DAAN 871: Data Visualization
Project Report

Anchal Gupta
Raj Vardhan

"Test Before You Tweet"

Purpose of visualization

Visualization is the use of the imagination through pictures or mental imagery to create visions of what we want in our lives and how to make them happen. It is a wonderful tool for preparing for anything, and everything. It invariably results in a higher level of performance. Patterns, trends and correlations that might go undetected in text-based data can be exposed and recognized easier with data visualization software. Tableau is one of the fastest growing data visualization tools that aims to help people see and understand data.

It's technology, however, that truly lit the fire under data visualization. Computers made it possible to process large amounts of data at lightning-fast speeds. Today, data visualization has become a rapidly evolving blend of science and art that is certain to change the corporate landscape over the next few years.

Because of the way the human brain processes information, using charts or graphs to visualize large amounts of complex data is easier than poring over spreadsheets or reports. Data visualization is a quick, easy way to convey concepts in a universal manner – and we can experiment with different scenarios by making slight adjustments. It can provide us the following:

Improved Insight

- Data visualization can provide insight that traditional descriptive statistics cannot.

Faster Decision Making

- Speed is key, and data visualization aides in the understanding of vast quantities of data by applying visual representations to the data.

Firstly, we used Tableau for exploring our data to get an insight of the trends in the field of Data Science and do some statistical analysis.

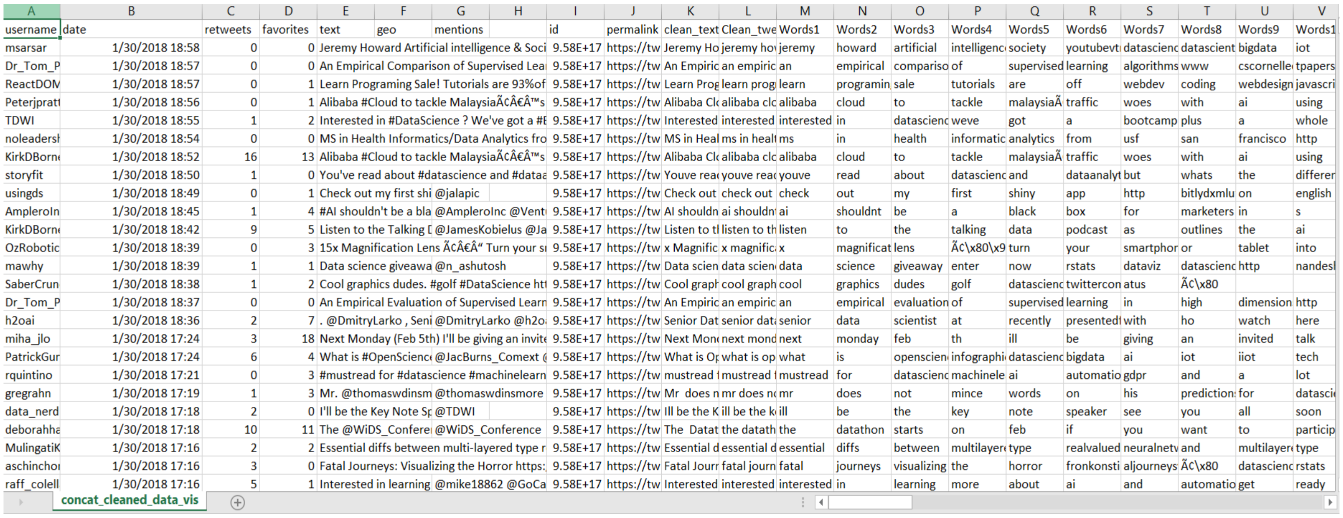
Our project is to provide users a platform where they can explore the past trends of Tweets in the field of "Data Science". Also, we have tried to extend the feature of Tableau reporting to a webpage where User can enter the tweet and get a prediction of Likes.

Data Collection

The connected society we live in today has allowed online users to willingly share opinions on an unprecedented scale. Motivated by the advent of mass opinion sharing, it is then crucial to devise algorithms that efficiently identify the emotions expressed within the opinionated content. Recently, Twitter has received a lot of interest and attention from a wide range of internet users across the globe. One of the main reasons for using Twitter is the ease of expressing opinions on diverse topics such as "Data Science". Such ease of use, coupled with the widespread use of connected portable devices, has made Twitter the primary channel for users to voluntarily share opinions, feelings, news, activities, interests, and other types of event-related information happening around them. Consequently, social networks have become some of the richest data repositories online.

We collected tweets on the topic "Data Science" for the year 2018(Jan to Dec). The file contains information about the creation date, number of retweets and likes, tweet text, mentions, hashtags etc. As the twitter data is raw data, we used Python to clean it. Below is the snapshot of data before and after cleaning:

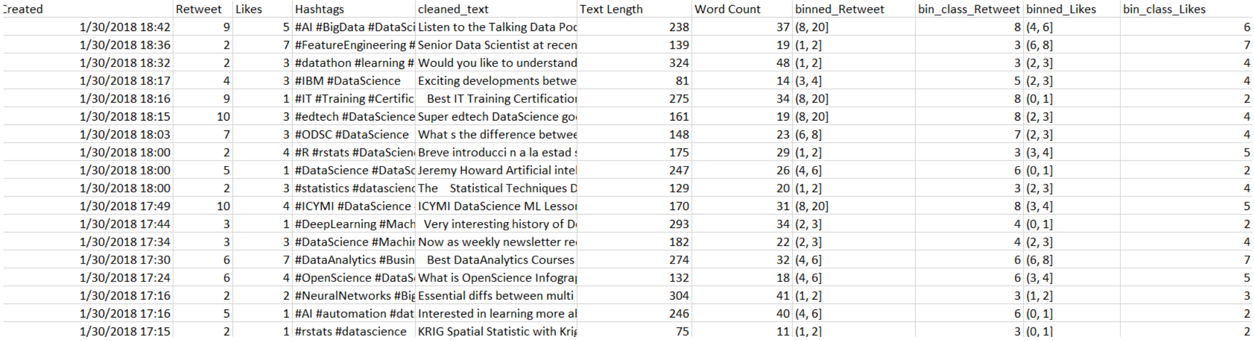
Raw Data:



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
username	date	retweets	favorites	text	geo	mentions		id	permalink	clean_text	Clean_twe	Words1	Words2	Words3	Words4	Words5	Words6	Words7	Words8	Words9	Words1
msarsar	1/30/2018 18:58	0	0	Jeremy Howard Artificial intelligence & Soci				9.58E+17	https://tw	Jeremy Ho	jeremy ho	jeremy	howard	artificial	intelligenc	society	youtubevt	datascienc	datascient	bigdata	iot
Dr_Tom_P	1/30/2018 18:57	0	0	An Empirical Comparison of Supervised Lea				9.58E+17	https://tw	An Empiric	an empiric	an	empirical	compariso	of	supervised	learning	algorithms	www	cscornelle	tpapers
ReactDOM	1/30/2018 18:57	0	1	Learn Programing Sale! Tutorials are 93%of				9.58E+17	https://tw	Learn Prog	learn prog	learn	programin	sale	tutorials	are	off	webdev	coding	webdesign	javascri
Peterjpratt	1/30/2018 18:56	0	1	Alibaba #Cloud to tackle MalaysiaÃ¢Â€Â™s				9.58E+17	https://tw	Alibaba Cl	alibaba clc	alibaba	cloud	to	tackle	malaysiaÃ	traffic	woes	with	ai	using
TDWI	1/30/2018 18:55	1	2	Interested in #DataScience ? We've got a #E				9.58E+17	https://tw	Interested	interested	interested	in	datascienc	weve	got	a	bootcamp	plus	a	whole
noleadersh	1/30/2018 18:54	0	0	MS in Health Informatics/Data Analytics fro				9.58E+17	https://tw	MS in Heal	ms in healt	ms	in	health	informatic	analytics	from	usf	san	francisco	http
KirkDBorne	1/30/2018 18:52	16	13	Alibaba #Cloud to tackle MalaysiaÃ¢Â€Â™s				9.58E+17	https://tw	Alibaba Cl	alibaba clc	alibaba	cloud	to	tackle	malaysiaÃ	traffic	woes	with	ai	using
storyfit	1/30/2018 18:50	1	0	You've read about #datascience and #dataa				9.58E+17	https://tw	Youve rea	youve rea	youve	read	about	datascienc	and	dataanalyt	but	whats	the	differen
usingds	1/30/2018 18:49	0	1	Check out my first shi		@jalapic		9.58E+17	https://tw	Check out	check out	check	out	my	first	shiny	app	http	bitlydxmlu	on	english
AmpleroIn	1/30/2018 18:45	1	4	#AI shouldn't be a bla		@AmpleroInc @Vent		9.58E+17	https://tw	AI shouldn	ai shouldn	ai	shouldnt	be	a	black	box	for	marketers	in	s
KirkDBorne	1/30/2018 18:42	9	5	Listen to the Talking D		@JamesKobielus @Ja		9.58E+17	https://tw	Listen to t	listen to th	listen	to	the	talking	data	podcast	as	outlines	the	ai
OzRobotic	1/30/2018 18:39	0	3	15x Magnification Lens Ã¢Â€Â" Turn your s				9.58E+17	https://tw	x Magnific	x magnific	x	magnificat	lens	Ã¢\x80\x9	turn	your	smartphor	or	tablet	into
mawhy	1/30/2018 18:39	1	1	Data science giveawa		@n_ashutosh		9.58E+17	https://tw	Data scien	data scien	data	science	giveaway	enter	now	rstats	dataviz	datascienc	http	nandesh
SaberCrun	1/30/2018 18:38	1	2	Cool graphics dudes. #golf #DataScience htt				9.58E+17	https://tw	Cool graph	cool graph	cool	graphics	dudes	golf	datascienc	twittercon	atus	Ã¢\x80		
Dr_Tom_P	1/30/2018 18:37	0	0	An Empirical Evaluation of Supervised Learn				9.58E+17	https://tw	An Empiric	an empiric	an	empirical	evaluation	of	supervised	learning	in	high	dimension	http
h2oai	1/30/2018 18:36	2	7	. @DmitryLarko , Seni		@DmitryLarko @h2o		9.58E+17	https://tw	Senior Dat	senior data	senior	data	scientist	at	recently	presentedt	with	ho	watch	here
miha_jlo	1/30/2018 17:24	3	18	Next Monday (Feb 5th) I'll be giving an invite				9.58E+17	https://tw	Next Mon	next mond	next	monday	feb	th	ill	be	giving	an	invited	talk
PatrickGur	1/30/2018 17:24	6	4	What is #OpenScienc		@JacBurns_Comext @		9.58E+17	https://tw	What is Op	what is op	what	is	openscien	infographi	datascienc	bigdata	ai	iot	iiot	tech
rquintino	1/30/2018 17:21	0	3	#mustread for #datascience #machinelearn				9.58E+17	https://tw	mustread f	mustread f	mustread	for	datascienc	machinele	ai	automatio	gdpr	and	a	lot
gregrahn	1/30/2018 17:19	1	3	Mr. @thomaswdinsm		@thomaswdinsmore		9.58E+17	https://tw	Mr does n	mr does nc	mr	does	not	mince	words	on	his	prediction	for	datasci
data_nerd	1/30/2018 17:18	2	0	I'll be the Key Note Sp		@TDWI		9.58E+17	https://tw	Ill be the K	ill be the k	ill	be	the	key	note	speaker	see	you	all	soon
deborahha	1/30/2018 17:18	10	11	The @WiDS_Confere		@WiDS_Conference		9.58E+17	https://tw	The Datat	the datath	the	datathon	starts	on	feb	if	you	want	to	particip
MulingatiK	1/30/2018 17:16	2	2	Essential diffs between multi-layered type r				9.58E+17	https://tw	Essential d	essential d	essential	diffs	between	multilayer	type	realvalued	neuralnetv	and	multilayer	type
aschinchor	1/30/2018 17:16	3	0	Fatal Journeys: Visualizing the Horror https:				9.58E+17	https://tw	Fatal Jourr	fatal journ	fatal	journeys	visualizing	the	horror	fronkonsti	aljourneys	Ã¢\x80	datascienc	rstats
raff_colell	1/30/2018 17:16	5	1	Interested in learning		@mike18862 @GoCa		9.58E+17	https://tw	Interested	interested	interested	in	learning	more	about	ai	and	automatio	get	ready

concat_cleaned_data_vis

Cleaned Data:



Created	Retweet	Likes	Hashtags	cleaned_text	Text Length	Word Count	binned_Retweet	bin_class_Retweet	binned_Likes	bin_class_Likes
1/30/2018 18:42	9	5	#AI #BigData #DataSci	Listen to the Talking Data Poc	238	37	(8, 20]	8	(4, 6]	6
1/30/2018 18:36	2	7	#FeatureEngineering #	Senior Data Scientist at recen	139	19	(1, 2]	3	(6, 8]	7
1/30/2018 18:32	2	3	#datathon #learning #	Would you like to understand	324	48	(1, 2]	3	(2, 3]	4
1/30/2018 18:17	4	3	#IBM #DataScience	Exciting developments betwe	81	14	(3, 4]	5	(2, 3]	4
1/30/2018 18:16	9	1	#IT #Training #Certific	Best IT Training Certificatio	275	34	(8, 20]	8	(0, 1]	2
1/30/2018 18:15	10	3	#edtech #DataScience	Super edtech DataScience go	161	19	(8, 20]	8	(2, 3]	4
1/30/2018 18:03	7	3	#ODSC #DataScience	What s the difference betwee	148	23	(6, 8]	7	(2, 3]	4
1/30/2018 18:00	2	4	#R #rstats #DataScien	Breve introducci n a la estad s	175	29	(1, 2]	3	(3, 4]	5
1/30/2018 18:00	5	1	#DataScience #DataSc	Jeremy Howard Artificial intel	247	26	(4, 6]	6	(0, 1]	2
1/30/2018 18:00	2	3	#statistics #datascienc	The Statistical Techniques D	129	20	(1, 2]	3	(2, 3]	4
1/30/2018 17:49	10	4	#ICYMI #DataScience	ICYMI DataScience ML Lesso	170	31	(8, 20]	8	(3, 4]	5
1/30/2018 17:44	3	1	#DeepLearning #Mach	Very interesting history of D	293	34	(2, 3]	4	(0, 1]	2
1/30/2018 17:34	3	3	#DataScience #Machir	Now as weekly newsletter re	182	22	(2, 3]	4	(2, 3]	4
1/30/2018 17:30	6	7	#DataAnalytics #Busin	Best DataAnalytics Courses	274	32	(4, 6]	6	(6, 8]	7
1/30/2018 17:24	6	4	#OpenScience #DataS	What is OpenScience Infogra	132	18	(4, 6]	6	(3, 4]	5
1/30/2018 17:16	2	2	#NeuralNetworks #Big	Essential diffs between multi	304	41	(1, 2]	3	(1, 2]	3
1/30/2018 17:16	5	1	#AI #automation #dat	Interested in learning more al	246	40	(4, 6]	6	(0, 1]	2
1/30/2018 17:15	2	1	#rstats #datascience	KRIG Spatial Statistic with Kri	75	11	(1, 2]	3	(0, 1]	2

Created: Creation date of the Tweet

Retweet: Count of Retweets on the Tweet

Likes: Count of Likes on the Tweet

Text: Tweet sentence

User Interaction

With the visualizations in our story, a user will be able to get an idea of the following:

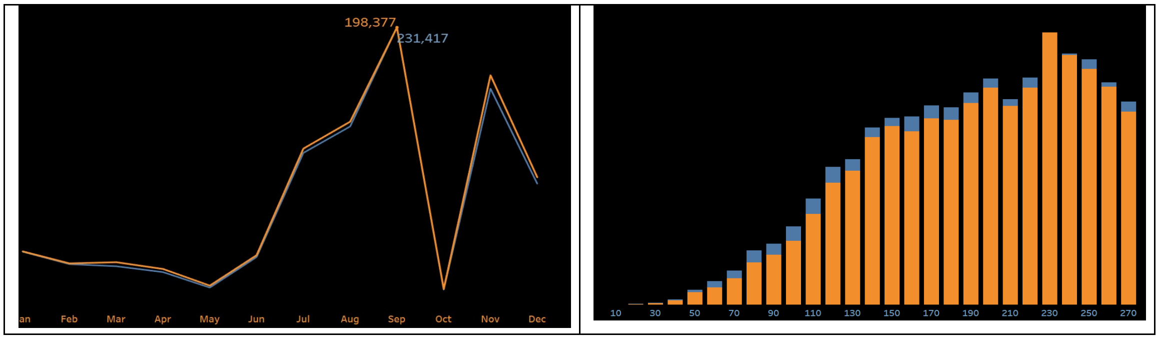
- A trend of Retweets and Likes on Tweets in the field of Data Science for the Year 2018(Jan to Dec).
- Dependency of Retweets and Likes on text length and the count of words in a Tweet.
- The most frequent words which are used by others.

- Also, a user can enter a tweet and get a prediction of Likes he/she can get.

Design Principles

Charts

- Use of line charts for time-series analysis.
- Use of stacked bar charts for frequency distribution.

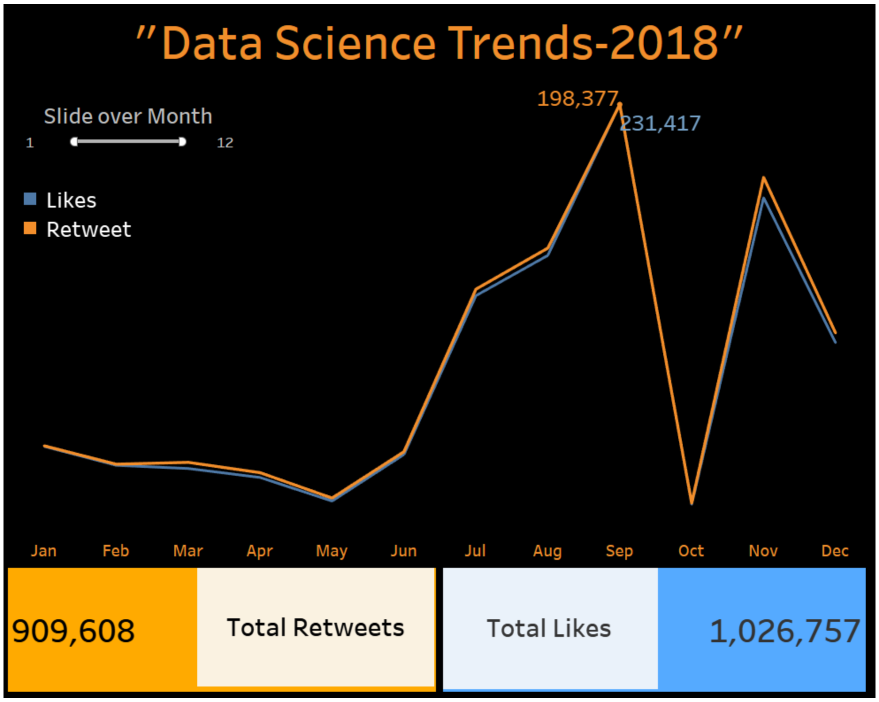


Color Maps

- Theme of yellow and blue has been kept uniform to understand visualizations of Retweets and Likes respectively.

Communication

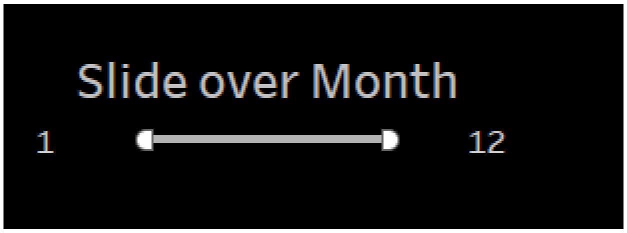
- Annotations, Legends have been used to provide the clear information to user.



Techniques

Filters

- To analyze data over months, a range filter has been used.

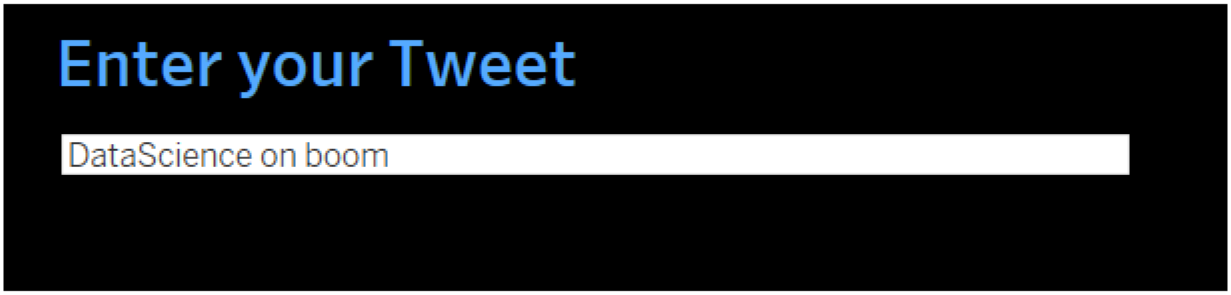


Slide over Month

1 12

Parameter

- We used text parameter where user can enter the tweet.

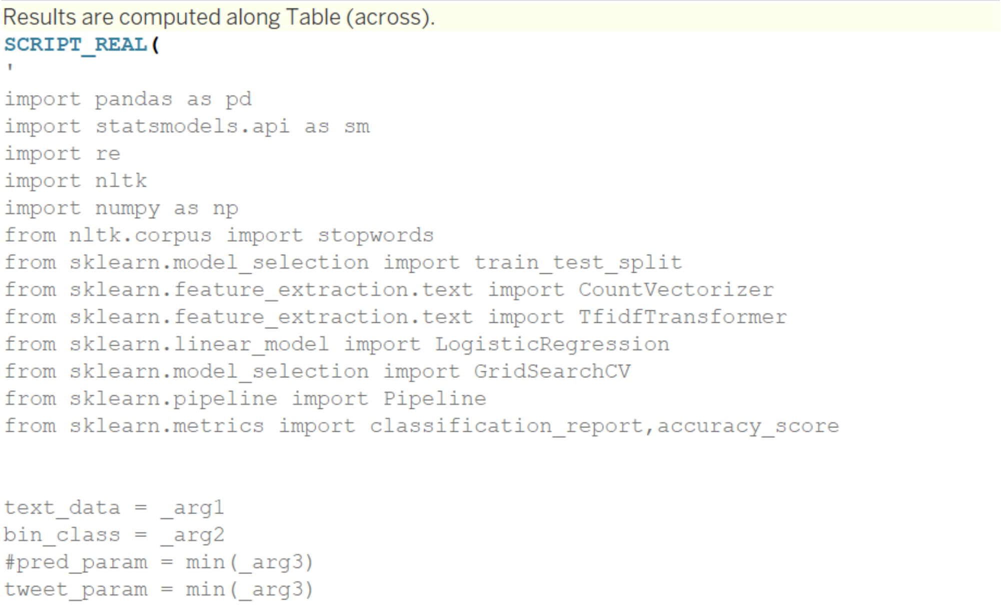


Enter your Tweet

DataScience on boom

Calculated Field

- We created a Calculated field for prediction of retweets on text entered by user, counting of words and computing text length.



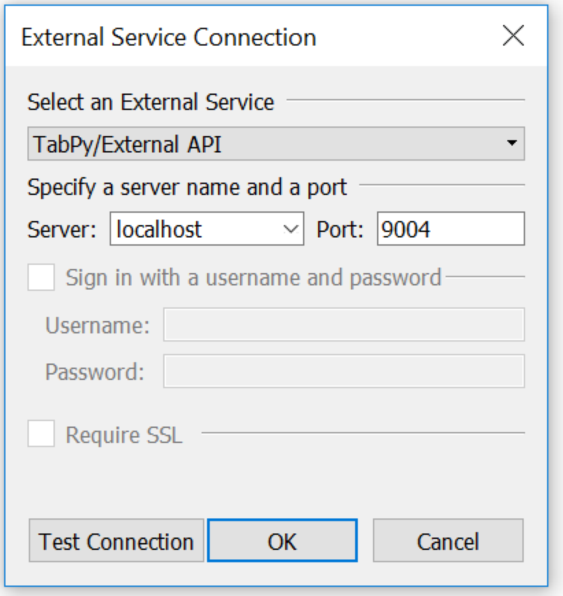
Results are computed along Table (across).

```
SCRIPT_REAL(
'
import pandas as pd
import statsmodels.api as sm
import re
import nltk
import numpy as np
from nltk.corpus import stopwords
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import GridSearchCV
from sklearn.pipeline import Pipeline
from sklearn.metrics import classification_report,accuracy_score


text_data = _arg1
bin_class = _arg2
#pred_param = min(_arg3)
tweet_param = min(_arg3)
```

Tabpy

- Connected Python to Tableau for applying Logistic regression to our data and providing predictions of retweets.



External Service Connection

Select an External Service

TabPy/External API

Specify a server name and a port

Server: localhost Port: 9004

□ Sign in with a username and password

Username:

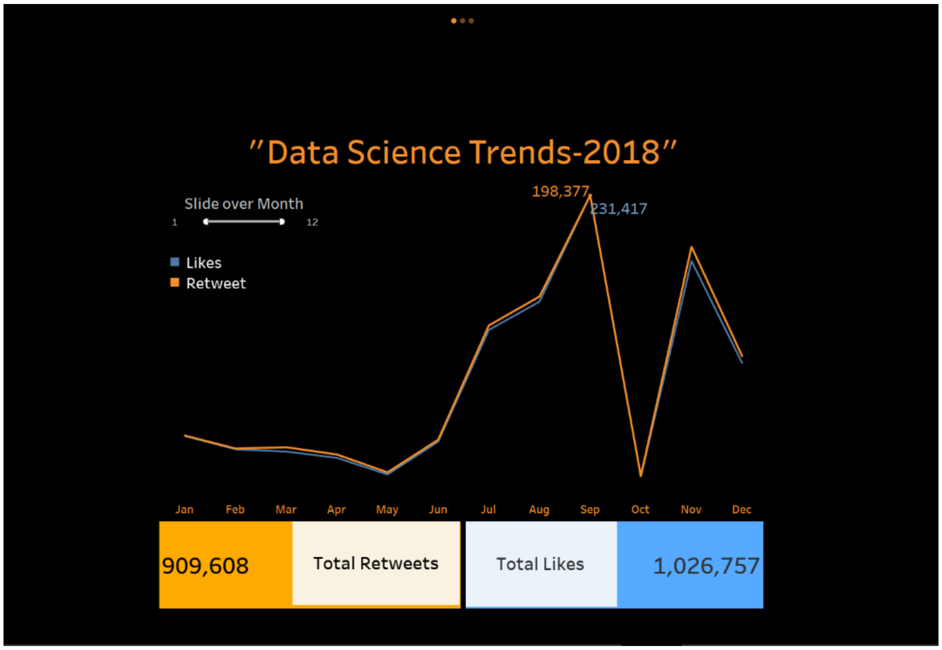
Password:

□ Require SSL

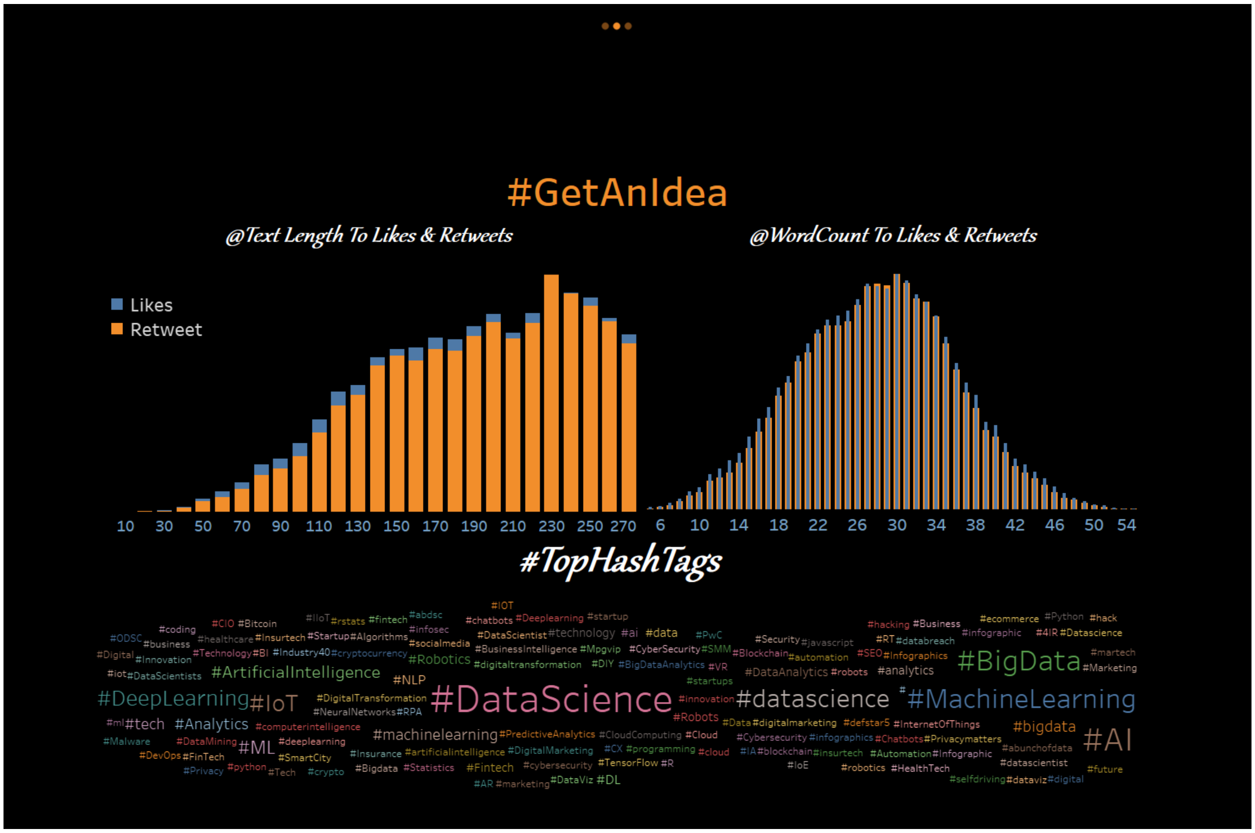
Dashboards/Story

- Multiple dashboards have been created to combine different sheets and create an interactive Story.

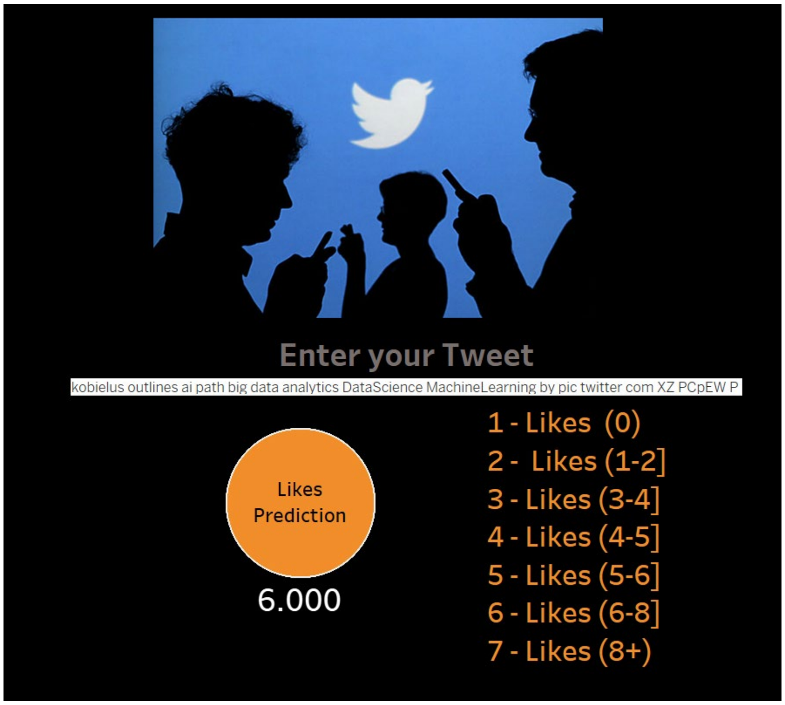
The story explains us the trends over the year 2018 for Data Science Tweets.



Next, we tried to determine the factors which can increase the likes and retweets.



The below dashboard allows a user to enter a Tweet and get a prediction of Likes.



Enter your Tweet

kobielus outlines ai path big data analytics DataScience MachineLearning by pic twitter com XZ PCpEW P

Likes Prediction

6.000

1 - Likes (0)
2 - Likes (1-2]
3 - Likes (3-4]
4 - Likes (4-5]
5 - Likes (5-6]
6 - Likes (6-8]
7 - Likes (8+)

We tested our model on 3 types of tweets and found the following results:

Tweet	**Predicted Likes**	**Factor Consideration**
What We Expect to See in perspectives of BigData #DataScience #AI #Python #RStats #TensorFlow #JavaScript #Analytics #architecture #DevOps #DataEngineering #ML #Java #ReactJS #VueJS #GoLang #CloudComputing #Serverless #infoq #com #articles #infoq #retrospective	6 (6-8]	More Words and large Text Length, Use of frequent words
AI is evolving	2 (1-2]	Small Sentence
AI is evolving. Learning is fun with #AI #BigData #MachineLearning #NLP	3 (2-3]	Use of frequent words

Future Work:

We are trying to apply models other than Logistic and trying to improve the accuracy and predictions. We will integrate the best model to this story which can give us faster and accurate results.