# Модерирование онлайн-дискуссий

Волкова Юлия Михайловна

Руководитель: Кузнецов Антон Михайлович,
старший преподаватель, департимент информатики

Консультант: Аленичева Алиса Романовна,
исследователь, JetBrains Research

Высшая Школа Экономики
Санкт-Петербург

Июнь, 2021

# Содержание

**Мотивация**

**Обзор литературы**

**Цель и задачи**

**Этапы работы**

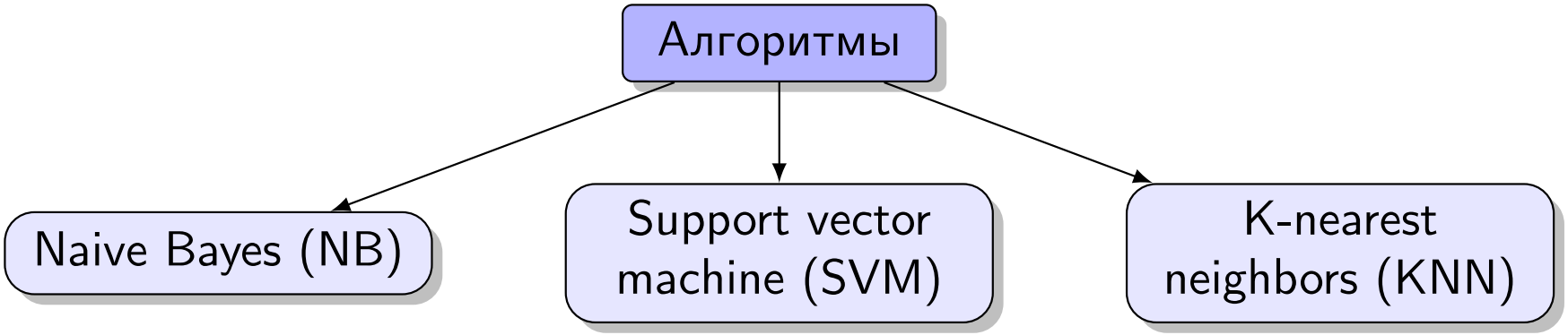
**Заключение**

**Список литературы**

# Мотивация

- По мере того как онлайн-платформа развивается и на нее приходит все больше и больше пользователей, у основателя со временем возникает потребность в контролировании этого процесса. Возникает необходимость в модераторах, в людях которые будут поддерживать порядок в сообществе, создавать комфортную атмосферу для присутствующих (Панов М.А., 2014)
- Модерирование платформы напрямую связано с Natural Language Processing (NLP, Обработка естественного языка)
- Модерирование онлайн-платформы → задача классификации текста, анализ тональности текста на российских данных

# Задача классификации текста (Kadhim A. I., 2019)




Дополнительные подходы:

- Logistic regression (LR)
- Random forests (RF)

# Задача классификации текста

| | Technique | | |
|---|---|---|---|
| | KNN | NB | SVM |
| (Tilve A. el al, 2017) | 0.8597 | **0.8843** | 0.8628 |
| (Tilve A. el al, 2017) | **0.7512** | 0.4590 | 0.7011 |
| (Kadhim A. I., 2019) | 70.85 | 79.7 | **81.35** |

Таблица 1: Результаты исследований

Разные методы работают по-разному в зависимости от набора данных, однако чаще всего хорошо себя показывает:

- KNN (Kadhim A. I., 2019)
- SVM (Romanov A. el al., 2019)

# Feature weighting

## Мешок слов (BoW)

текст представляется в виде мешка (мультимножества) его слов без какого-либо учета грамматики и порядка слов, но с сохранением информации об их количестве

## TF-IDF (Jones K. S., 1972)

$$TF(t,d) = \frac{n_t}{\sum n_k},$$

где $n_t$ - число вхождений слова $t$ в документ, знаменатель - общее число слов в данном документе

$$IDF(t,D) = log(\frac{|D|}{|d_i \in D | t \in d_i|}),$$

где $|D|$ -число документов в коллекции, $|d_i \in D | t \in d_i|$ - число документов из коллекции $D$, в которых встречается $t$ (когда $n_t \neq 0$)

# Цель и задачи

## Цель

Взять оживленную онлайн-платформу на русском языке, применить алгоритмы NLP с целью дополнения информации о контенте платформы

# Задачи

- Выбрать онлайн-платформу с русским контентом
- Изучить структуру сайта и достать необходимую информацию для дальнейшего анализа
- Обработать полученный текст и попробовать несколько методов подготовки данных (BoW VS TF-IDF)
- Провести эксперименты с применением различных методов классификации текстов на основе разной предобработки текста ({NB, LR, SVM, KNN, RF} & {BoW, TF-IDF})
  **Гипотеза**: Лучшим классификатором будет KNN или SVM на основе TF-IDF
- Выбор лучшего классификатора, проведение стат.теста на статистическую значимость результатов
- Применить ансамбль методов для более точного предсказания
- Изучить обученную модель для анализа тональности комментариев
- Применить модель для определения эмоционально окраски комментариев и типа пользователей на основе окраски их комментариев
- Сформировать некоторую форму, которая выдавала бы информацию о категории поста, эмоциональной окраске комментариев под постом с учетом "типа"пользователей

# Сбор данных

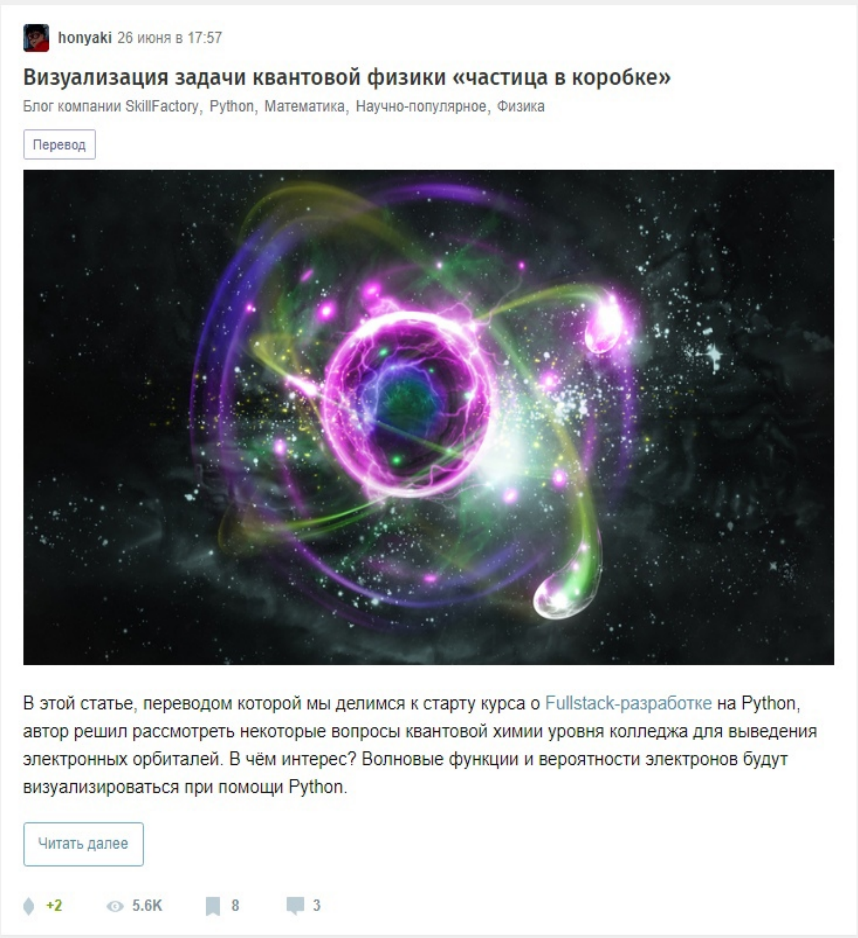



Рис. 1: Информация о посте на сайте Хабр

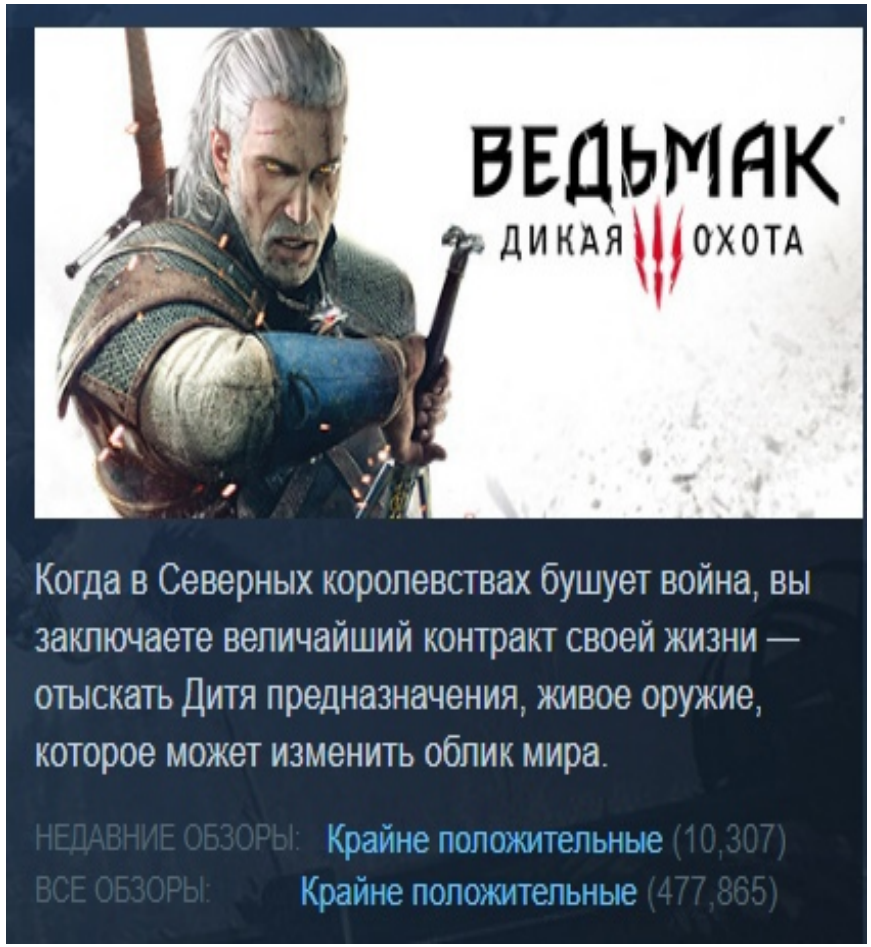
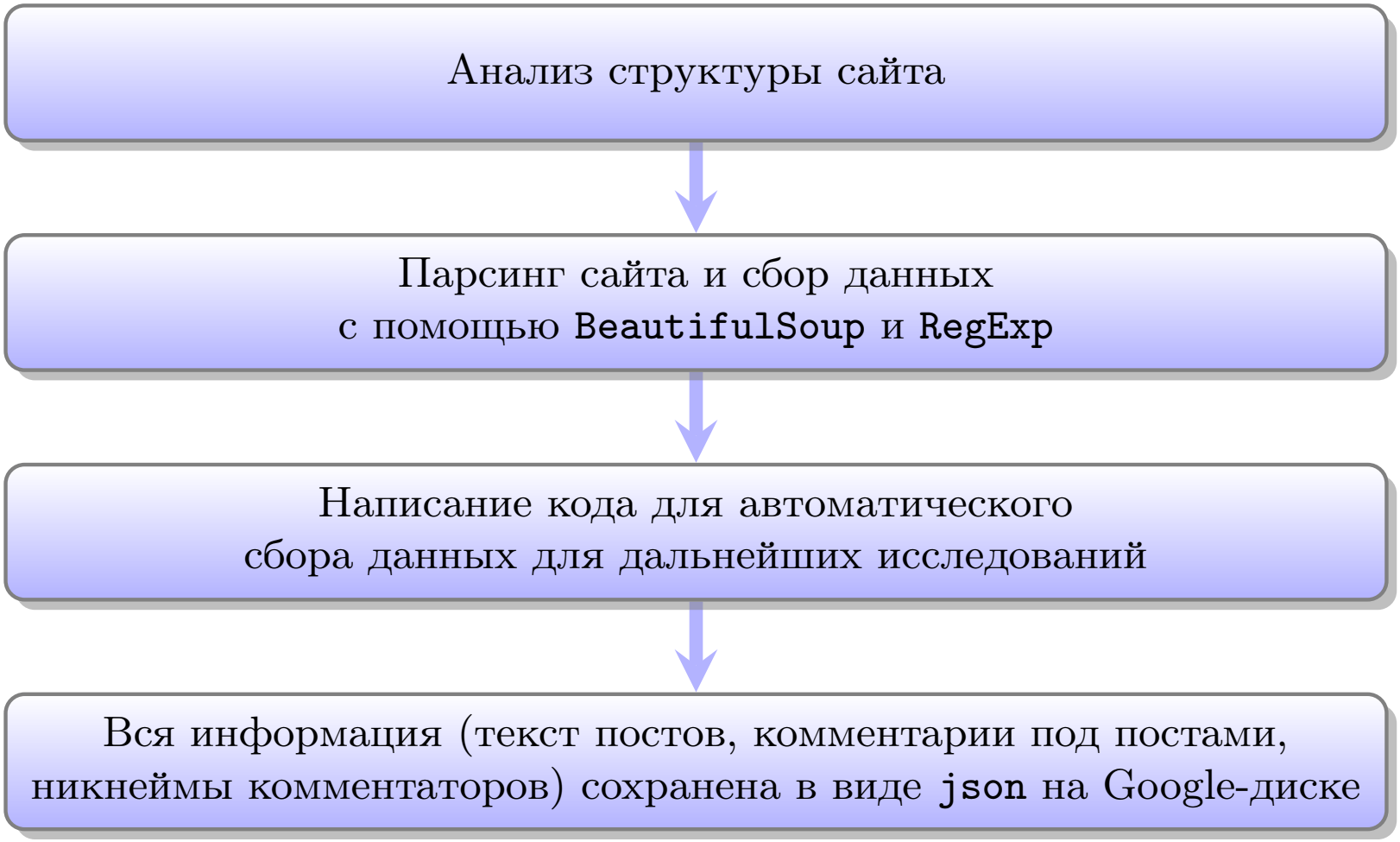
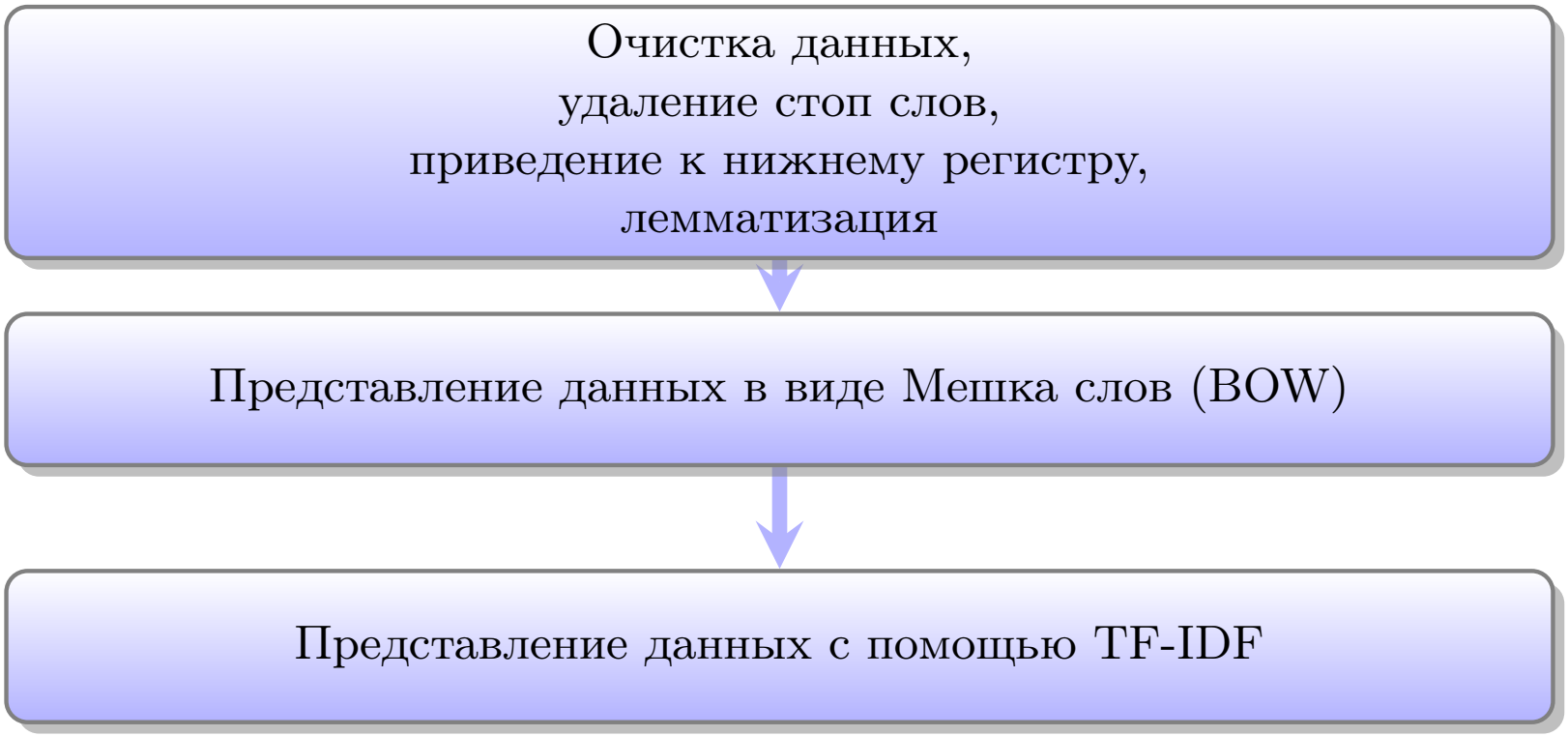



Рис. 2: Информация о тональности комментариев (Steam)

# Сбор данных

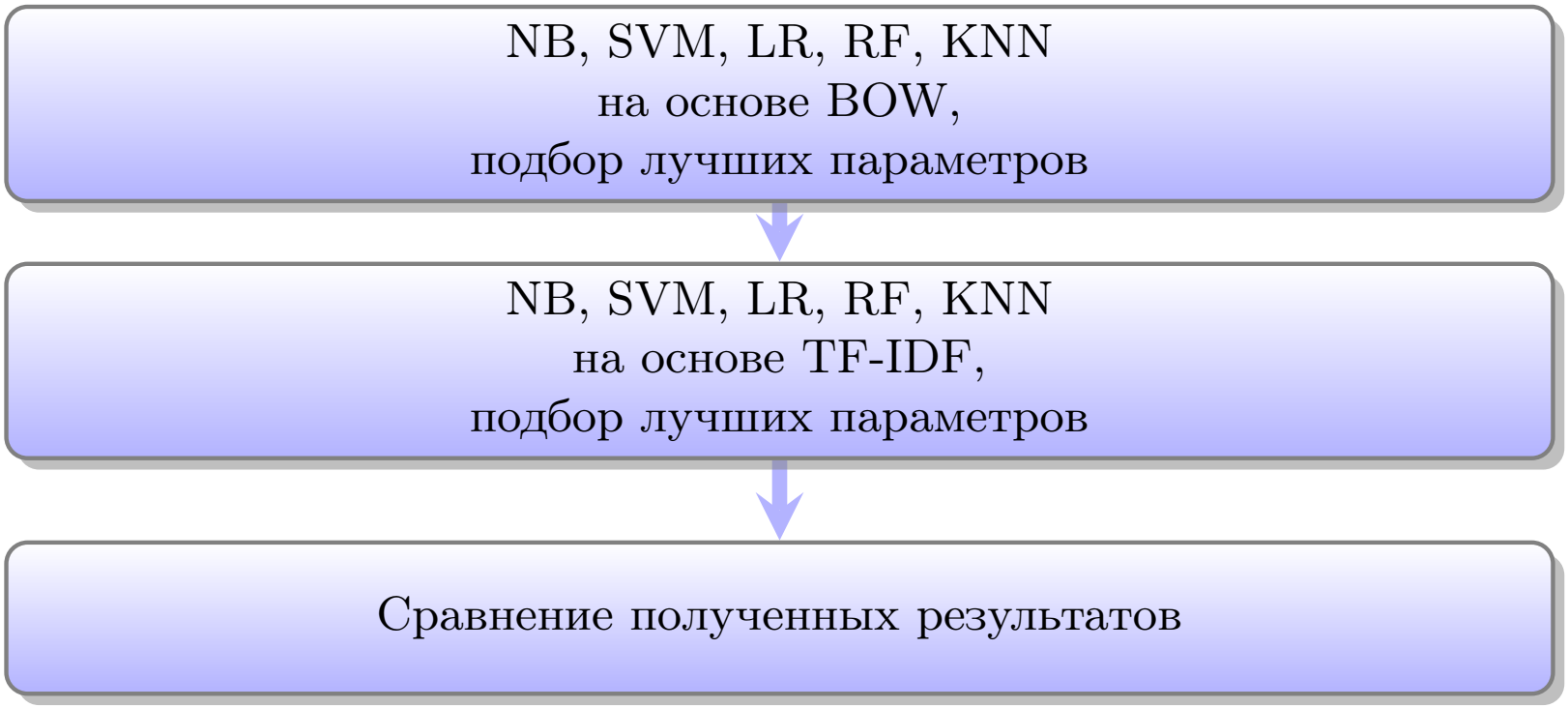

Анализ структуры сайта
Парсинг сайта и сбор данных
с помощью BeautifulSoup и RegExp
Написание кода для автоматического
сбора данных для дальнейших исследований
Вся информация (текст постов, комментарии под постами,
никнеймы комментаторов) сохранена в виде json на Google-диске

# Обработка данных



# Задача классификации текста

Для задачи классификации были взяты посты из пяти категорий: Научно-популярное, Программирование,Математика, Разработка игр, Здоровье
Количество постов в тренировочном датасете: **9874**
Количество постов в тестовом датасете: **4254**



| Метод | Классы | Precision | Recall | F1 | Accuracy |
|---|---|---|---|---|---|
| Naive Bayes (BOW + лемматизация, без стоп-слов, voc_limit = 1000) | Научно-популярное | 0.74 | 0.65 | 0.69 | |
| | Программирование | 0.80 | 0.73 | 0.76 | |
| | Математика | 0.76 | 0.72 | 0.73 | 0.76 |
| | Разработка игр | 0.88 | 0.80 | 0.83 | |
| | Здоровье | 0.66 | 0.92 | 0.77 | |
| Naive Bayes (TFIDF + лемматизация, без стоп-слов, ngram_range(1, 2)) | Научно-популярное | 0.78 | 0.72 | 0.75 | |
| | Программирование | 0.80 | 0.85 | 0.82 | |
| | Математика | 0.81 | 0.79 | 0.80 | **0.81** |
| | Разработка игр | 0.89 | 0.85 | 0.87 | |
| | Здоровье | 0.76 | 0.84 | 0.80 | |
| Logistic Regression (BOW (voc_limit = 1500) + лемматизация, без стоп-слов) | Научно-популярное | 0.74 | 0.69 | 0.72 | |
| | Программирование | 0.66 | 0.68 | 0.67 | |
| | Математика | 0.77 | 0.78 | 0.78 | 0.75 |
| | Разработка игр | 0.84 | 0.86 | 0.85 | |
| | Здоровье | 0.71 | 0.72 | 0.71 | |
| Logistic Regression (TFIDF + ngram_range(1, 2)) | Научно-популярное | 0.93 | 0.86 | 0.89 | |
| | Программирование | 0.78 | 0.77 | 0.77 | |
| | Математика | 0.82 | 0.87 | 0.84 | **0.81** |
| | Разработка игр | 0.77 | 0.82 | 0.79 | |
| | Здоровье | 0.73 | 0.71 | 0.72 | |
| Random Forest (BOW (voc_limit = 1500) + лемматизация, без стоп-слов) | Научно-популярное | 0.78 | 0.75 | 0.76 | |
| | Программирование | 0.88 | 0.90 | 0.89 | |
| | Математика | 0.75 | 0.72 | 0.73 | 0.79 |
| | Разработка игр | 0.82 | 0.84 | 0.83 | |
| | Здоровье | 0.72 | 0.73 | 0.72 | |
| SVM (линейный) (BOW (voc_limit = 1500) + лемматизация, без стоп-слов) | Научно-популярное | 0.67 | 0.63 | 0.65 | |
| | Программирование | 0.62 | 0.61 | 0.61 | |
| | Математика | 0.71 | 0.70 | 0.71 | 0.68 |
| | Разработка игр | 0.78 | 0.84 | 0.81 | |
| | Здоровье | 0.63 | 0.63 | 0.63 | |
| SVM (линейный) (TFIDF + ngram_range(1, 2)) | Научно-популярное | 0.79 | 0.82 | 0.81 | |
| | Программирование | 0.77 | 0.73 | 0.75 | |
| | Математика | 0.91 | 0.89 | 0.90 | **0.82** |
| | Разработка игр | 0.82 | 0.88 | 0.85 | |
| | Здоровье | 0.80 | 0.77 | 0.79 | |
| KNN (neighbors = 5) (BOW) | Научно-популярное | 0/58 | 0.66 | 0.62 | |
| | Программирование | 0.75 | 0.77 | 0.76 | |
| | Математика | 0.56 | 0.79 | 0.65 | 0.66 |
| | Разработка игр | 0.75 | 0.61 | 0.67 | |
| | Здоровье | 0.73 | 0.48 | 0.58 | |
| KNN (neighbors = 200) (TF-IDF) | Научно-популярное | 0.78 | 0.83 | 0.81 | |
| | Программирование | 0.91 | 0.58 | 0.70 | |
| | Математика | 0.76 | 0.80 | 0.78 | **0.79** |
| | Разработка игр | 0.72 | 0.91 | 0.80 | |
| | Здоровье | 0.83 | 0.84 | 0.84 | |

Таблица 2: Результаты экспериментов

Baseline: Naive Bayes на TF-IDF
Метод сравнения: 5x2cv paired t-test

$H_0$: Модель №1 и Модель №2 имеют одинаковую точность (статистически незначимую)
$H_1$: Модель №1 и Модель №2 показывают статистически значимо разные результаты

| Model №1 | Model №2 | t-stat | p-value | reject ? (5%) | reject? (10%) |
|---|---|---|---|---|---|
| NB | LR | -2.351 | 0.060 | no | yes |
| NB | SVM | -4.569 | 0.006 | yes | yes |
| NB | KNN | 2.261 | 0.073 | no | yes |

Таблица 3: Статистическая значимость между результатми методов на TF-IDF

# Ансамбль методов

Модели, которые вошли в ансамбль:

- MultinomialNB(alpha = 0.01)
- SGDClassifier(max_iter = 2000, tol = 1e-5, loss = 'log')
- LinearSVC()
- KNeighborsClassifier(n_neighbors = 200)

Метод: MaxVoting

| | **precision** | **recall** | **f1-score** | **support** |
|---|---|---|---|---|
| Научно-популярное | 0.80 | 0.85 | 0.82 | 789 |
| Программирование | 0.78 | 0.75 | 0.77 | 768 |
| Здоровье | 0.78 | 0.75 | 0.77 | 896 |
| Математика | 0.81 | 0.85 | 0.83 | 895 |
| Разработка игр | 0.90 | 0.88 | 0.89 | 899 |
| **accuracy** | | | **0.82** | 4247 |
| macro avg | 0.82 | 0.82 | 0.82 | 4247 |
| weighted avg | 0.82 | 0.82 | 0.82 | 4247 |

Таблица 4: Точность ансамбля методов

# Результат работы ансамбля методов

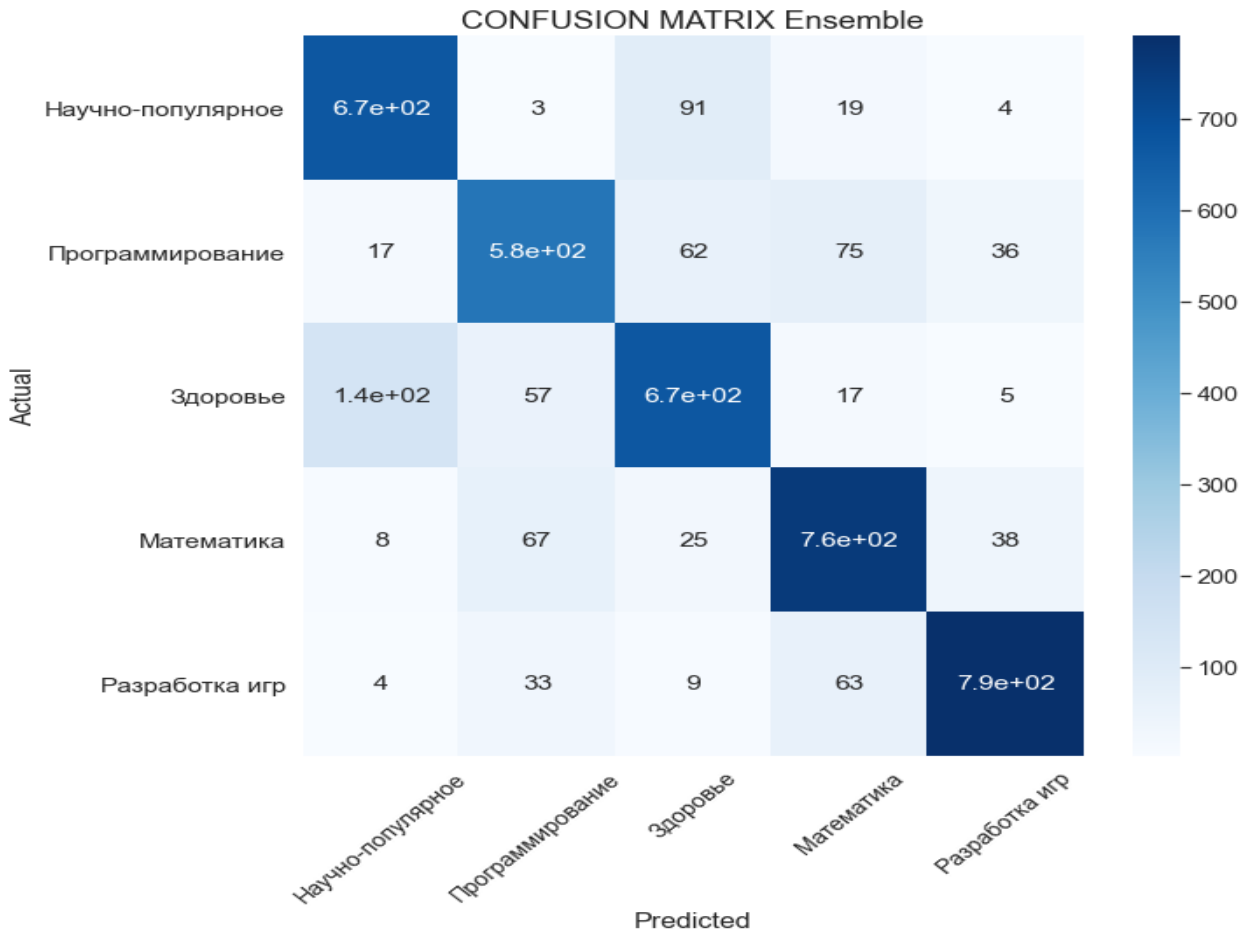



Рис. 3: Confusion matrix

# Анализ тональности комментариев

Для русского языка есть очень удобная библиотека на языке Python — Dostoevsky
Модель была обучена на наборе данных `RuSentiment`),
0.71 F1 score
Библиотека классифицирует текст на 5 категорий:

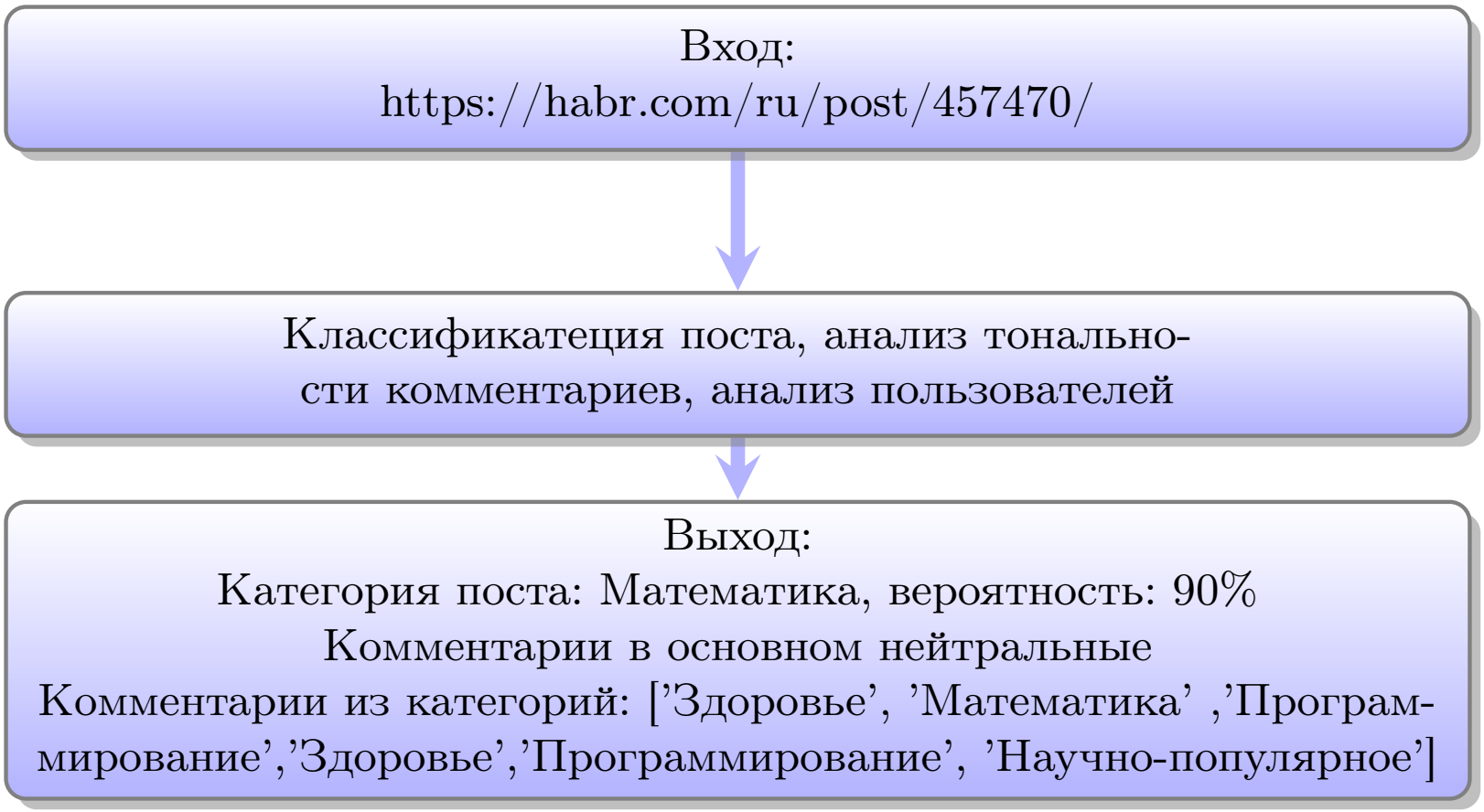
- Негативное настроение;
- Позитивное настроение;
- Позитивное настроение;
- Нейтральное поведение;
- Речевой акт (формальные поздравления, благодарственные и поздравительные посты);
- Класс «пропустить» для неясных случаев.

# Эмоциональная окраска комментариев

| | | Тип комментатора | | |
|---|---|---|---|---|
| | | Negative | Positive | Neutral |
| Тип комментария | Negative | -1 | -1.5 | -1 |
| | Positive | +1.5 | +1 | +1 |
| | Neutral | 0 | 0 | 0 |

Таблица 5: Учет типа пользователя в оценке тональности комментариев под постом

# Пример


Вход:
https://habr.com/ru/post/457470/
Классификатеция поста, анализ тональности комментариев, анализ пользователей
Выход:
Категория поста: Математика, вероятность: 90%
Комментарии в основном нейтральные
Комментарии из категорий: ['Здоровье', 'Математика' ,'Программирование','Здоровье','Программирование', 'Научно-популярное']

# Заключение

- Подтвердилась гипотеза о лучшем классификаторе (SVM + TF-IDF)
- Дополнена информация о постах онлайн-платформы Хабр с целью предоставить более полную информацию пользователям о контенте платформы

Ссылка на GIT:
https://github.com/YUMVOLKOVA/Online-Discussions/tree/main/Habr

# Список литературы


- Панов М. А. Управление интернет-ресурсами: проблемы современного модерирования //Проблемы современной экономики (Новосибирск). – 2014. – №. 22-2.
- Kadhim A. I. Survey on supervised machine learning techniques for automatic text classification //Artificial Intelligence Review. – 2019. – Т. 52. – №. 1. – С. 273-292.
- Tilve A. K. S., Jain S. N. A survey on machine learning techniques for text classification //Int. J. Eng. Sci. Res. Technol. – 2017. – Т. 3. – С. 513-520.
- Romanov A., Lomotin K., Kozlova E. Application of natural language processing algorithms to the task of automatic classification of Russian scientific texts. Data Sci. J. 18 (37), 1–17. – 2019.
- Jones K. S. A statistical interpretation of term specificity and its application in retrieval //Journal of documentation. – 1972.
- https://github.com/bureaucratic-labs/dostoevsky

# Тип комментаторов

| commentator | positive | mean | negative | mean | neutral | mean | Score | main_class |
|---|---|---|---|---|---|---|---|---|
| Fedcomp | 0 | 0.057384 | 4 | 0.501624 | 1 | 0.394879 | -1.611618 | negative |
| mentin | 0 | 0.082607 | 2 | 0.503938 | 1 | 0.375210 | -0.632667 | negative |
| xBrowser | 0 | 0.067713 | 2 | 0.515596 | 0 | 0.339574 | -1.031191 | negative |
| maldalik | 0 | 0.038255 | 2 | 0.511791 | 1 | 0.512958 | -0.510625 | negative |
| AlexArt84 | 0 | 0.074161 | 2 | 0.552921 | 0 | 0.407887 | -1.105841 | negative |

Таблица 6: Информация о пользователях