Анализ онлайн популярности новостей

Научно-исследовательский семинар
"Методология проектной работы"

Сидоренко Марина
Сотникова Анастасия

БЭАД222

November 9, 2023

Contents

1 Введение

В рамках данного задания по дисциплине Научно-исследовательский семинар "Методология проектной работы" мы должны были провести базовый анализ выбранного нами набора данных, содержащего информацию об онлайн популярности новостей. В процессе выполнения данного задания мы хотела научиться основным инстурментам анализа данных, моделям предсказания и, самое главное, научиться правильно интерпретировать и описывать результаты и делать верные выводы.

2 Постановка задачи

Цель проекта: выявить, какие признаки могут влиять популярность новостей на различных платформах, а также построить модели, которые бы могли неплохо предсказывать итоговую популярность новости.

Задачи:

1. Ознакомление с набором данных
2. Первичная обработка данных
3. Проведение разведочного анализа данных
4. Построение и анализ моделей

3 Описание данных

Наш набор данных содержит 13 файлов формата *csv*. 12 из них содержат данные о популярности новостей по платформам и по тематикам на протяжении первых двух дней с момента публикации. Содержание этих файлов описано в таблице ниже:

Название колонки	**Тип данных**	**Описание данных**
IDLink	numeric	Уникальный идентификатор новости
TS1	numeric	Популярность новости в промежутке 0-20 мин
TS2	numeric	Популярность новости в промежутке 20-40 мин
...	...	...
TS144	numeric	Итоговая популярность спустя 2 дня

Последний файл содержит датафрейм с описание каждой из новостей:

Название колонки	**Тип данных**	**Описание данных**
IDLink	numeric	Уникальный идентификатор новости
Title	string	Название статьи
Headline	string	Заголовок статьи
Source	string	Источник новости
Topic	string	Тематика новости
PublishDate	timestamp	Дата публикации новости
SentimentTitle	numeric	Sentiment score названия
SentimentHeadline	numeric	Sentiment score заголовка
Facebook	numeric	Итоговая популярность на Facebook
GooglePlus	numeric	Итоговая популярность на Google+
LinkedIn	numeric	Итоговая популярность на LinkedIn

Всего в датасете представлены данные по 4 тематикам: экономика, Обама, Палестина и Майкрософт.

Sentiment score – это показатель, оценивающий эмоциональную окрашенность текста (в нашем случае названия и заголовка новости). Он варьируется от -1 до 1, где -1 – сильная негативная окрашенность, 1 – сильная положительная окрашенность, 0 – отсутствие эмоциональной окрашенности.

4 Разведочный анализ данных

4.1 Первичная обработка данных

В первую очередь мы проверили наличие незаполненных полей в датафреймах. Пропуски были обнаружены в таблице NewsFinal в колонках Headline и Source в количестве 15 и 279 соответственно. Так как этих значений достаточно мало, а данные признаки не представляют большого интереса для дальнейшего исследования, то на эти пропуски можно закрыть глаза.

Мы решили посмотреть на то, как распределены наиболее интересующие нас признаки: а именно итоговая популярность новостей и с этой целью построили гистограммы (*Рисунок 1-3*).

Figure 1: Распределение популярности новостей на Google+ с учетом тематики

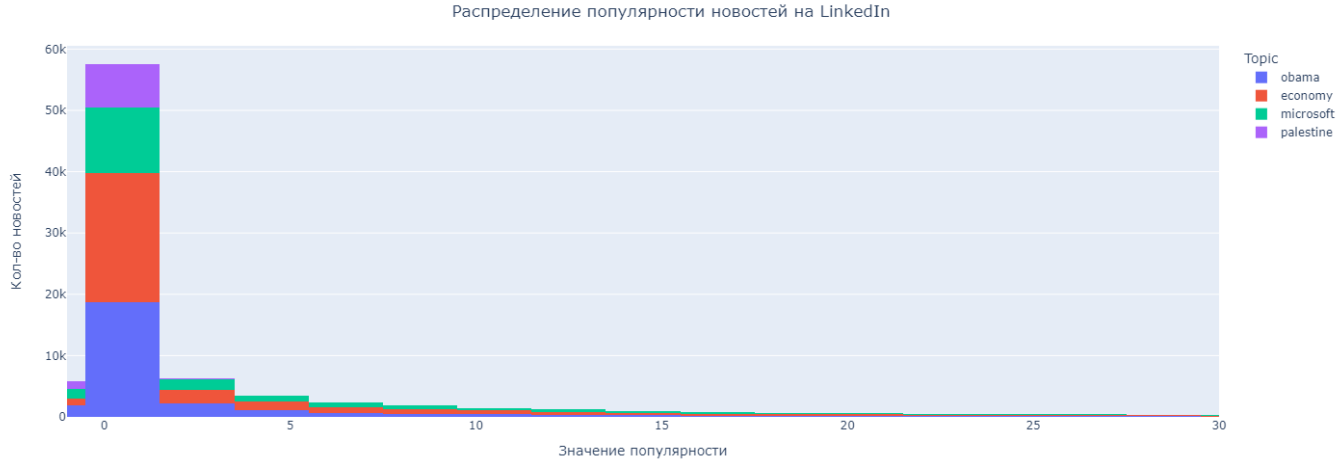

Figure 2: Распределение популярности новостей на LinkedIn с учетом тематики

Figure 3: Распределение популярности новостей на Facebook с учетом тематики

Также мы добавили в датафрейм новый признак – срднее значение итоговой популярности, тем самым усреднив показатели по платформам. Распределение данной величины показано на *Рисунке 4*.

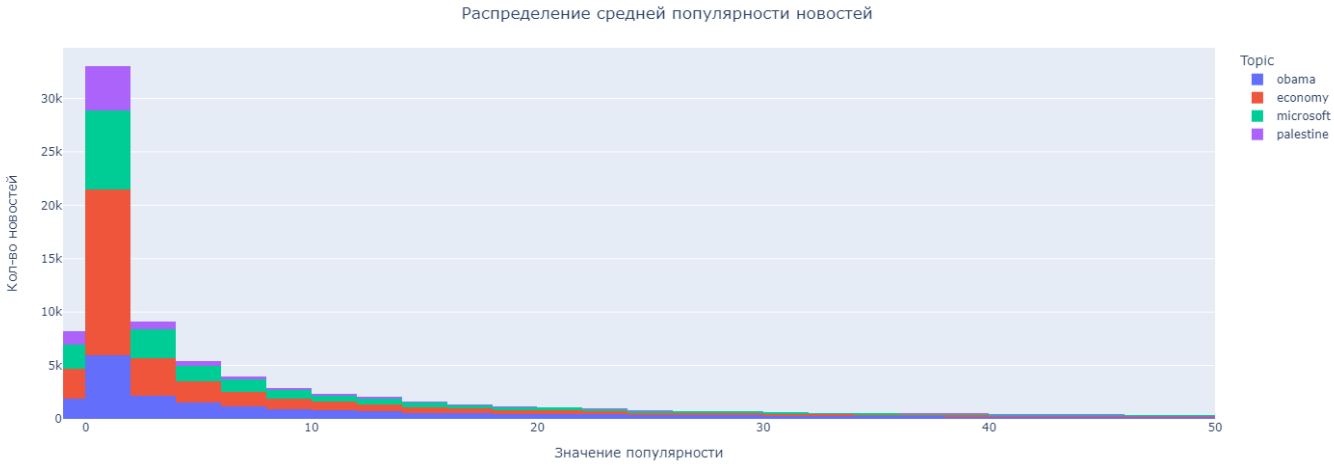

Figure 4: Распределение средней популярности новостей с учетом тематики

Несложно заметить, что основная масса наблюдений сконцентрирована около небольших значений популярности, а большинство остальных наблюдений с большой вероятностью являются выбросами. Для начала мы решили избавиться от тех записей, у которых значение популярности равняется -1. Далее мы построили графики типа "ящик с усами", чтобы проанализировать выбросы (*Рисунок 5*).

Как можно заметить, выбросов действительно катастрофически много. Почистим наши данные от них следующим образом: оставим только те наблюдения, популярность которых на Google+, LinkedIn, Facebook не превосходит 10, 15 и 60 соответственно. Посмотрим, как изменилась ситуация после манипуляций (*Рисунок 6*).

Некоторые выбросы все еще присутствуют, но в целом ситуация стала значительно лучше. После данной обработки у нас осталось 56890 записей.

Далее посмотрим на то, как распределены новости по тематикам (*Рисунок 7*). Заметно, что больше всего новостей написано на тему экономики, а меньше всего – Палестины.

Обратимся теперь к остальным 12 файлам. Так как все они имеют одинаковую структуру, то логично посмотреть на один из них. Исходя из предыдущего анализа, наиболее приятное

Figure 5: Распределение популярности по платформам

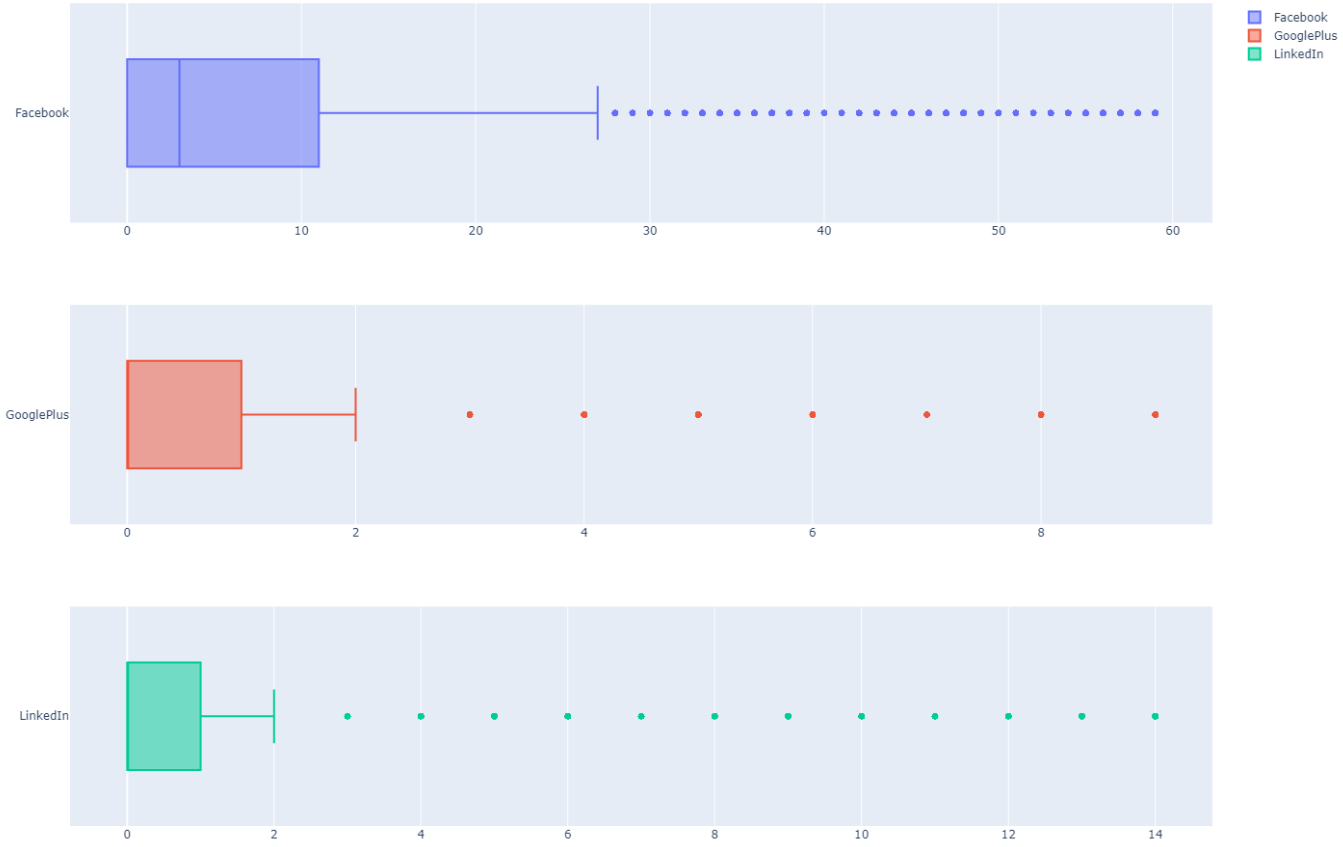

Figure 6: Распределение популярности по платформам

Figure 7: Распределение новостей по тематикам

распределение имеют значения популярности на Facebook, а больше всего новостей по теме экономики, поэтому возьмем файл *Facebook_ Economy.csv*. Сразу заметим, что все значения в датафрейме числовые, а пропуски отсутствуют.

Для анализа распределния построим ящик с усами для итогового уровня популярности (*Рисунок 8*). Далее исключим нерелевантные значения (-1) и избавимся от сильных выбросов, проверим итоговое распределение (*Рисунок 9*).

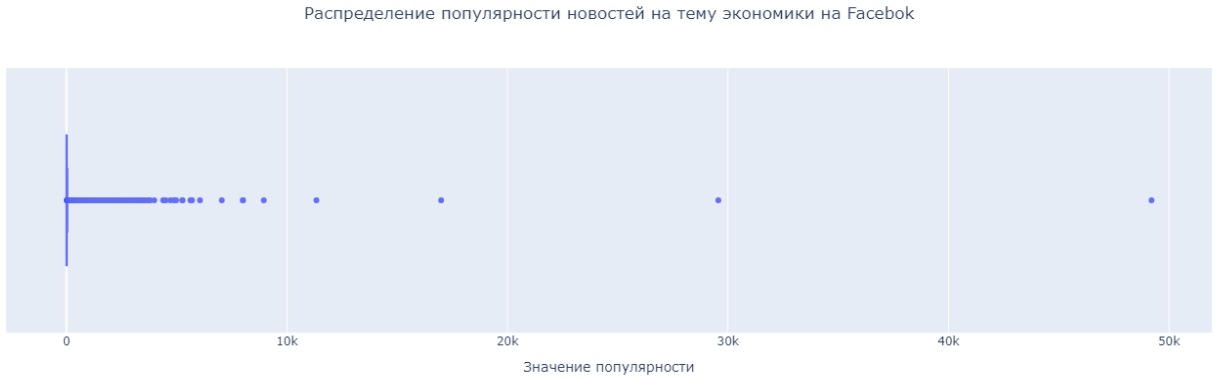

Figure 8: Распределение итоговой популярности новостей для новостей на тему экономики на Facebook

4.2 Анализ зависимостей

Перейдем к анализу зависимостей между признаками в датасете. В первую очередь с этой целью мы построили корреляционную матрицу всех релевантных числовых значений из таблицы *News_ Final* и нового призака – *TitleLength*, отображающего длину названия статьи в символах (*Рисунок 10*). Мы добавили новый признак, так как у нас появилась гипотеза о том, что данный признак может влиять на популярность новости, так как, например, более длинные названия могут привлекать большее внимание.

Заметим, что признаки, отображающие Sentiment score, слабо скоррелированны между собой, а также прослеживается взаимосвязь между итоговыми показателями популярности на платформах, что абсолютно логично.

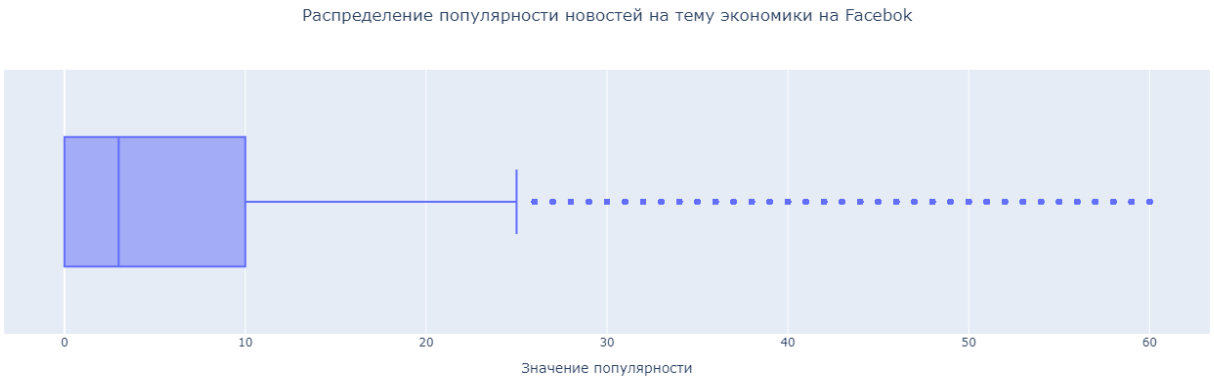

Figure 9: Распределение итоговой популярности новостей для новостей на тему экономики на Facebook

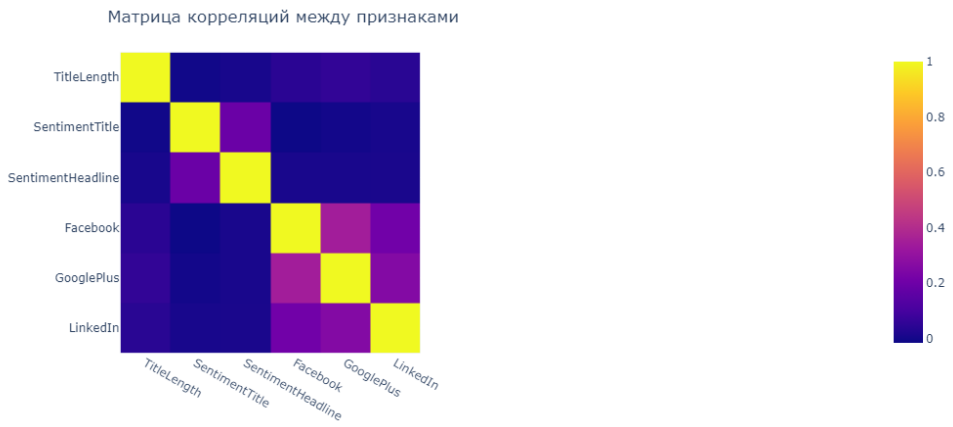

Figure 10: Корреляционная матрица

Также далее мы проиллюстрировали зависимости средней итоговой популярности и таких признаков, как SentimentTitle, SentimentHeadline, TitleLength (*Рисунок 11*). Сильной зависимости не вырисовывается, но заметно, что при определенных значениях признаков значительно большее количество новостей достигает более высоких значений популярности.

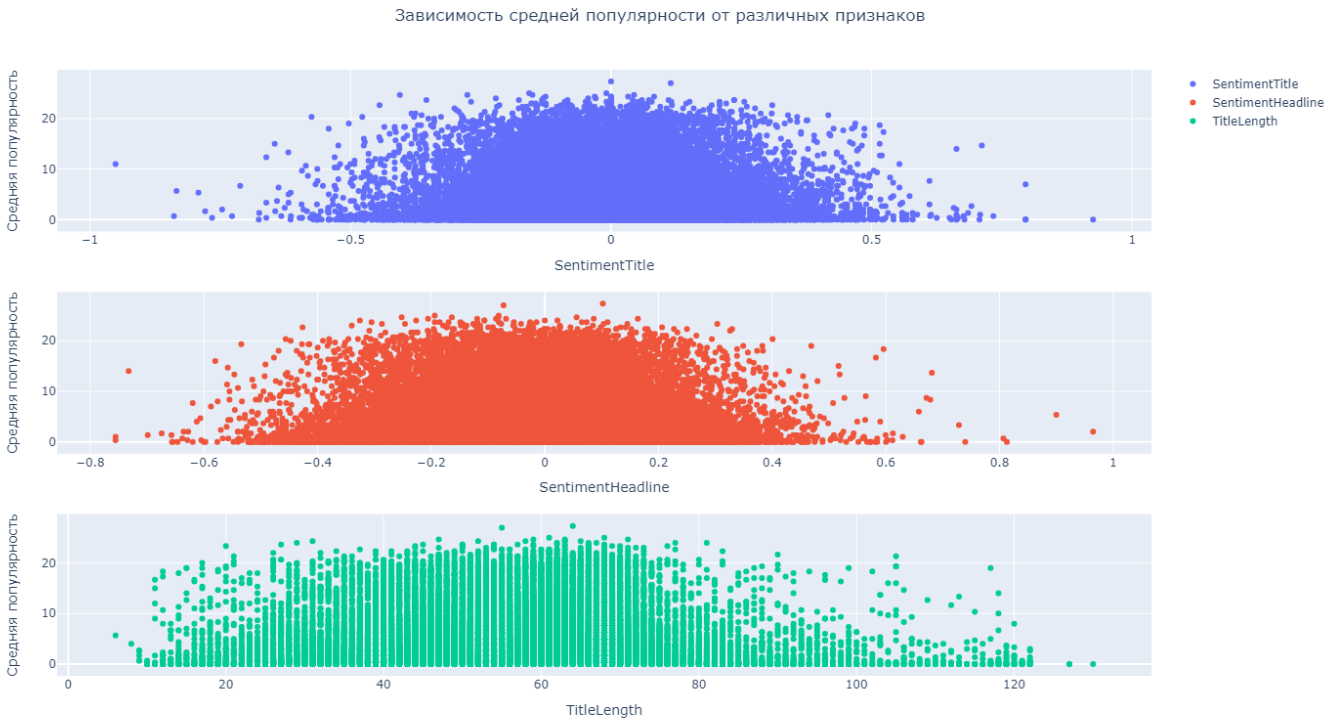

Figure 11: Графики зависимостей

Обратимся теперь к файлу *Facebook_ Economy*. Мы предположили, что существует прямая зависимость между популярностью новости в первые 12 часов и итоговой популярностью. То есть, если за первые 12 часов новость не набрала сильной популярности, то и дальше у нее не очень хорошие перспективы. И аналогично, если новость стала достаточно популярной в первые 12 часов, то в дальнейшем ее распространение будет более активным, а итоговая популярность – более высокой. Желаемая зависимость изображена на *Рисунке 12*.

Figure 12: Зависимость между популярностью спустя 12 часов и спустя 2 дня

5 Построение моделей

В качестве основологающей модели нашей работы мы выбрали модель линейной регрессии, так как она лучше всего подходит для прогнозирования данных с учетом зависимостей, имеющихся в нашем датасете.

На данном этапе работы мы неоднократно выполняли следующие шаги:

1. Выбор нужных признаков (один из которых обязательно зависимый)
2. Разделение выборки на тестовую и тренировочную, чтобы потом можно было оценить качество модели
3. Инициализация модели тренировочной выборкой
4. "Подгон" модели и получение результатов
5. Анализ полученной модели

5.1 Прогнозирование популярности по длине названия

В первую очередь мы решили попробовать прогнозировать итоговую популярность по количеству символов в названии новостной статьи. В итоге мы получили модель с коэффициентов детерминации $R^2 = 35.3\%$. Это означает, что 35.3% дисперсии зависимой переменной объясняется нашей моделью. Также заметим, что p-value коэффициента перед независимой переменной = 0.000, что говорит о его статистической значимости.

OLS Regression Results

Dep. Variable:	MeanPopularity	**R-squared (uncentered):**	0.353
Model:	OLS	**Adj. R-squared (uncentered):**	0.353
Method:	Least Squares	**F-statistic:**	1.983e+04
Date:	Mon, 06 Nov 2023	**Prob (F-statistic):**	0.00
Time:	11:22:15	**Log-Likelihood:**	-1.0736e+05
No. Observations:	36416	**AIC:**	2.147e+05
Df Residuals:	36415	**BIC:**	2.147e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Title	0.0610	0.000	140.813	0.000	0.060	0.062

Omnibus:	12109.040	**Durbin-Watson:**	2.003
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	33201.045
Skew:	1.813	**Prob(JB):**	0.00
Kurtosis:	5.956	**Cond. No.**	1.00

Figure 13: Саммари модели без константы

Далее мы сделали предикт значений по тестовой выборке и оценили точность модели с помощью метрик MAE (mean absolute error) и MSE (mean squared error). В итоге получили MAE ≈ 3.45, MSE ≈ 22.02.

Данная модель была построена без свободного члена, в силу чего мы получили неотцентрованный коэффициент детерминации. Мы также построили модель со свободным членым, результаты представлены на *Рисунке 14*. В данной ситуации мы имеем $R^2 \approx 1\%$, что является очень плохим показателем. Также нами были рассчитаны метрики: MAE ≈ 3.47, MSE ≈ 21.92. По данным метрикам данная модель не отличается от предыдущей.

OLS Regression Results

Dep. Variable:	MeanPopularity	**R-squared:**	0.009
Model:	OLS	**Adj. R-squared:**	0.009
Method:	Least Squares	**F-statistic:**	313.8
Date:	Mon, 06 Nov 2023	**Prob (F-statistic):**	6.31e-70
Time:	11:22:24	**Log-Likelihood:**	-1.0729e+05
No. Observations:	36416	**AIC:**	2.146e+05
Df Residuals:	36414	**BIC:**	2.146e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	1.3145	0.120	10.998	0.000	1.080	1.549
Title	0.0379	0.002	17.715	0.000	0.034	0.042

Omnibus:	12280.615	**Durbin-Watson:**	2.003
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	34027.216
Skew:	1.836	**Prob(JB):**	0.00
Kurtosis:	5.991	**Cond. No.**	277.

Figure 14: Саммари модели с константой

5.2 Прогнозирование популярности по эмоциональной окраске

Далее мы решили попробовать построить модель линейной регрессии для зависимого признака *MeanPopularity* и независимого признака *SentimentTitle*. Из графика зависимости средней популярности среди предложенных медиа от *SentimentTitle* было видно, что увеличение по модулю как отрицательных, так и положительных значений *SentimentTitle* у новости негативно сказывается на ее популярности. Также видно, что данное распределение имеет ось симметрии. Из этого был сделан вывод, что при построении модели линейной регрессии целесообразнее будет смотреть на значение *SentimentTitle* по модулю, тем самым взяв в качестве параметра именно степень эмоциональной окраски заголовка.

В результате построения модели мы получили неудовлетворительные значения коэффициента детерминации как в случае отсутствия константы, так и в случае ее наличия (*Рисунок 15-16*). Однако в последней ситуации, коэффициент перед независимым признаком оказался статистически незначимым, то есть такая модель вообще не будет релевантной. Что касается метрик: $\text{MAE}_1 \approx 3.36$, $\text{MSE}_1 \approx 27.44$; $\text{MAE}_2 \approx 3.46$, $\text{MSE}_2 \approx 21.48$.

OLS Regression Results

Dep. Variable:	MeanPopularity	**R-squared (uncentered):**	0.174
Model:	OLS	**Adj. R-squared (uncentered):**	0.174
Method:	Least Squares	**F-statistic:**	8379.
Date:	Wed, 08 Nov 2023	**Prob (F-statistic):**	0.00
Time:	13:58:45	**Log-Likelihood:**	-1.2235e+05
No. Observations:	39823	**AIC:**	2.447e+05
Df Residuals:	39822	**BIC:**	2.447e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
SentimentTitle	17.5555	0.192	91.539	0.000	17.180	17.931

Omnibus:	10218.591	**Durbin-Watson:**	1.790
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	23443.151
Skew:	1.460	**Prob(JB):**	0.00
Kurtosis:	5.366	**Cond. No.**	1.00

Figure 15: Саммари модели без константы

OLS Regression Results

Dep. Variable:	MeanPopularity	**R-squared:**	0.000
Model:	OLS	**Adj. R-squared:**	0.000
Method:	Least Squares	**F-statistic:**	1.874
Date:	Wed, 08 Nov 2023	**Prob (F-statistic):**	0.171
Time:	13:58:47	**Log-Likelihood:**	-1.1763e+05
No. Observations:	39823	**AIC:**	2.353e+05
Df Residuals:	39821	**BIC:**	2.353e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	3.3605	0.033	103.226	0.000	3.297	3.424
SentimentTitle	0.3265	0.238	1.369	0.171	-0.141	0.794

Omnibus:	13368.811	**Durbin-Watson:**	2.009
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	36741.870
Skew:	1.832	**Prob(JB):**	0.00
Kurtosis:	5.953	**Cond. No.**	10.4

Figure 16: Саммари модели с константой

Аналогичная модель была построена с использованием независимой переменной *SentimentHeadline* (*Рисунок 17-18*). $\text{MAE}_1 \approx 3.35$, $\text{MSE}_1 \approx 25.58$; $\text{MAE}_2 \approx 3.44$, $\text{MSE}_2 \approx 20.94$.

OLS Regression Results

Dep. Variable:	MeanPopularity	**R-squared (uncentered):**	0.206
Model:	OLS	**Adj. R-squared (uncentered):**	0.206
Method:	Least Squares	**F-statistic:**	1.033e+04
Date:	Wed, 08 Nov 2023	**Prob (F-statistic):**	0.00
Time:	13:58:52	**Log-Likelihood:**	-1.2179e+05
No. Observations:	39823	**AIC:**	2.436e+05
Df Residuals:	39822	**BIC:**	2.436e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
SentimentHeadline	18.2239	0.179	101.618	0.000	17.872	18.575

Omnibus:	10167.240	**Durbin-Watson:**	1.845
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	23045.517
Skew:	1.460	**Prob(JB):**	0.00
Kurtosis:	5.315	**Cond. No.**	1.00

Figure 17: Саммари модели без константы

5.3 Прогнозирование итоговой популярности по промежуточной

В рамках последней модели мы решили попробовать предсказывать значение итоговой популярности по уровню популярности на момент 12 часов с публикации. Аналогично предыдщим случаям сразу были построены модель с константой и без (*Рисунок 19-20*). В обоих случаях имеем неплохие коэффициенты детерминации: $R_1^2 = 63.4\%$, $R_2^2 = 55.8\%$, то есть обе модели объясняют более половины дисперсии зависимого признака. Для первой модели: $\text{MAE} \approx 5.52$, $\text{MSE} \approx 168.21$; для второй: $\text{MAE} \approx 6.67$, $\text{MSE} \approx 150.97$. Интересно, что модели уступают друг другу по одной из метрик.

OLS Regression Results

Dep. Variable:	MeanPopularity	**R-squared:**	0.000
Model:	OLS	**Adj. R-squared:**	-0.000
Method:	Least Squares	**F-statistic:**	0.0007070
Date:	Wed, 08 Nov 2023	**Prob (F-statistic):**	0.979
Time:	13:58:54	**Log-Likelihood:**	-1.1784e+05
No. Observations:	39823	**AIC:**	2.357e+05
Df Residuals:	39821	**BIC:**	2.357e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	3.4131	0.037	93.459	0.000	3.342	3.485
SentimentHeadline	0.0067	0.254	0.027	0.979	-0.491	0.504

Omnibus:	13310.929	**Durbin-Watson:**	1.996
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	36372.491
Skew:	1.827	**Prob(JB):**	0.00
Kurtosis:	5.928	**Cond. No.**	11.0

Figure 18: Саммари модели с константой

OLS Regression Results

Dep. Variable:	TS144	**R-squared (uncentered):**	0.634
Model:	OLS	**Adj. R-squared (uncentered):**	0.634
Method:	Least Squares	**F-statistic:**	3.284e+04
Date:	Sun, 05 Nov 2023	**Prob (F-statistic):**	0.00
Time:	16:04:20	**Log-Likelihood:**	-74671.
No. Observations:	18985	**AIC:**	1.493e+05
Df Residuals:	18984	**BIC:**	1.494e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
TS36	1.4635	0.008	181.205	0.000	1.448	1.479

Omnibus:	15466.426	**Durbin-Watson:**	1.837
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	373518.473
Skew:	3.880	**Prob(JB):**	0.00
Kurtosis:	23.297	**Cond. No.**	1.00

Figure 19: Саммари модели без константы

OLS Regression Results

Dep. Variable:	TS144	**R-squared:**	0.558
Model:	OLS	**Adj. R-squared:**	0.558
Method:	Least Squares	**F-statistic:**	2.399e+04
Date:	Sun, 05 Nov 2023	**Prob (F-statistic):**	0.00
Time:	16:07:42	**Log-Likelihood:**	-73643.
No. Observations:	18985	**AIC:**	1.473e+05
Df Residuals:	18983	**BIC:**	1.473e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.3484	0.093	46.598	0.000	4.166	4.531
TS36	1.3015	0.008	154.875	0.000	1.285	1.318

Omnibus:	15936.578	**Durbin-Watson:**	2.009
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	387131.140
Skew:	4.067	**Prob(JB):**	0.00
Kurtosis:	23.573	**Cond. No.**	12.2

Figure 20: Саммари модели с константой

6 Выводы

В ходе выполнения анализа мы пришли к следующим выводам:

6.1 Распределение данных

- Самая популярная среди предложенных тем - экономика, меньше всего данных доступно на тему Палестины
- В датасете довольна большая часть данных является выборосами, непригодными для построения моделей
- В колонках Headline и Source в таблице NewsFinal имеются пропуски в данных

6.2 Анализ зависимостей

- Имеется взаимосвязь между итоговыми популярностями одной и той же новости в разных медиа
- Между признаками SentimentHeadline и SentimentTitle имеется небольшая корреляция
- Высокая степень окраски информации в заголовке и хэдлайне новости как и в положительную, так и в отрицательную сторону негативно влияет на ее популярность
- Как правило, пользователи предпочитают новости с умеренной длиной заголовка – отклонение в большую и меньшую сторону ведет к снижению популярности
- Существует прямая зависимость между популярностью новости на тему экономики в Facebook и ее итоговой популярностью

6.3 Результат построения моделей

- Зависимость популярности новости от длины ее названия слабо аппроксимируется линейной функцией
- Зависимость популярности новости от ее эмоциональной окраски слабо аппроксимируется линейной функцией
- Построенная модель линейной регрессии для прогнозирования значения итоговой популярности по уровню популярности спустя 12 часов с момента публикаци дала достойный результат - данная зависимость может быть аппроксимирована линейной функцией

Вот и все ☺