News Classification using NLP

DECEMBER 12

BrainStation, Vancouver
Written by: Harsh Sharma

BrainStation®

Project Problem Statement

Classification of News based on the Headline Text given.

There are wide variety of news available on the internet and different media platforms. The range of the types of the news varies from a world level to a personal level. In this project I'm making machine learning (ML) models to identify the right category of the news, just by their respective headlines.

How will this help the business or the society: Gone are the times in which people used to wait for their newspapers to come in the morning and they would read it with their morning tea. This is the 21st century and the so-called information age, where nobody wants a hardcopy of anything to deal. Everybody wants to shift their lives digitally in every possible terms.


Cows lose their jobs as milk prices drop

By By Scott Calvert

EASTON, Md. — As his 100 dairy cows lumbered over for their Monday afternoon milking, farmer Eric Foster pondered his sudden misfortune. Those Holstein and Jerseys, profit machines during a recent milk

This unlikely combination of forces has hit Foster's milking parlor and other dairy operators with a vengeance. After soaring in 2007 and remaining high in much of 2008, milk prices paid to farmers have collapsed and are expected to remain dismal, even as feed and fuel stay fairly costly. Some dairy farmers have shut

Economists are concerned

Although, consumers sh benefit from somewhat lo prices at the supermarket, drop in the value of milk p cents a gion outlook for d


Woman missing since she got lost


ANIMAL WATCH

The owls are not what they seem

ARLINGTON, North Military Road, 3000 block, 5:20 p.m. Nov. 10. Responding to a call about an injured owl on the side of the road, an officer found a large mushroom.

Do the above snippets of news make any sense to you?? No, right?
This is the reason why relevancy of news is important to customers and classification of news is the footstep of that process.

Why target news to target audience is important: Everybody wants to have relevance and reality in their lives, and nobody has time to read a big newspaper on their mobile device. Also, competition has grown to such an extent in every industry that nobody willing to pay a whole sum of money on every type of news. Corporations want efficiency and maximum profit. That's why ML is crucial in this area for classifying the news into the right category and then recommending it.

Background on the subject matter area:

News apps or websites already have different type of categories in their features. Which means they are already doing a good job in terms of classifying the news, but what I feel they are lacking in is targeting right news at the right audience, which as discussed earlier is very necessary. Data Science are the right fit for these kinds of problems because a lot of information and trends can be extracted from the data provided. Things like patterns of the writing styling of different categories, kind of authors (their backgrounds) that publish certain kind of news and popularity of certain kind of news along with time demand some data analysis and machine learning to be applied on to arrive at right results.

Details on the dataset:

Headline of the news articles	Short description of news articles	Category	Authors	Date
This is the main column/feature in the dataset	Short description of the news articles giving some more context to the meaning of the headline.	There are 42 types of news in the dataset, like world news, political new, entertainment...	This columns has the names of the authors There can be more than 1 author for a news	Date in which the news was posted on the website.
On an average, there are 9 words in a headline	9.4 % of the total rows, information is missing in terms of short description	This is the dependent or the targeted factor.	Job occupations are also mentioned for some of the authors	News vary from 2012 till 2022

The dataset which is being dealt with here came from Kaggle. Originally data is from Huffpost which is an American liberal news portal with localized and international relations. Their official site provides news, blogs, satires, and original content, and covers politics, business, entertainment and 39 more variety of featuring columnists. However, this news classification dataset is not uploaded by Huffpost itself on Kaggle, but by Risabh Misra who is senior ML engineer at Twitter. And this dataset was updated lastly in September 2022.

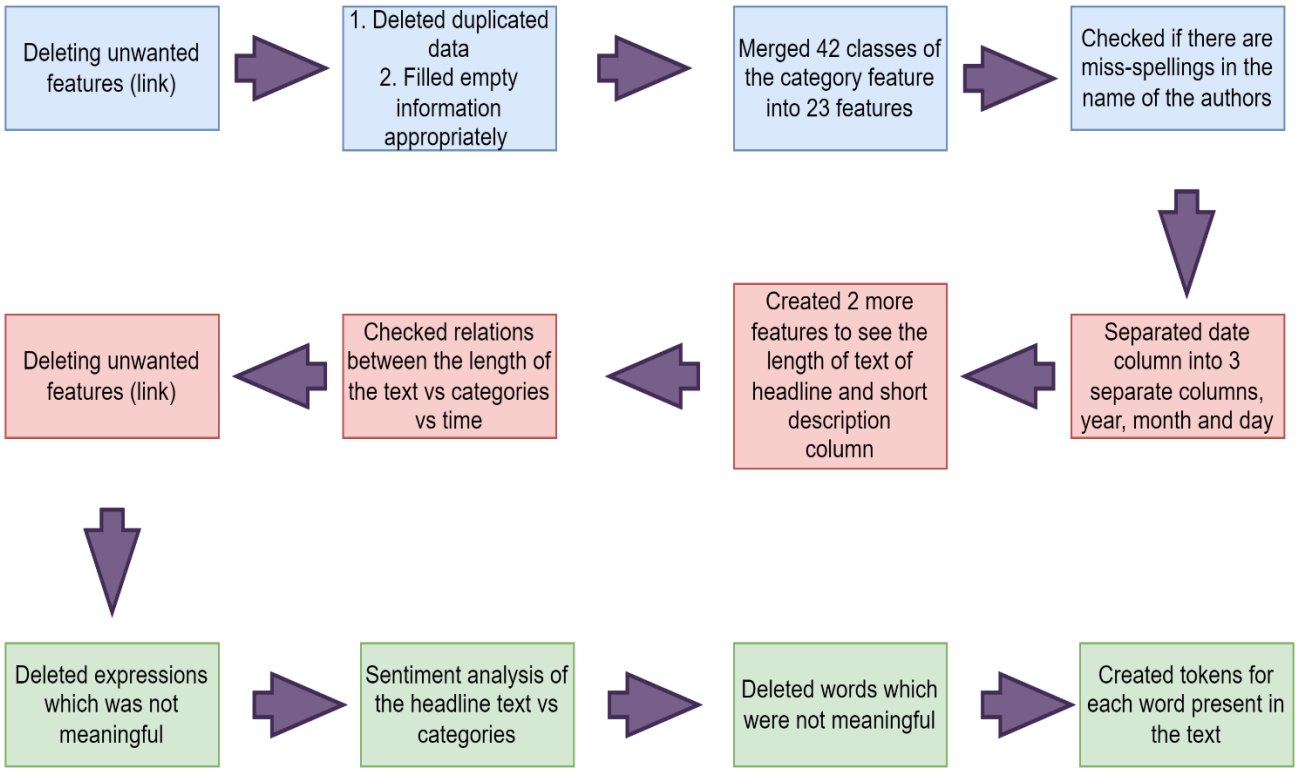
Data file provided on Kaggle is in **json** format with size of **87.3 MB.**

It has:

1. 209527 rows and 6 columns
2. 99 duplicated rows
3. No null values, but 9.4% empty strings in short description column and 17.8% empty strings in authors column.

Summary of Preprocessing and cleaning:

Preprocessing: The text of headline and short description contains a lot of unwanted text like words which do not have any meaning. These words must be deleted or omitted to make the model more efficient to successfully classify the category of the news. There were several steps taken to clean the data like **stop-word removal, stemming, lemmatization and ultimately tokenizing the text using TF-IDF.** Firstly, only the headline text is only dealt with to see the accuracy of models, then text of short description is also added to have a check on the accuracy and ultimately prefer the highest performing model.

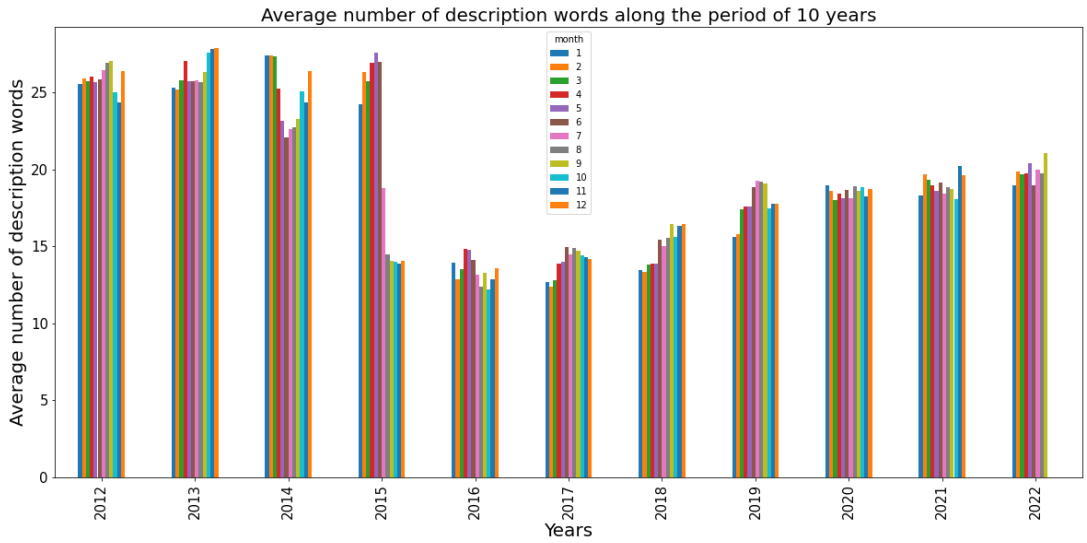


Flowchart explaining the process of cleaning and analysis of data

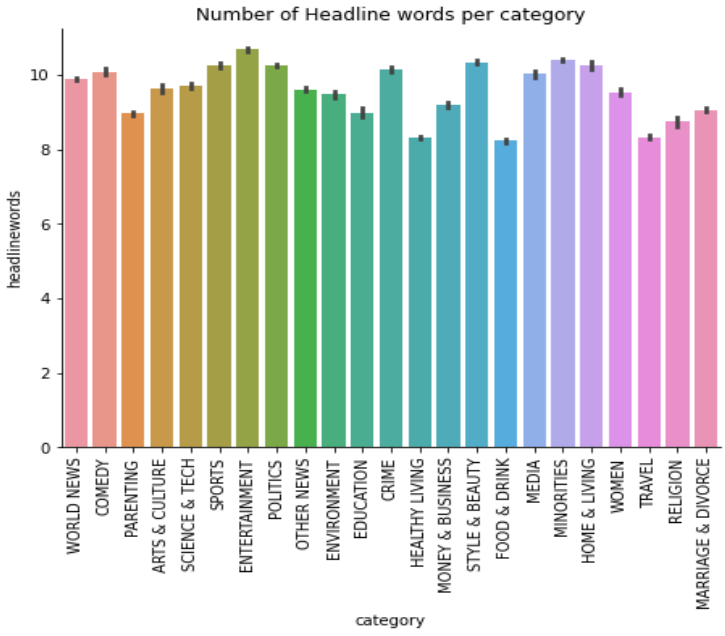
1. **Number of words** in headline and short description were checked to see the relation of length of text for different categories and along the period of 11 years.
2. A short sentiment analysis **(Polarity and subjectivity)** was checked to see the writing behavior of different classes of category of news.
3. Words like **and, for, the, a,** is and so on were deleted to just keep the text which makes sense to the algorithm.
   Signs or expressions like **/?,<>.+=….** Were also excluded from the text of headline and short description.

Insights, modelling, and results:

Insights:

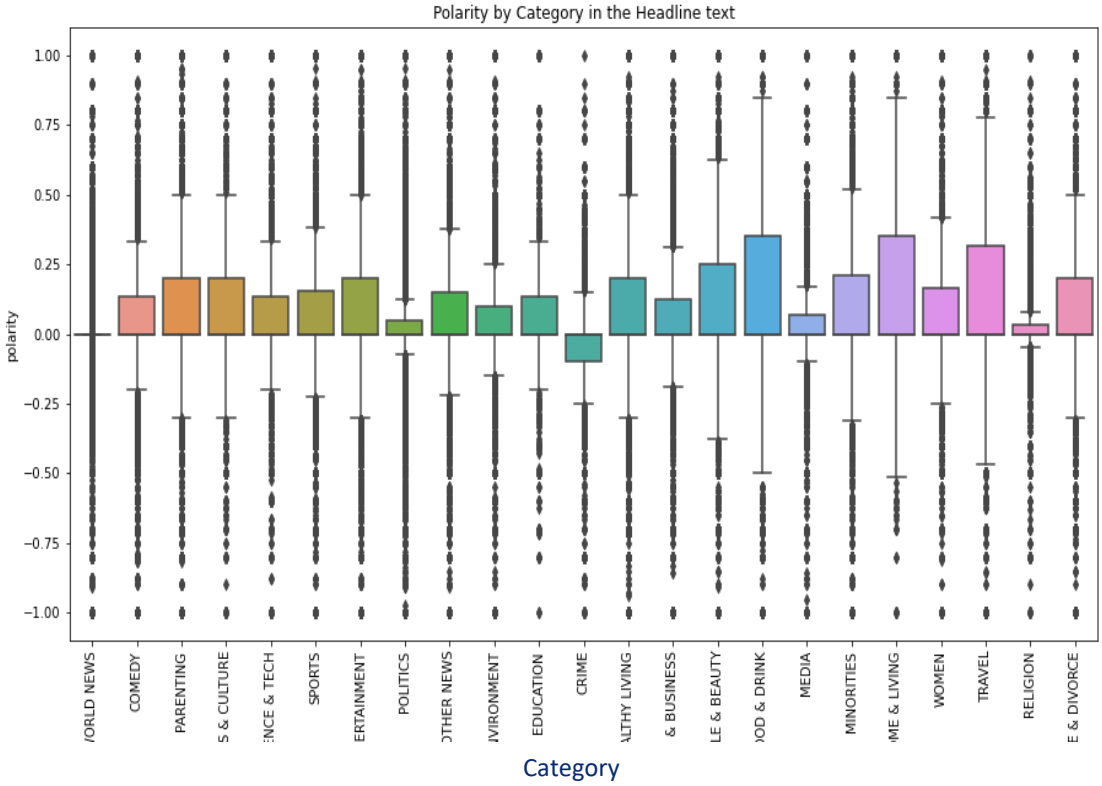


As can be seen from above image that the average number of description words decreased almost half of the value it was in mid-2015. However, after dropping lowest point in sept2017, the average number of words started increasing thereafter.
Although the average number of words decreased in between, but average number of headline words increased by a small amount each year.



If I talk about number of words by category, **ENTERTAINMENT** and **MINORITIES** were the top classes to have to lengthy texts. On the same page **HEALTHY LIVING** and **FOOD & DRINK** were categories with shortest text in their headlines.

Overall, there were **9 words** on average in a headline of any type of news and average **25 words in the description** of the news.



The above figure shows the use of positive and negative words in all the categories. Polarity measures the positiveness above 0 and vice versa. As it can be expected that **CRIME** has the greatest number of negative words in its bag. **FOOD & DRINK** and **HOME & LIVING** are the ones who have most positive words followed by **TRAVEL**.

Modelling:

There were 42 categories which I merged into 23 which had a little bit of similar context. As this was multi-class problem, logistic regression and SVM would not work well.
Therefore, I chose **random forest classifier, XGBoost classifier and a neural network** model to classify text of headlines and descriptions into **23 categories.**

Results:

Models	Accuracy
Random Forest Classifier	53.18%
XGBoost Classifier	55%
Neural Network	62.40%

Findings and conclusions:

For me the biggest challenge in this project has been to classify headlines of news into 42 classes. I already knew it would not have been easy with quite simple algorithms to classify the news just by headlines. That's why I kept it short to 23 categories.
However, I expected that the accuracy of my models would be par 70%. It's still achievable with some advanced algorithms.
Sentiment of some classes like crime, politics, house & living and so on were as expected.
Interestingly, classes which have higher average number of words in their headlines, tend to have shorter average number of words in their short descriptions.

After we are successful with classifying news into categories, what are the implications and prospects!?
Below are some of my thoughts…

1. Classifying news is just the first part of process of recommending news to someone and dig deeper into network analysis.
2. Fake news keeps on circulating on different online platforms. Next step is to detect fake news in each type of news. Every type of news, for example, politics have different level of complexity associated to them. So, it would be easy to detect fakeness once we are able to detect them.
3. What is the impact of certain kind of news on the economy.

THANK YOU