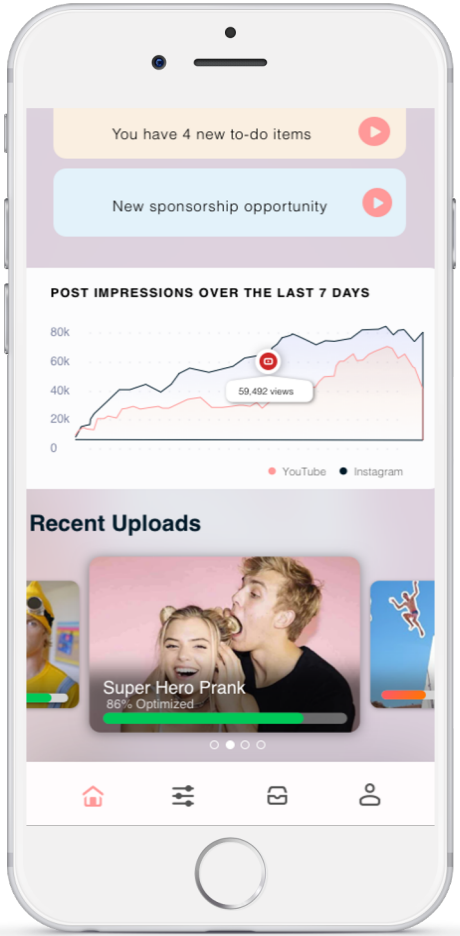
Predicting Click-Through-Rates For YouTube Videos Based on Video Title

George Paskalev

CAPSTONE PROJECT

FLATIRON SCHOOL



THE CLIENT.

TACTIQ

TACTIQ is an app that makes the increasingly complex world of social influence more manageable. Through machine-learning and personal guidance, **TACTIQ** takes away the stress of optimization, organization, and monetization across all platforms so creators can simply create.

DATASET

Provided by TACTIQ's parent agency, the dataset includes information on 15,000 videos by 50 creators.

FOCUS

TACTIQ works primarily with YouTube creators.

MACHINE LEARNING

AI-generated 'blueprint' to map out the path to success.

THE GOAL

Train models that guide YouTubers to optimize their videos based on predictions and recommendations.

THE BUSINESS PROBLEM

THE TACTIQ MVP NEEDS A FEATURE TO PREDICT A VIDEO'S CTR BASED ON ITS TITLE*

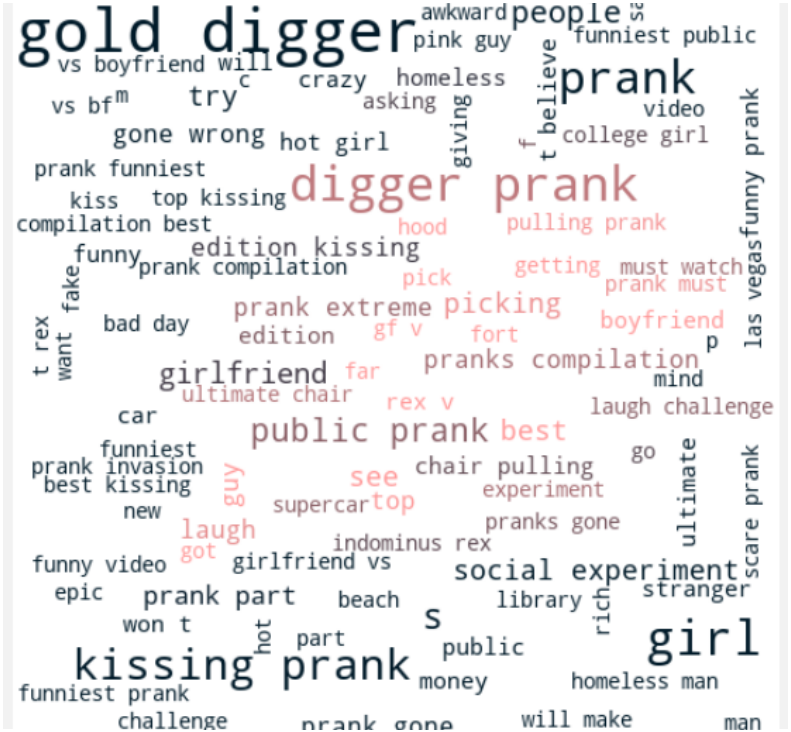
* A video's title and thumbnail both play a crucial role in achieving a successful Click-Through-Rate. The title alone is not enough to predict CTR with 100% accuracy.

LET'S GET STARTED.

POPULAR WORDS IN PRANK TITLES.

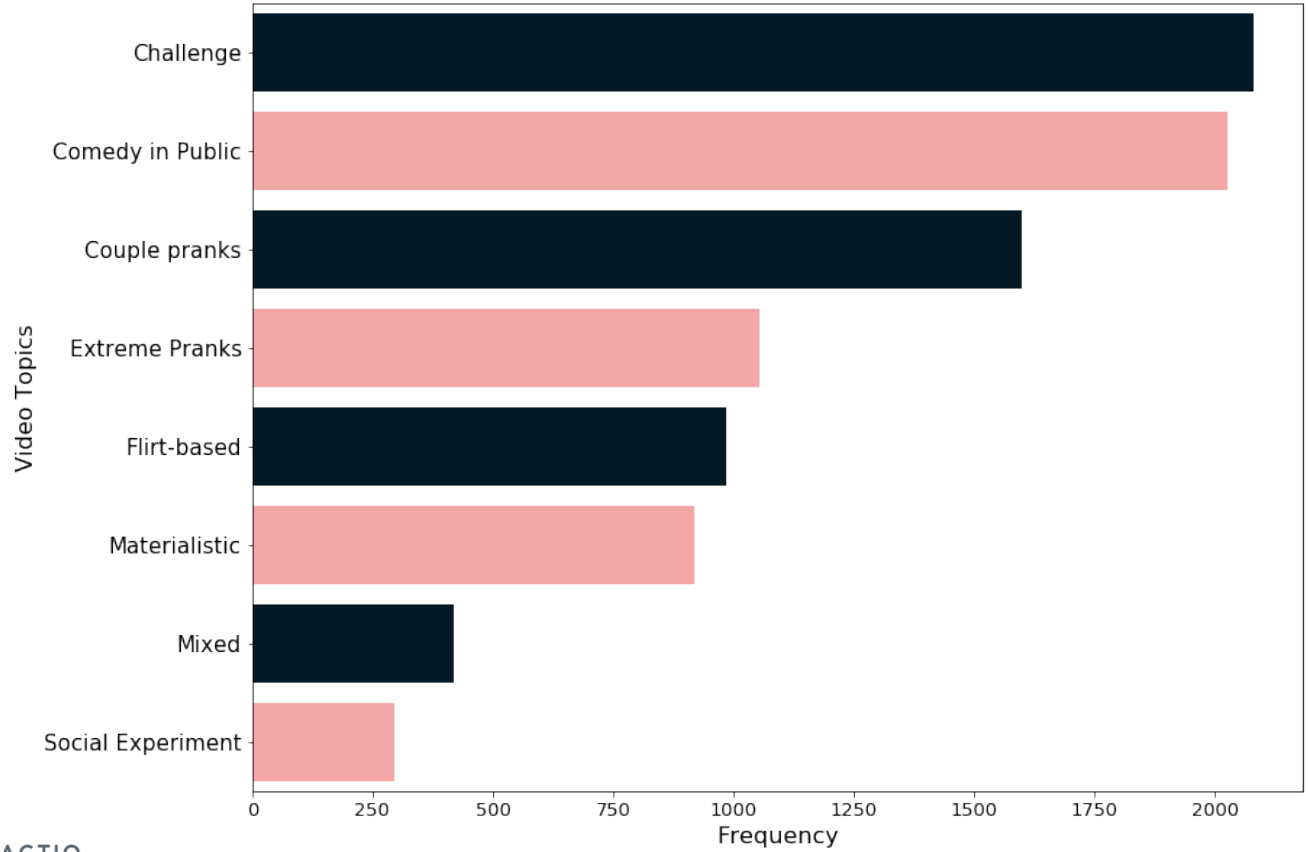
PRE-PROCESSING. The title of over 15k videos were transformed using standard NLP techniques (removing stop words, tokenization, etc.)

CLASSIFICATION. Each video was initially assigned to have either a Low, Average, or High CTR based on its CTR.



BEYOND INDIVIDUAL WORDS.

TOPIC DISTRIBUTION



TECHNIQUES for topic extraction that I used include: LDA and non-negative matrix factorization (NMF).

TOPIC DISTRIBUTION was pretty identical across all three classes.

THE SOLUTION

LET'S GET TO MODELING.

LET'S GET TO MODELING.

THE MODELS.

MY APPROACH.

CLASSIFICATION.

1 TRAIN 9 VANILLA CLASSIFIERS

2 PICK THE BEST-PERFORMING BASED ON ACCURACY.

3 TWEAK.

REGRESSION.

1 USE A RANDOM FORREST REGRESSOR WITH TF-IDF

2 TRY IT WITH WORD2VEC

3 TRY XGBOOST WITH WORD2VEC

VANILLA MODELS

CLASSIFICATION

From all vanilla classifiers, I chose to move forward with SVM, SGDC, and Multinomial Bayes; always using TF-IDF as my vectorizer.

	Model	Accuracy (Train)	Accuracy (Test)
0	SVM w/ CV	0.733726	0.62642
1	SVM w/ TF	0.782391	0.622869
2	Linear SVC w/ CV	0.797441	0.578835
3	Linear SVC w/ TF	0.781136	0.591619
4	SGDC w/ CV	0.771855	0.590909
5	SGDC w/ TF	0.738743	0.612926
6	Multinomial Bayes w/ CV	0.72156	0.585227
7	Multinomial Bayes w/ TF	0.683306	0.619318
8	Random Forrest w/ CV	0.955976	0.566761
9	Random Forrest w/ TF	0.955976	0.585938

OPTIMIZATION

TWEAKING THE MODELS.

- Utilized Grid Search to obtain best parameters for all models.
- Removed words that appear frequently among all classes.
- Rearranged the classes based on different threshold to achieve optimal test accuracy.
- Best results: 72% accuracy with cut-off being 2% CTR, and 70.1% with 6% CTR

SGDC

A linear classifier using the SGD method, trained for binary classification where 6% CTR is the cut-off for Low and High CTR.

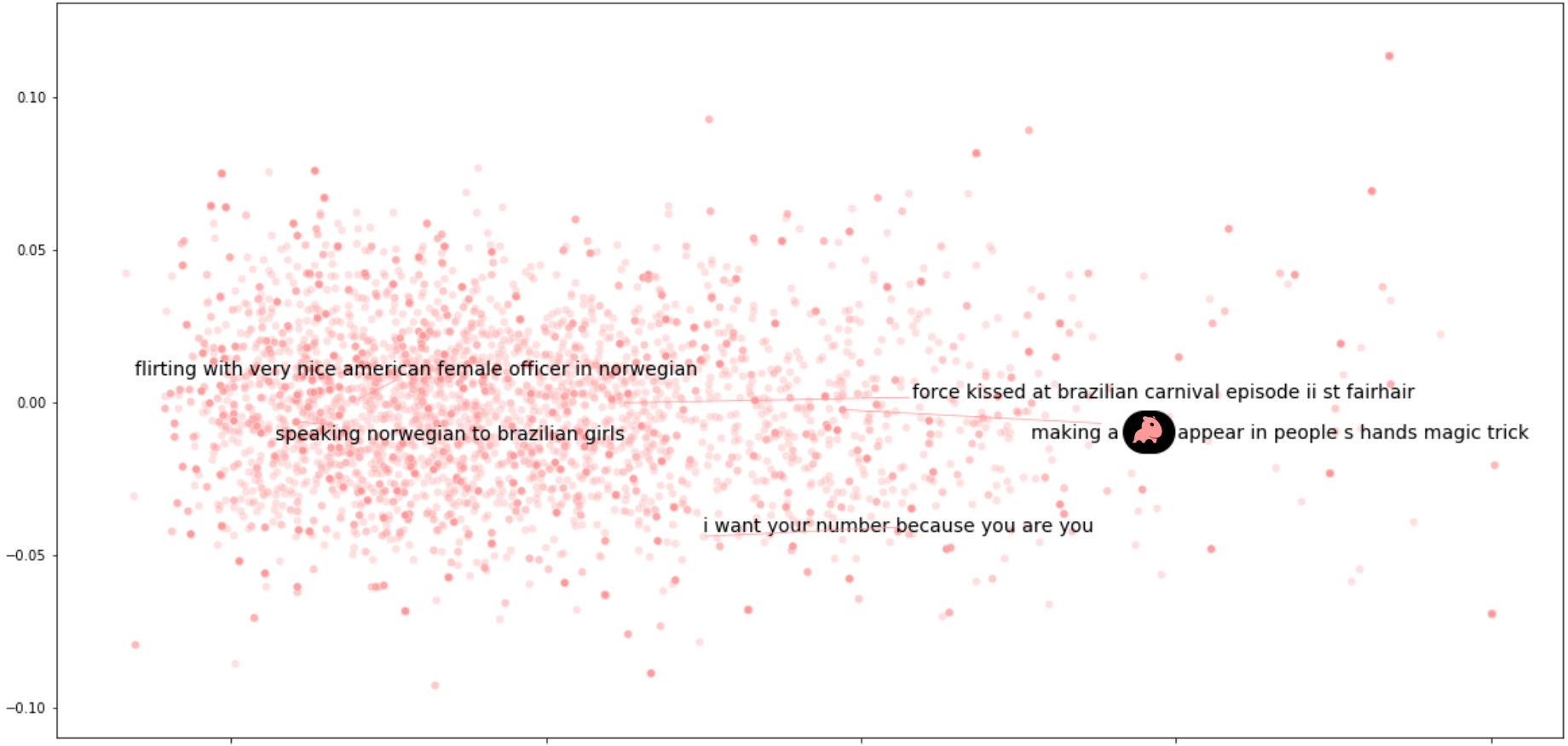
NEXT.

ON TO REGRESSION MODELS

ON TO REGRESSION MODELS

ANOTHER APPROACH.

WORD2VEC WITH GENSIM.



THE RESULTS.

REGRESSION PERFORMANCE

RANDOM FORREST
REGRESSOR WITH TF-IDF

RMSE: 3.50%

RANDOM FORREST
REGRESSOR WITH WORD2VEC

RMSE: 3.48%

XGB REGRESSOR:

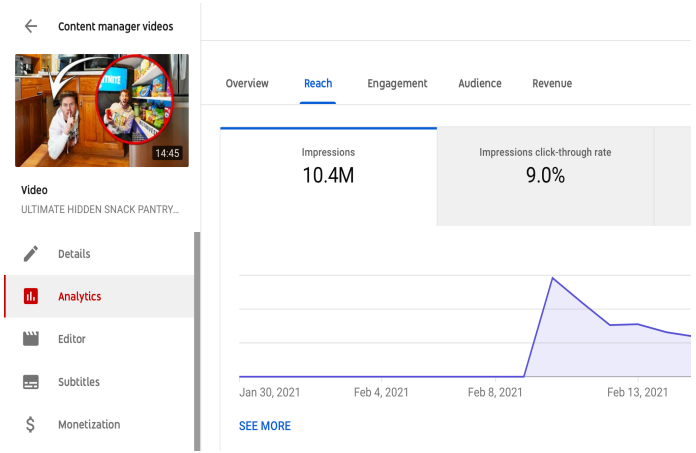
RMSE: 11.80%

THE WINNER:

MY LINEAR CLASSIFIER USING SGD.

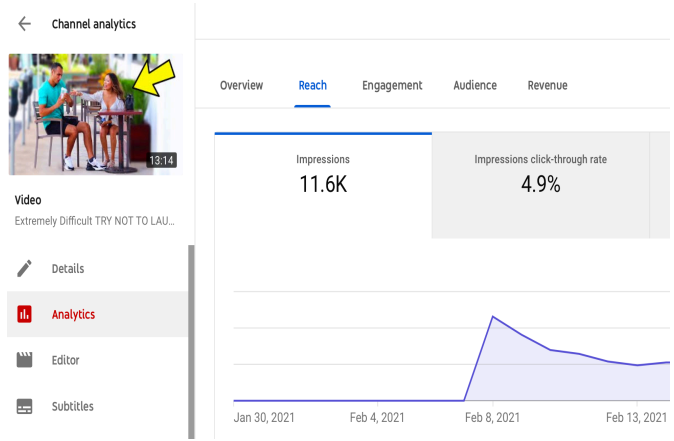
```
#Prediction #1
sgdc_final.predict([preprocess_title('ULTIMATE HIDDEN SNACK PANTRY! *Secret Entrance*')])
```

```
array(['High CTR'], dtype='<U8')
```




```
#PREDICTION #2
sgdc_final.predict([preprocess_title('Extremely Difficult TRY NOT TO LAUGH Challenge 😂🔥😹')])
```

```
array(['Low CTR'], dtype='<U8')
```



OPTIMIZATION

FUTURE WORK:

- Create a classifier that predicts CTR based on thumbnails.
- Train current model on CTR for different time frames: first 48 hours, first 7 days, etc.
- Train current model with titles in different languages.
- Train current model on titles from different video genres such as Beauty, Lifestyle, Gaming, etc.

$

RECOMMENDATION

The average CPM for a pranks channel is $1.69.

Over 80% of channels manage to achieve a CTR lower than 6%.

50% only get as high as 3.8% CTR. For a 100k views, that generates $6.42.

If these 50% use TACTIQ's tool, they can get above 6%, which would generate $10.42. Over 60% revenue increase.

THE END

THANK YOU.

THANK YOU.