Springboard

Web content sentimental analysis: News

Karolina Bzdusek

Springboard

June 30th, 2019

Contents

1. Introduction

Automatization is a crucial part of any business. On the Internet we have immense source of data and information. To maintain web content is highly important four owner of the various websites to have labeled their data correctly. To define categories, or keywords that would help users to get around in their content.

We have concentrate of one of the above mentioned applications - tagging articles from news.

This dataset contains 200853 news headlines from the year 2012 to 2018 obtained from HuffPost. The model trained on this data set is used to identify tags for untracked news articles.

The question that we want the answer is the following:

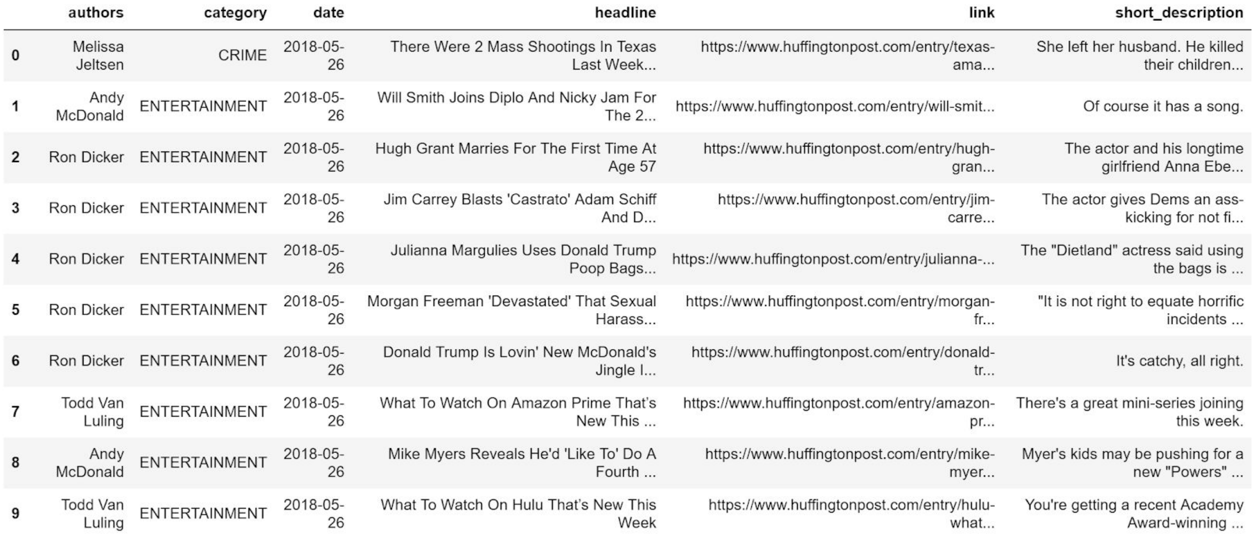
Can we categorize news articles based on their headlines and short descriptions?

Source of data set: https://rishabhmisra.github.io/publications/

Data contain 6 features : authors, category, date, headlie, link and short description. Features link and date has no values in terms of predicting categories[1], so we would not consider them.

Typically authors is focus in some area within they are publishing their articles. However, in their main focus is or example 'crime', it does not mean that this is exclusive category for them. They may also publish within categories such as 'parents' or 'politics'. This could lead to a leakage of the data. Therefore he feature 'author' we will not consider as well and our model would be based purely on 'headline' and 'short description" features.

[1] This is not entirely true, if something shocking happens and it "it is all over the news" then date could play a role in weighting some category more.



	authors	category	date	headline	link	short_description
0	Melissa Jeltsen	CRIME	2018-05-26	There Were 2 Mass Shootings In Texas Last Week...	https://www.huffingtonpost.com/entry/texas-ama...	She left her husband. He killed their children...
1	Andy McDonald	ENTERTAINMENT	2018-05-26	Will Smith Joins Diplo And Nicky Jam For The 2...	https://www.huffingtonpost.com/entry/will-smit...	Of course it has a song.
2	Ron Dicker	ENTERTAINMENT	2018-05-26	Hugh Grant Marries For The First Time At Age 57	https://www.huffingtonpost.com/entry/hugh-gran...	The actor and his longtime girlfriend Anna Ebe...
3	Ron Dicker	ENTERTAINMENT	2018-05-26	Jim Carrey Blasts 'Castrato' Adam Schiff And D...	https://www.huffingtonpost.com/entry/jim-carre...	The actor gives Dems an ass-kicking for not fi...
4	Ron Dicker	ENTERTAINMENT	2018-05-26	Julianna Margulies Uses Donald Trump Poop Bags...	https://www.huffingtonpost.com/entry/julianna-...	The "Dietland" actress said using the bags is ...
5	Ron Dicker	ENTERTAINMENT	2018-05-26	Morgan Freeman 'Devastated' That Sexual Harass...	https://www.huffingtonpost.com/entry/morgan-fr...	"It is not right to equate horrific incidents ...
6	Ron Dicker	ENTERTAINMENT	2018-05-26	Donald Trump Is Lovin' New McDonald's Jingle I...	https://www.huffingtonpost.com/entry/donald-tr...	It's catchy, all right.
7	Todd Van Luling	ENTERTAINMENT	2018-05-26	What To Watch On Amazon Prime That's New This ...	https://www.huffingtonpost.com/entry/amazon-pr...	There's a great mini-series joining this week.
8	Andy McDonald	ENTERTAINMENT	2018-05-26	Mike Myers Reveals He'd 'Like To' Do A Fourth ...	https://www.huffingtonpost.com/entry/mike-myer...	Myer's kids may be pushing for a new "Powers" ...
9	Todd Van Luling	ENTERTAINMENT	2018-05-26	What To Watch On Hulu That's New This Week	https://www.huffingtonpost.com/entry/hulu-what...	You're getting a recent Academy Award-winning ...

As it is an NLP task, the dimensionality is high (each unique word is a feature), therefore some cleaning of the text is needed. Other possibility is stemming/lemmatization/stop words techniques and choice of our tokenizer as well. These are ways to reduce dimensionality before transforming words to vectors. Another would be to determine features that are the most helpful to recognize categories, in other words dimension reduction (using a technique such as PCA).

2. Exploratory Data Analysis

2.1 Data Wrangling

As we are dealing with NLP, it is necessary to transform text into numbers. First of all we want to preprocess our text data. First we get rid of punctuation and we are using TweetTokenizer to create tokens. Other hyperparameters that we are using are stopwords.

Two approaches would be used : using nltk library and SpaCy library. Text features would be transform to TFIDF matrix.

All text features are labelled. Originally data contained 41 features. There are some labels that are almost the same and were merged as the same category. Such as:

- 'THE WORLDPOST', 'WORLDPOST'
- 'ARTS' ,'CULTURE & ARTS', 'ARTS & CULTURE'
- 'PARENTING’, 'PARENTS'
- 'STYLE', 'STYLE & BEAUTY'
- 'COLLEGE', 'EDUCATION'
- 'TASTE', 'FOOD & DRINK'

This resulted to have 34 categories instead of 41. Categories are not equally represented which leads to solving multi-classification problem with imbalanced dataset.

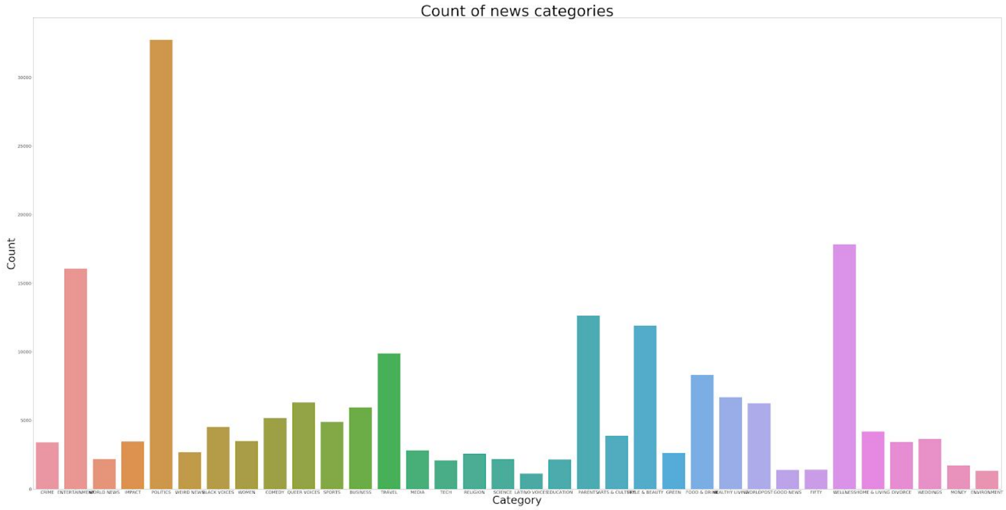
2.2 Data Storytelling

Let’s take a look on those categories:

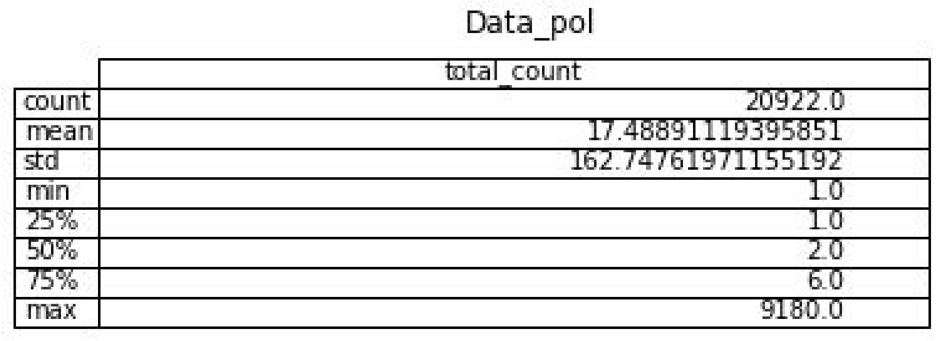
Category	Occurrences
POLITICS	32739
WELLNESS	17827
ENTERTAINMENT	16058
PARENTS	12632
STYLE & BEAUTY	11903
TRAVEL	9887
FOOD & DRINK	8322
HEALTHY LIVING	6694
QUEER VOICES	6314
WORLDPOST	6243
BUSINESS	5937
COMEDY	5175

SPORTS	4884
BLACK VOICES	4528
HOME & LIVING	4195
ARTS & CULTURE	3878
WEDDINGS	3651
WOMEN	3490
IMPACT	3459
DIVORCE	3426
CRIME	3405
MEDIA	2815
WEIRD NEWS	2670
GREEN	2622
RELIGION	2556
SCIENCE	2178
WORLD NEWS	2177
EDUCATION	2148
TECH	2082
MONEY	1707
FIFTY	1401
GOOD NEWS	1398
ENVIRONMENT	1323
LATINO VOICES	1129

Here you can see histogram of previous mentioned categories and their occurrences through histogram graph.



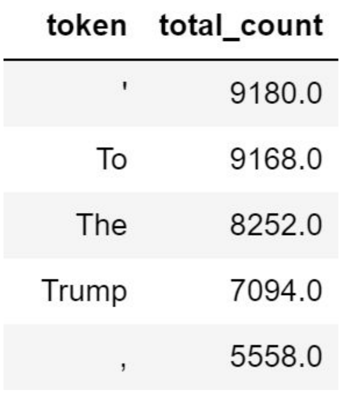
Let’s look closer on category ‘politics’ (category with the highest number of articles). How many unique words does the feature ‘headline’ contain? (after preprocessing).



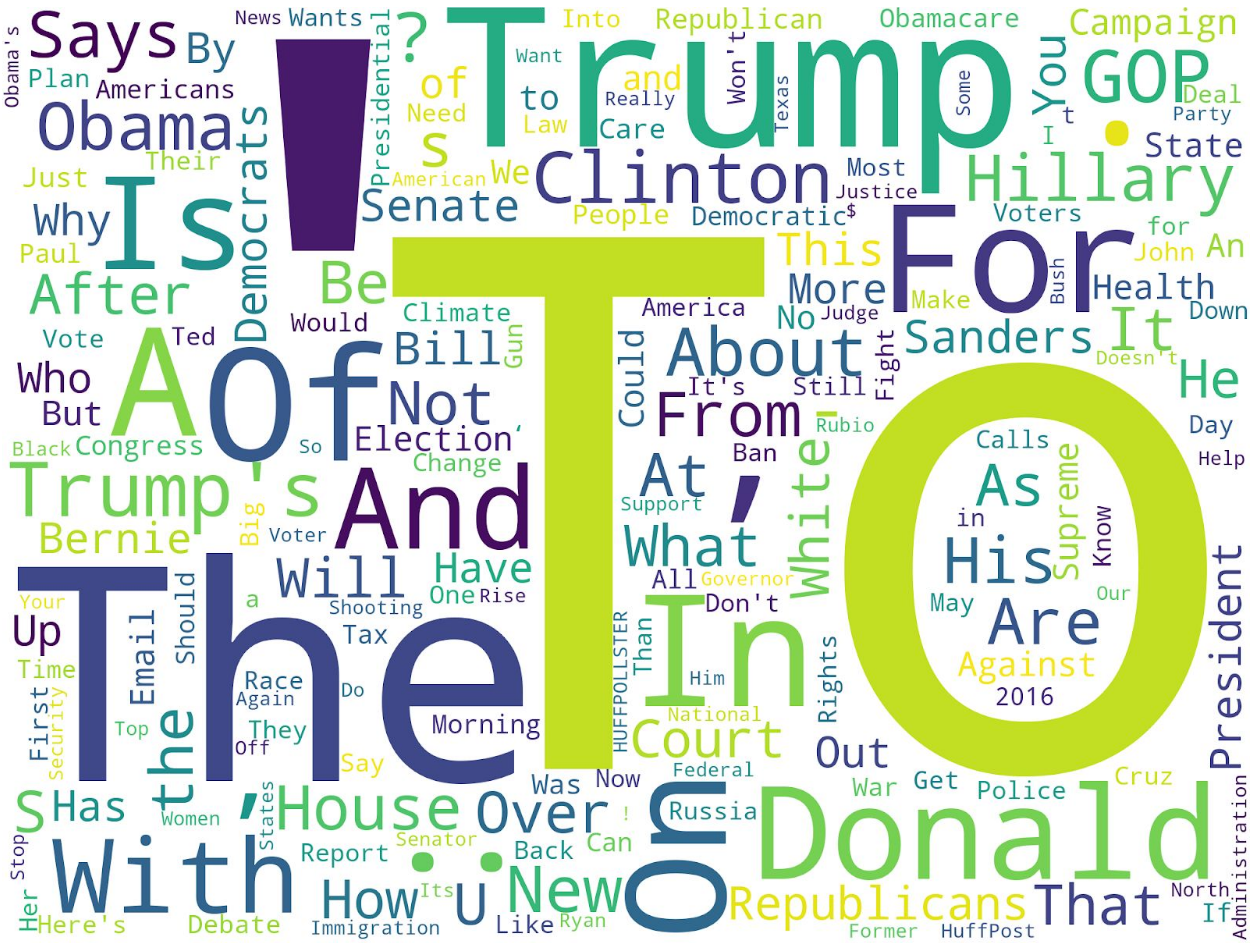
Data_pol

	total_count
count	20922.0
mean	17.48891119395851
std	162.74761971155192
min	1.0
25%	1.0
50%	2.0
75%	6.0
max	9180.0

Just for category politics we have 20922 unique tokens (and we have 34 categories!). Headline contain emojis (e.g. tacos emoji). This tokens where counted after our preprocessing. Unique words statistics table shows us, that only small portion of data set is occuring more than six times and maximal frequency(=occurrence) is 9180.



token	total_count
'	9180.0
To	9168.0
The	8252.0
Trump	7094.0
,	5558.0



Note: How to understand WordCloud? Feature color you should omit entirely. The only relation is that the bigger font size the more frequent word is.

2.3 Inferential Statistics

In our data set of News we have 43 categories. We want to find out, whether some features is more important than others in order to predict whether it belongs to one category and the others. We are comparing in this example categories ‘Politics’ and “Media’. For the sake of demonstration we will pick just one word ("Donald").

Firstly, we need to find out whether the CLT (Central Limit Theorem) applies:

1) We have 32739 rows for ‘Politics’ and 2815 rows for ‘Media, so N is large enough
2) We assume that independent condition is satisfied as well

We state our null and alternative hypothesis and then we test them. We set $\boldsymbol{\alpha}$ = 0.5

$\mathbf{H_0}$: $\mu_{politics} = \mu_{medial}$

$\mathbf{H_A}$: $\mu_{politicsl} \neq \mu_{medial}$

In this Jupyter Notebook you can find details about testing our hypothesis. We here concluded our findings.

Our 95% confidence interval is <0.011909, 0.0316317>. And therefore we can reject null hypothesis. This is a result we would expect, to see word 'Donald' more often in 'Politics' then in 'Media' category.