



Министерство науки и высшего образования  
Российской Федерации Федеральное государственное  
бюджетное образовательное учреждение высшего образования  
«Московский государственный технический университет  
имени Н.Э. Баумана (национальный исследовательский  
университет)» (МГТУ им. Н.Э. Баумана)

---

ФАКУЛЬТЕТ \_\_\_\_\_ Информатика и системы управления и искусственный интеллект

КАФЕДРА \_\_\_\_\_ Системы обработки информации и управления

---

## Рубежный контроль №2

### По курсу

### «Методы машинного обучения»

### «Методы обработки текстов»

Студент

\_\_\_\_\_  
*подпись, дата*

**Костарев А. П.**

\_\_\_\_\_  
*фамилия, и .о.*

Преподаватель

\_\_\_\_\_  
*подпись, дата*

**Гапанюк Ю. Е.**

\_\_\_\_\_  
*фамилия, и .о.*

2024 г.

## **Задание**

Необходимо решить задачу классификации текстов на основе любого выбранного Вами датасета (кроме примера, который рассматривался в лекции). Классификация может быть бинарной или многоклассовой. Целевой признак из выбранного Вами датасета может иметь любой физический смысл, примером является задача анализа тональности текста.

Необходимо сформировать два варианта векторизации признаков - на основе CountVectorizer и на основе TfidfVectorizer.

Классификатор 1 (согласно варианту): LinearSVC

Классификатор 2 (согласно варианту): LogisticRegression

# Выполнение задания

✓  
2s

```
[1] from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
from sklearn.model_selection import train_test_split
from sklearn.svm import LinearSVC
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
import pandas as pd
import time
```

✓  
0s

```
[2] # Загрузка данных
df = pd.read_csv('Elon_musk_articles.csv')
```

✓  
0s

```
[3] df.head(10)
```

author	published_date	link	clean_url	excerpt	summary	rights	article_rank	topic	country	language	authors	media	twitter_account	article_score
i Verge	30-09-2022 00:30	https://www.newsbreak.com/news/2766864647685h...	newsbreak.com	Elon Musk's deposition for the Twitter v. Musk...	#Linux Elon Musk#Linux Business#Business Leads...	newsbreak.com	4828	news	PH	en	Ad Robertson,The Verge	https://img.particlenews.com/img/id/2727hy_OG...	@newsbreak	23.075157
ia Sori	01-10-2022 06:46	https://www.hindustantimes.com/world-news/elon-musk-in-an-event-devoted-to-robots-and-t...	hindustantimes.com	Elon Musk in an event devoted to robots and Twitter...	Published on Oct 01, 2022 11:16 AM ISTYin vL...	hindustantimes.com	980	world	IN	en	Malika Sori	https://images.hindustantimes.com/img/2022/10/...	@tweets	23.064362
ia Sori	01-10-2022 02:20	https://www.hindustantimes.com/world-news/elon-musk-tweets-trust-elon-musk-and-tweetar...	hindustantimes.com	Elon Musk Tweets Trust Elon Musk and Tweetar...	Messages between Tesla chief Elon Musk and imp...	hindustantimes.com	980	news	IN	en	Malika Sori	https://images.hindustantimes.com/img/2022/10/...	@tweets	22.901016
ne Tag	02-10-2022 17:26	https://www.thething.com/elon-musks-mom-maye-...	thething.com	Elon Musk learned that being a billionaire doe...	Everyone knows Elon Musk as a pioneer, a busin...	thething.com	14062	news	US	en	Nadine Tag	https://static.thethingimages.com/wordpress/...	@thethingcom	22.876207
K. Bel	30-09-2022 00:02	https://www.newsbreak.com/news/2766803421377ie...	newsbreak.com	A branch of Elon Musk's private messages have...	#Linux Business#Business Leadership#Linux Comp...	newsbreak.com	4828	news	PH	en	K. Bel	https://img.particlenews.com/img/id/2QXVPM_OG...	@newsbreak	22.777958
9i Bal	30-09-2022 21:06	https://tchistory.in/chat-between-elon-musk-an...	tchistory.in	According to recent reports, chats between Eo...	According to recent reports, chats between Eo...	tchistory.in	23310	news	IN	en	Adil Bal	https://tchistory.in/wp-content/uploads/2022/0...	tchistoryin	22.714160
d Zare	30-09-2022 11:26	https://www.newsbreak.com/news/276741376563h...	newsbreak.com	According to a series of text messages that we...	#Internal Communications#Elcom Valley#Busine...	newsbreak.com	4828	news	PH	en	Mehdi Zare	https://img.particlenews.com/img/id/2eUa_OG...	@newsbreak	22.673359
NaN	01-10-2022 04:26	https://www.indiatoday.in/technology/story/te...	indiatoday.in	Hundreds of text messages between Elon Musk an...	Text messages between Elon Musk and key figure...	indiatoday.in	1817	news	IN	en	NaN	https://akm-img-a-in.tosshub.com/indiatoday/im...	@indiatoday	22.672602
Sandali agina	30-09-2022 11:18	https://www.newsbreak.com/news/2767432963421ie...	newsbreak.com	"I think a new social media company is needed..."	Terms of Use#Privacy Policy#Do Not Sell My Info#...	newsbreak.com	4828	news	PH	en	Sandali Handigama	https://img.particlenews.com/img/id/2eUa_OG...	@newsbreak	22.655766
Ti-City Herald	30-09-2022 17:13	https://www.newsbreak.com/news/276765928936ie...	newsbreak.com	Streaming powerhouse Netflix once had the mark...	Elon Musk on Tuesday revised his offer to buy ...	newsbreak.com	4828	news	PH	en	Newsbreak, Ti-City Herald	https://img.particlenews.com/img/id/0RfYt_OG...	@newsbreak	22.566383

✓  
0s

[4] df.info()

```
⇒ <class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 17 columns):
#   Column                Non-Null Count  Dtype
---  -
0   _id                    5000 non-null   object
1   title                  5000 non-null   object
2   author                 4438 non-null   object
3   published_date         5000 non-null   object
4   link                   5000 non-null   object
5   clean_url              5000 non-null   object
6   excerpt                4962 non-null   object
7   summary                4986 non-null   object
8   rights                 4966 non-null   object
9   article_rank           5000 non-null   int64
10  topic                  5000 non-null   object
11  country                5000 non-null   object
12  language               5000 non-null   object
13  authors                4438 non-null   object
14  media                  4962 non-null   object
15  twitter_account        3538 non-null   object
16  article_score          5000 non-null   float64
dtypes: float64(1), int64(1), object(15)
memory usage: 664.2+ KB
```

✓  
0s

[5] # проверим пропуски в данных и устраним их  
na\_mask = df.isna()  
na\_counts = na\_mask.sum()  
na\_counts

```
⇒ _id                    0
   title                  0
   author                 562
   published_date         0
   link                   0
   clean_url              0
   excerpt                38
   summary                14
   rights                 34
   article_rank           0
   topic                  0
   country                0
   language               0
   authors                562
   media                  38
   twitter_account        1462
   article_score          0
dtype: int64
```

```
✓ [6] df.dropna(inplace=True)
0s na_mask = df.isna()
na_counts = na_mask.sum()
na_counts
```

```
⇒ _id 0
title 0
author 0
published_date 0
link 0
clean_url 0
excerpt 0
summary 0
rights 0
article_rank 0
topic 0
country 0
language 0
authors 0
media 0
twitter_account 0
article_score 0
dtype: int64
```

```
✓ [7] # Разделим набор данных на обучающую и тестовую выборки
0s X, Y = df['summary'], df['topic']
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.2, random_state=42)

time_arr = []
```

```
✓ [8] # векторизация признаков с помощью CountVectorizer
0s count_vect = CountVectorizer()
X_train_counts = count_vect.fit_transform(X_train)
X_test_counts = count_vect.transform(X_test)
```

```
✓ [9] # векторизация признаков с помощью TfidfVectorizer
0s tfidf_vect = TfidfVectorizer()
X_train_tfidf = tfidf_vect.fit_transform(X_train)
X_test_tfidf = tfidf_vect.transform(X_test)
```

```
✓ [11] # Произведем обучения двух классификаторов (по варианту) для CountVectorizer
1m
# LinearSVC
gbc = LinearSVC()
start_time = time.time()
gbc.fit(X_train_counts, y_train)
train_time = time.time() - start_time
time_arr.append(train_time)
pred_gbc_counts = gbc.predict(X_test_counts)
print("Точность (CountVectorizer + LinearSVC):", accuracy_score(y_test, pred_gbc_counts))

# Logistic Regression
lr = LogisticRegression(max_iter=2000)
start_time = time.time()
lr.fit(X_train_counts, y_train)
train_time = time.time() - start_time
time_arr.append(train_time)
pred_lr_counts = lr.predict(X_test_counts)
print("Точность (CountVectorizer + LogisticRegression):", accuracy_score(y_test, pred_lr_counts))

⇒ /usr/local/lib/python3.10/dist-packages/sklearn/svm/_base.py:1244: ConvergenceWarning: Liblinear failed to converge, increase the number of iterations.
warnings.warn(
Точность (CountVectorizer + LinearSVC): 0.7315541601255887
Точность (CountVectorizer + LogisticRegression): 0.7770800627943485
```


```

11s [13] # Произведем обучения двух классификаторов (по варианту) для TfidfVectorizer

# LinearSVC
gbc = LinearSVC()
start_time = time.time()
gbc.fit(X_train_tfidf, y_train)
train_time = time.time() - start_time
time_arr.append(train_time)
pred_gbc_tfidf = gbc.predict(X_test_tfidf)
print("Точность (TfidfVectorizer + LinearSVC):", accuracy_score(y_test, pred_gbc_tfidf))

# Logistic Regression
lr = LogisticRegression(max_iter=2000)
start_time = time.time()
lr.fit(X_train_tfidf, y_train)
train_time = time.time() - start_time
time_arr.append(train_time)
pred_lr_tfidf = lr.predict(X_test_tfidf)
print("Точность (TfidfVectorizer + LogisticRegression):", accuracy_score(y_test, pred_lr_tfidf))

```

 Точность (TfidfVectorizer + LinearSVC): 0.8037676609105181  
 Точность (TfidfVectorizer + LogisticRegression): 0.8084772370486656

```


0s [15] from tabulate import tabulate

data = [
    ["(CountVectorizer + LogisticRegression)", accuracy_score(y_test, pred_lr_counts), time_arr[0]],
    ["(CountVectorizer + LinearSVC)", accuracy_score(y_test, pred_gbc_counts), time_arr[1]],
    ["(TfidfVectorizer + LogisticRegression)", accuracy_score(y_test, pred_lr_tfidf), time_arr[2]],
    ["(TfidfVectorizer + LinearSVC)", accuracy_score(y_test, pred_gbc_tfidf), time_arr[3]]
]

sorted_data = sorted(data, key=lambda x: x[1], reverse=True)

# Вывод отсортированных данных в виде таблицы
print(tabulate(sorted_data, ['Связка', 'Точность валидации', 'Время обучения'], tablefmt="grid"))

```



Связка	Точность валидации	Время обучения
(TfidfVectorizer + LogisticRegression)	0.808477	1.77126
(TfidfVectorizer + LinearSVC)	0.803768	84.5727
(CountVectorizer + LogisticRegression)	0.77708	4.12995
(CountVectorizer + LinearSVC)	0.731554	52.7793