Whale & Jaguar

Data Analysis

News Categorization

Data source:
https://rishabhmisra.github.io/publications/

Data Scientist: Afonso Lima

Topics

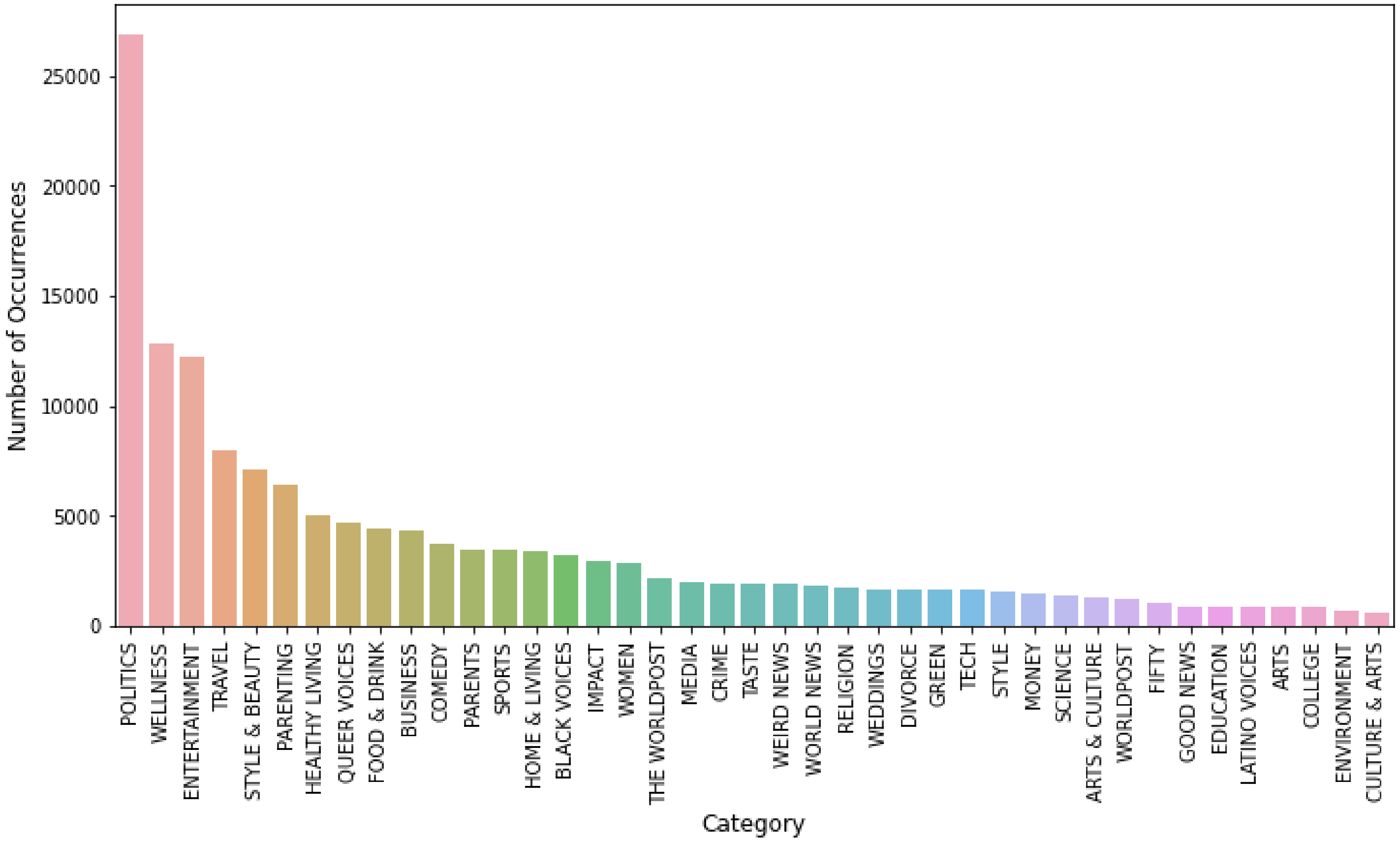
- Exploratory Data Analysis (EDA)
- Classification using supervised learning
- Classification using unsupervised learning
- Suggestions for improvements

Exploratory Data Analysis (EDA)

- The original Dataset has 200.852 rows
- After preprocessing phase, that turns the dataset more reliable, the treated dataset has 148.982 rows
- The new column considering the joining of "headline" and "short description" columns has 4.771.193 words.

Exploratory Data Analysis (EDA)

Categories



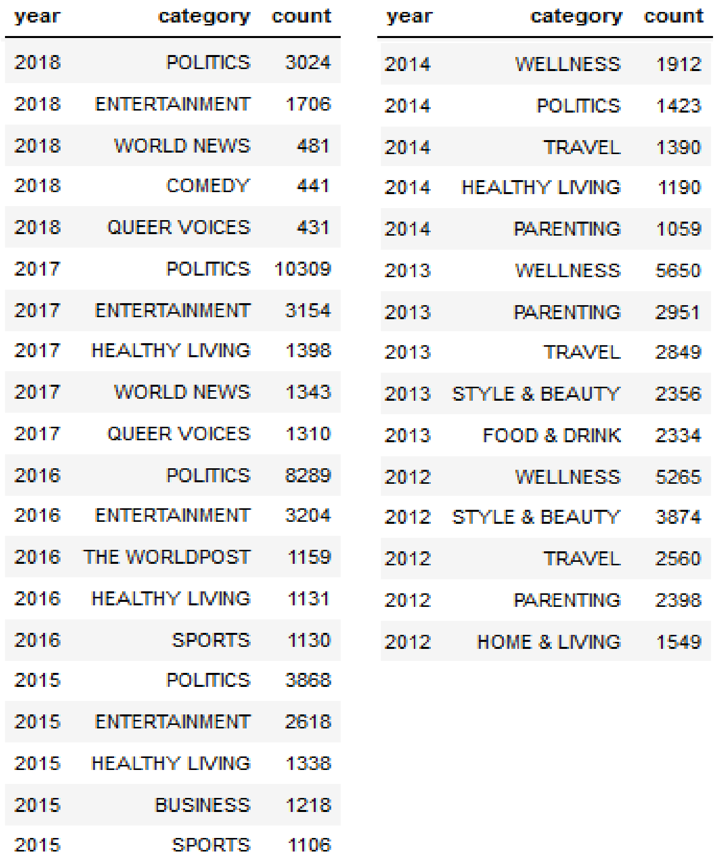
- 41 Categories
- Skewed distribution. Often this affects the classification algorithms

1 - POLITICS	18.06
2 - WELLNESS	8.61
3 - ENTERTAINMENT	8.18
4 - TRAVEL	5.38
5 - STYLE & BEAUTY	4.77
6 - PARENTING	4.30
7 - HEALTHY LIVING	3.40

Total = 52.7% in just 7 of 41 categories

Exploratory Data Analysis (EDA)

Categories by year



year	category	count
2018	POLITICS	3024
2018	ENTERTAINMENT	1706
2018	WORLD NEWS	481
2018	COMEDY	441
2018	QUEER VOICES	431
2017	POLITICS	10309
2017	ENTERTAINMENT	3154
2017	HEALTHY LIVING	1398
2017	WORLD NEWS	1343
2017	QUEER VOICES	1310
2016	POLITICS	8289
2016	ENTERTAINMENT	3204
2016	THE WORLDPOST	1159
2016	HEALTHY LIVING	1131
2016	SPORTS	1130
2015	POLITICS	3868
2015	ENTERTAINMENT	2618
2015	HEALTHY LIVING	1338
2015	BUSINESS	1218
2015	SPORTS	1106
2014	WELLNESS	1912
2014	POLITICS	1423
2014	TRAVEL	1390
2014	HEALTHY LIVING	1190
2014	PARENTING	1059
2013	WELLNESS	5650
2013	PARENTING	2951
2013	TRAVEL	2849
2013	STYLE & BEAUTY	2356
2013	FOOD & DRINK	2334
2012	WELLNESS	5265
2012	STYLE & BEAUTY	3874
2012	TRAVEL	2560
2012	PARENTING	2398
2012	HOME & LIVING	1549

Some useful information

- Since 2015 POLITICS and ENTERTAINMENT are the TOP two categories.
- Until 2014 they didn't appear in the TOP 5 list.
- WELLNESS was the TOP category until 2014.
- Since 2015 it didn't in the TOP 5 categories.

Exploratory Data Analysis (EDA)

Authors

Top 5 authors

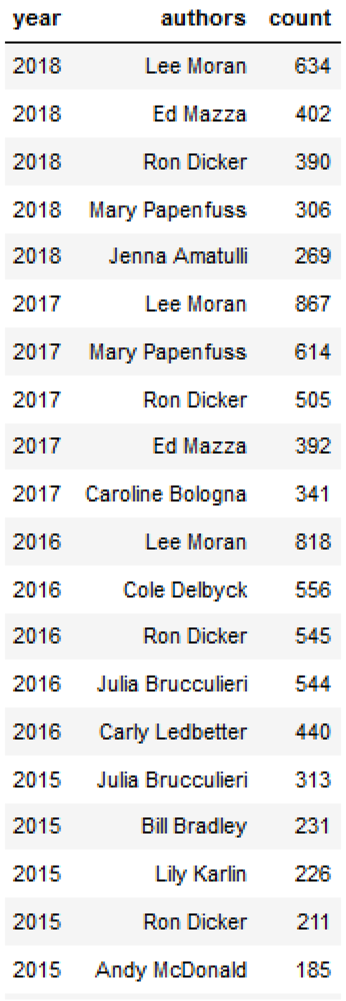
Author	# posts
Lee Moran	2423
Ron Dicker	1762
Reuters	1562
Ed Mazza	1194
Cole Delbyck	1140

Some useful information

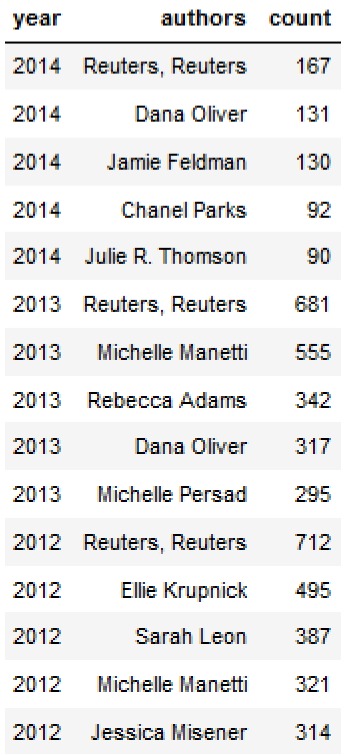
- Total Authors: 148.982
- 50 Top Authors: 25.02% of the total
- 100 Top Authors: 35.0% of the total
- 500 Top Authors: 55.0% of the total

Exploratory Data Analysis (EDA)

Authors by year



year	authors	count
2018	Lee Moran	634
2018	Ed Mazza	402
2018	Ron Dicker	390
2018	Mary Papenfuss	306
2018	Jenna Amatulli	269
2017	Lee Moran	867
2017	Mary Papenfuss	614
2017	Ron Dicker	505
2017	Ed Mazza	392
2017	Caroline Bologna	341
2016	Lee Moran	818
2016	Cole Delbyck	556
2016	Ron Dicker	545
2016	Julia Brucculieri	544
2016	Carly Ledbetter	440
2015	Julia Brucculieri	313
2015	Bill Bradley	231
2015	Lily Karlin	226
2015	Ron Dicker	211
2015	Andy McDonald	185



year	authors	count
2014	Reuters, Reuters	167
2014	Dana Oliver	131
2014	Jamie Feldman	130
2014	Chanel Parks	92
2014	Julie R. Thomson	90
2013	Reuters, Reuters	681
2013	Michelle Manetti	555
2013	Rebecca Adams	342
2013	Dana Oliver	317
2013	Michelle Persad	295
2012	Reuters, Reuters	712
2012	Ellie Krupnick	495
2012	Sarah Leon	387
2012	Michelle Manetti	321
2012	Jessica Misener	314

Some useful information

- Since 2015 Lee Moran is the TOP author with more than 20% publications over the second in the list.
- Some authors left TOP 5 list since 2015, although some of them are in the highest overall position (e.g. Reuters)

Classification using supervised learning

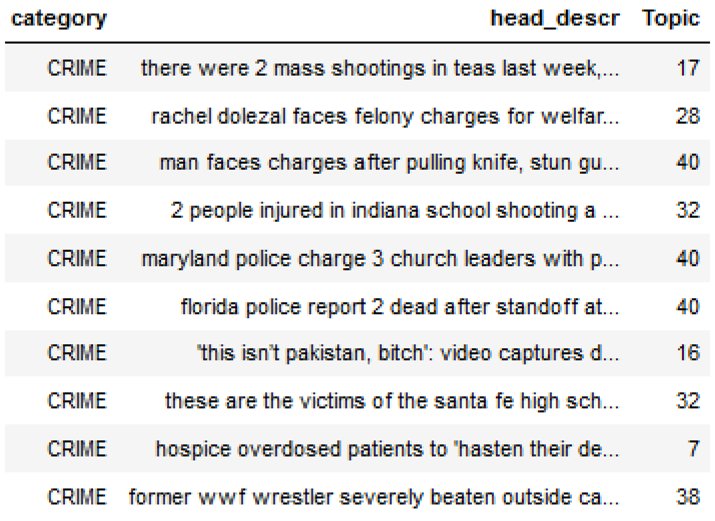
Classifier	Accuracy
Naïve Bayes (NB)	38%
Support Vector Machine (SVM)	55%
NB with GridSearch + Cross Validation	56%
SVM with GridSearch + Cross Validation	56%
NB with NLTK	49%

Some useful information

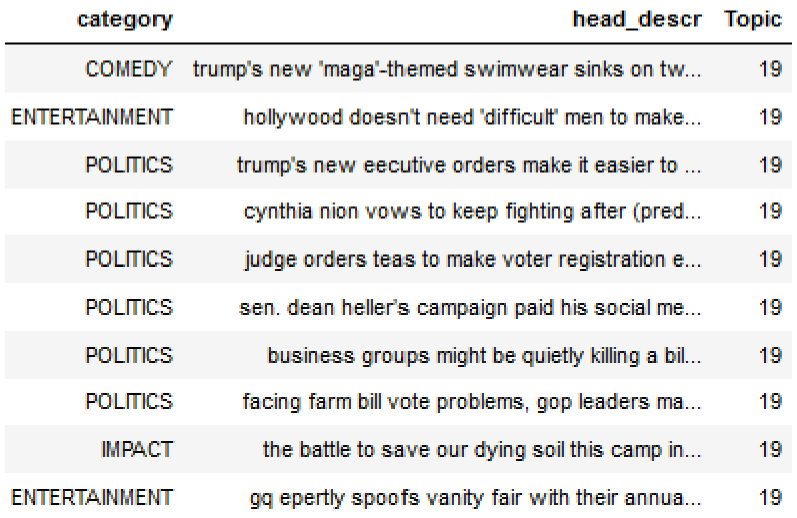
- Imbalanced classification problem due to skewed distribution.
- Due to high number of categories, classification algorithms didn't have a high accuracy.
- GridSearch and Cross Validation helped classifiers to get better tuning parameters settings and improving their accuracies.

Classification using unsupervised learning

Technique: Non-negative Matrix Factorization



category	head_descr	Topic
CRIME	there were 2 mass shootings in teas last week,...	17
CRIME	rachel dolezal faces felony charges for welfar...	28
CRIME	man faces charges after pulling knife, stun gu...	40
CRIME	2 people injured in indiana school shooting a ...	32
CRIME	maryland police charge 3 church leaders with p...	40
CRIME	florida police report 2 dead after standoff at...	40
CRIME	'this isn't pakistan, bitch': video captures d...	16
CRIME	these are the victims of the santa fe high sch...	32
CRIME	hospice overdosed patients to 'hasten their de...	7
CRIME	former wwf wrestler severely beaten outside ca...	38



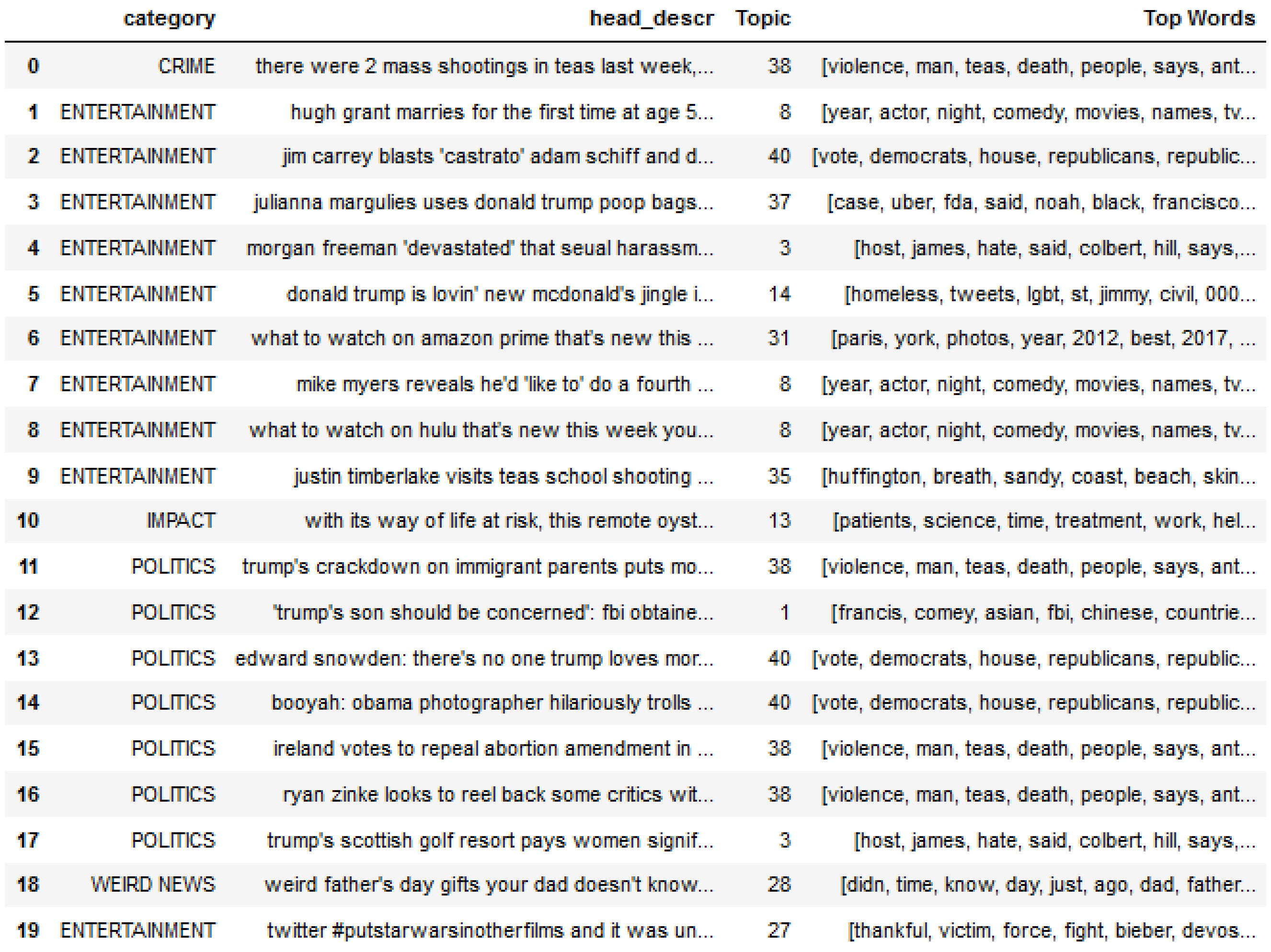
category	head_descr	Topic
COMEDY	trump's new 'maga'-themed swimwear sinks on tw...	19
ENTERTAINMENT	hollywood doesn't need 'difficult' men to make...	19
POLITICS	trump's new eecutive orders make it easier to ...	19
POLITICS	cynthia nion vows to keep fighting after (pred...	19
POLITICS	judge orders teas to make voter registration e...	19
POLITICS	sen. dean heller's campaign paid his social me...	19
POLITICS	business groups might be quietly killing a bil...	19
POLITICS	facing farm bill vote problems, gop leaders ma...	19
IMPACT	the battle to save our dying soil this camp in...	19
ENTERTAINMENT	gq epertly spoofs vanity fair with their annua...	19

Some useful information

- Imbalanced classification problem due to skewed distribution.
- Due to high number of categories, classification algorithms didn't have a good performance.
- It is possible to notice that existing categories and topics inferred by techniques aren't well aligned

Classification using unsupervised learning

Technique: Latent Dirichlet Allocation (LDA)



	category	head_descr	Topic	Top Words
0	CRIME	there were 2 mass shootings in teas last week,...	38	[violence, man, teas, death, people, says, ant...
1	ENTERTAINMENT	hugh grant marries for the first time at age 5...	8	[year, actor, night, comedy, movies, names, tv...
2	ENTERTAINMENT	jim carrey blasts 'castrato' adam schiff and d...	40	[vote, democrats, house, republicans, republic...
3	ENTERTAINMENT	julianna margulies uses donald trump poop bags...	37	[case, uber, fda, said, noah, black, francisco...
4	ENTERTAINMENT	morgan freeman 'devastated' that seual harassm...	3	[host, james, hate, said, colbert, hill, says,...
5	ENTERTAINMENT	donald trump is lovin' new mcdonald's jingle i...	14	[homeless, tweets, lgbt, st, jimmy, civil, 000...
6	ENTERTAINMENT	what to watch on amazon prime that's new this ...	31	[paris, york, photos, year, 2012, best, 2017, ...
7	ENTERTAINMENT	mike myers reveals he'd 'like to' do a fourth ...	8	[year, actor, night, comedy, movies, names, tv...
8	ENTERTAINMENT	what to watch on hulu that's new this week you...	8	[year, actor, night, comedy, movies, names, tv...
9	ENTERTAINMENT	justin timberlake visits teas school shooting ...	35	[huffington, breath, sandy, coast, beach, skin...
10	IMPACT	with its way of life at risk, this remote oyst...	13	[patients, science, time, treatment, work, hel...
11	POLITICS	trump's crackdown on immigrant parents puts mo...	38	[violence, man, teas, death, people, says, ant...
12	POLITICS	'trump's son should be concerned': fbi obtaine...	1	[francis, comey, asian, fbi, chinese, countrie...
13	POLITICS	edward snowden: there's no one trump loves mor...	40	[vote, democrats, house, republicans, republic...
14	POLITICS	booyah: obama photographer hilariously trolls ...	40	[vote, democrats, house, republicans, republic...
15	POLITICS	ireland votes to repeal abortion amendment in ...	38	[violence, man, teas, death, people, says, ant...
16	POLITICS	ryan zinke looks to reel back some critics wit...	38	[violence, man, teas, death, people, says, ant...
17	POLITICS	trump's scottish golf resort pays women signif...	3	[host, james, hate, said, colbert, hill, says,...
18	WEIRD NEWS	weird father's day gifts your dad doesn't know...	28	[didn, time, know, day, just, ago, dad, father...
19	ENTERTAINMENT	twitter #putstarwarsinotherfilms and it was un...	27	[thankful, victim, force, fight, bieber, devos...

Some useful information

- Imbalanced classification problem due to skewed distribution.
- Due to high number of categories, classification algorithms didn't have a good performance.
- It is possible to notice that existing categories and topics inferred by techniques aren't well aligned
- We can also see that some TOP WORDS aren't related with existing categories

Suggestions for improvements

- Decrease the amount of categories or create sub-categories.
- Balance dataset.
- To validate data before including in dataset in order to avoid missing data.

THANK YOU