Accuracy Score:     0.6442
Weighted F1 Score:     0.6365
Macro F1 Score:     0.6377
Micro Precision Score:     0.6400
Weighted Precision Score:     0.6520
Weighted Recall Score:     0.6400
```



```
Hashing Vectorizer, SVC

Accuracy:     0.7600
Balance Accuracy Score:     0.7628
Weighted F1 Score:     0.7592
Macro F1 Score:     0.7596
Micro Precision
Weighted Precis
Weighted Recall

Hashing Vectori

Accuracy:     0
Balance Accurac
Weighted F1 Sco
Macro F1 Score:
Micro Precision
Weighted Precis
Weighted Recall

Count Vectorize

Accuracy:     0.5600
Balance Accuracy Score:     0.5577
Weighted F1 Score:     0.5586
Macro F1 Score:     0.5572
Micro Precision Score:     0.5600
Weighted Precision Score:     0.5590
Weighted Recall Score:     0.5600

Tfidf Vectorizer, Linear SVC

Accuracy:     0.8000
Balance Accuracy Score:     0.7949
Weighted F1 Score:     0.7961
Macro F1 Score:     0.7947
Micro Precision Score:     0.8000
Weighted Precision Score:     0.8167
Weighted Recall Score:     0.8000

Tfidf Vectorizer, Random Forest

Accuracy:     0.6400
Balance Accuracy Score:     0.6442
Weighted F1 Score:     0.6365
Macro F1 Score:     0.6377
Micro Precision Score:     0.6400
```



```
LSTM
epoch,accuracy,loss
 0,0.6255601644515991,0.8017727732658386
 1,0.7994127869606018,0.46683648228645325
 2,0.8414464592933655,0.4009841978549957
 3,0.8609179258346558,0.37481769919395447
 4,0.8796167373657227,0.3605458736419678
 5,0.8890434503555298,0.3279246985912323
 6,0.903569757938385,0.3035697937011719
 7,0.918559730052948,0.3018531799316406
 8,0.9324679374694824,0.29524660110473633
 9,0.9377221465110779,0.2805999517440796
loss : 0.59,  acc : 0.71
```

Machine Learning (Cont.)

Additionally, there are two options for users to preprocess their text data by utilizing stemming and lemmatization. Stemming, the process of reducing a word to its stem (i.e. cats -> cat or going -> go) and lemmatization, a technique that utilizes part of speech tagging to reduce a word to its root (i.e. was -> be or has -> have) sometimes offer better performance.

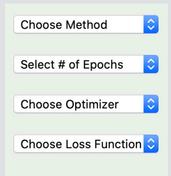
Hyper-parameter Tuning

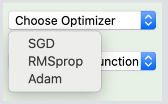
Hyper-parameter tuning is another useful technique that searches for through different combinations of parameters to find the ideal model architecture with the best performance. Our tool currently allows users to simply run randomized search on a model of their choosing. The parameters were manually hardcoded into a dictionary and the parameters for the selected model are retrieved when the the 'Tune Hyperparameters' button is pressed.

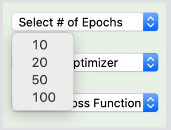
Deep Learning Methods

Currently there are two architectures implemented: LSTMs and RNNs. These models were chosen as they had the shortest training time among the several available architectures. Also, these architectures do not need GPU capabilities to run efficiently compared to other models.

The users can currently choose from a set of options the number of epochs, optimizer and loss function for each model. The users will have the accuracy and loss function for each epoch.







Future Work

We see a lot of potential and growth for this tool. This includes giving the user more granular control of choosing the model architecture. This includes which hyperparameters they wish to change or building models that utilize several different types of input. Additionally, there are a lot of deep learning architectures that can be added. Next steps for the tool also include more text processing methods and text embeddings. We also plan on integrating existing frameworks such as ML Flow, Optuna, etc. to provide a more user friendly experience. This project holds much potential that we hope to exploit in the near future.

References

Agarap, Abien Fred. (2018). Statistical Analysis on E-Commerce Reviews, with Sentiment Classification using Bidirectional Recurrent Neural Network.

Women's E-Commerce Clothing Reviews, 2018, June 30, 2021. https://www.kaggle.com/datasets/nicapotato/womens-ecommerce-clothing-reviews