

Разработка механизма ранжирования комментариев к постам на основе методов машинного обучения для соцсети ВКонтакте

Команда Team Target

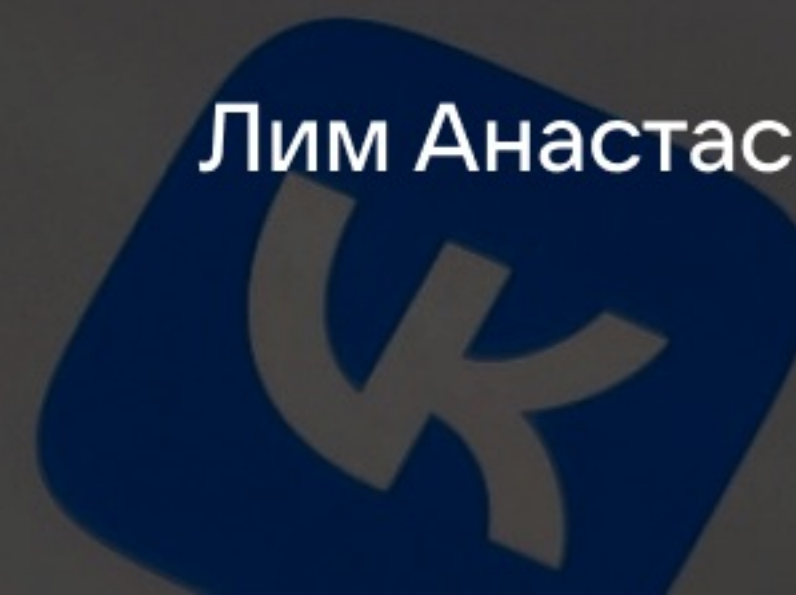
Усенко Ксения

Смирнов Илья

Карпунь Дарья

Кудряшов Никита

Лим Анастасия



Предложенный нами механизм ранжирования комментариев на основе методов машинного обучения для ВКонтакте с точностью не менее 80% позволит корректно выделить топовые комментарии, что сможет повысить ценность контента и улучшить пользовательский опыт к концу 2023.

Контекст

на данный момент у ВКонтакте существует механизм ранжирования комментариев, однако он не всегда работает корректно, пропуская в топ комментарии плохого качества, которые ухудшают общее восприятие поста и сайта в целом

★

79% людей замечало нерелевантные комментарии в топе

★

опубликовано более 6,3 млрд единиц контента


★

на 30% повысился охват лент

★

просмотрено более 251 млрд метров ленты

Инициативы



выделение уникальных критериев для повышения точности ранжирования комментариев:

1. число символов

2. пунктуация


3. тональность

4. связь текста и поста

5. сарказм

6. сленг

7. ошибки в тексте



механизм взаимодействия с комментаторами:

геймификация процесса
вытеснение из топов исключительно негативных комментариев

Результаты

	число символов	
пунктуация	①	
②		тональность
		③

Топ факторов, влияющих на ранжирование

Метрика NDCG

RankNet: 88,5%



Executive summary

Анализ проблемы

Анализ данных

Реализация

Результат

Наша команда

Комментарии повышают ценность контента и влияют на его восприятие. Улучшение механизмов ранжирования возможно через понимание предпочтений пользователей

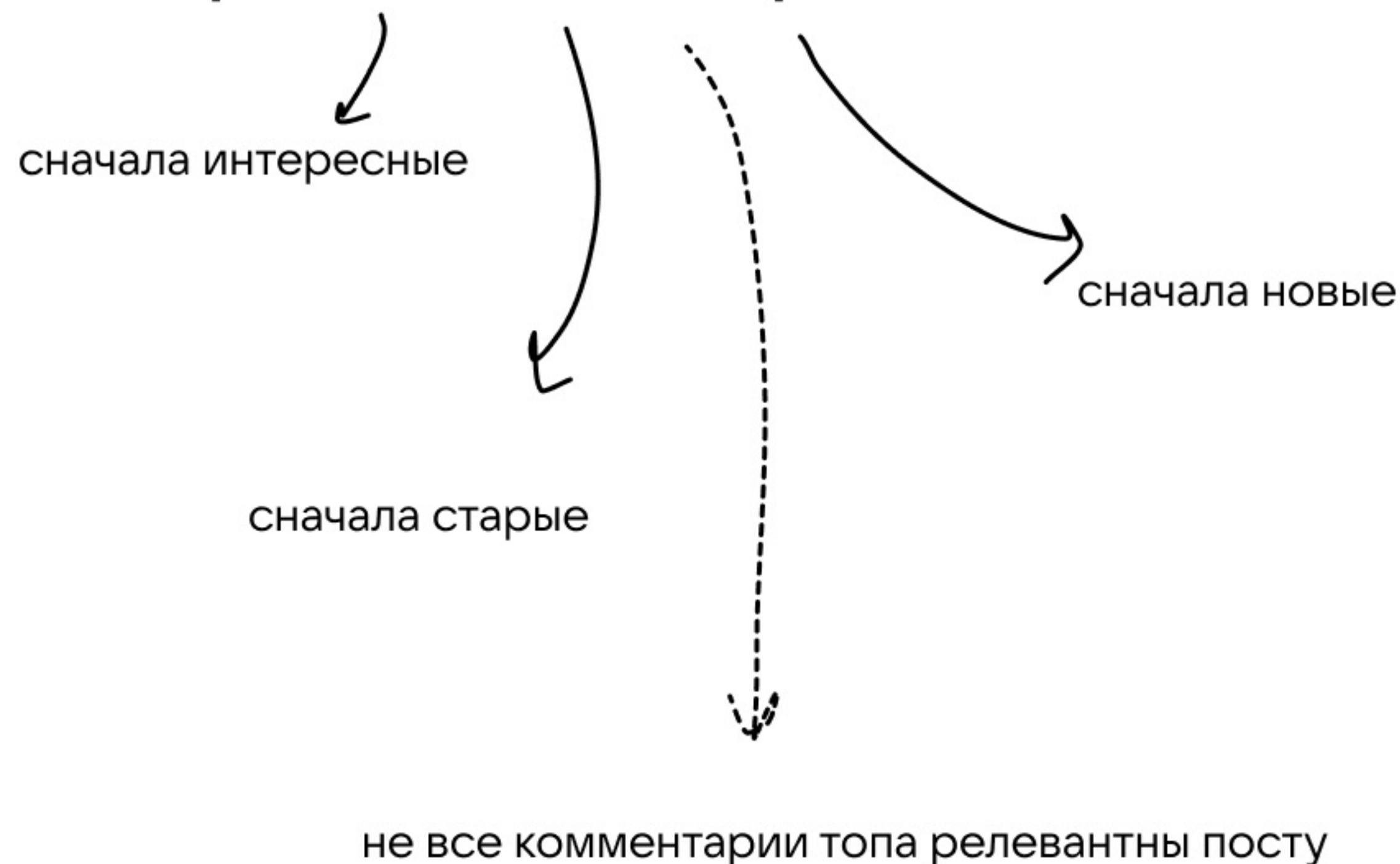
За 2022 год:

на **30%** повысился охват лент

опубликовано более **6,3** млрд единиц контента

совокупный доход сообществ VK **4** млрд рублей

Ранжирование комментариев сейчас





Выделение социальных групп, общественных тенденций, паттернов поведения и реакций помогают найти ключевые признаки при первичном анализе проблемы ранжирования

Анализ ЦА

Отношение пользователей к комментариям под постами в ВКонтакте



57% пользователей активно читает комментарии к постам в ВКонтакте



77% пользователей замечало нерелевантные комментарии в топе

Какие комментарии получают наиболее высокую оценку пользователей



62% высоко оценит в первую очередь саркастический комментарий



39% наиболее склонно оценить позитивный комментарий под постом



44,5% предпочитает читать согласие / не согласие с мнением автора поста

Общественное поведение постоянно меняется, влияя на то, как люди выражают свои мысли и эмоции в комментариях



Изменение приоритетов в комментариях — нормальное явление, отражающее постоянную эволюцию общественных ценностей и взглядов пользователей

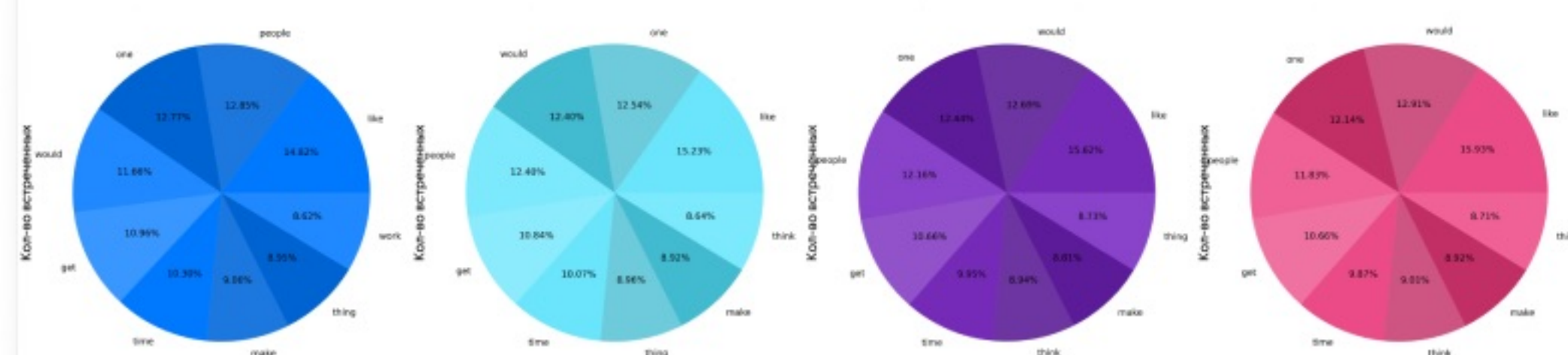


Задача ранжирования не имеет единого решения, так как ее результат зависит от множества факторов и контекста



Необходимо уделить особое внимание выделению уникальных критериев, которые будут релевантными для конкретной задачи

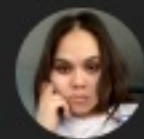
Анализ проблемы



[Executive summary](#)[Анализ проблемы](#)[Анализ данных](#)[Реализация](#)[Результат](#)[Наша команда](#)**Hacker News**

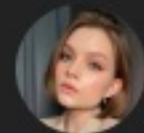
5 дней назад

Датасет к кейсу содержит в себе информацию о 88700 постах. Каждому посту соответствуют 5 комментариев с уникальным рейтингом от 0 до 4

**Анастасия Лим**

У этого крутого комментария ранг 0, вот он и на первом месте в топе

20 минут назад

[Ответить](#) [Поделиться](#)**Ксения Усенко**

У моего комментария ранг 1, но я не топ-1 (

17 минут назад

[Ответить](#) [Поделиться](#)**Илья Смирнов**

А я Илья. Я написал не очень интересный коммент, вот он и на 3 месте

27 минут назад

[Ответить](#) [Поделиться](#)**Дарья Карпунь**

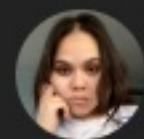
супер крутой кейс спасибо вк пишу без пунктуация привет 4 место

5 минут назад

[Ответить](#) [Поделиться](#)**Никита Кудряшов**

класс.

1 минут назад

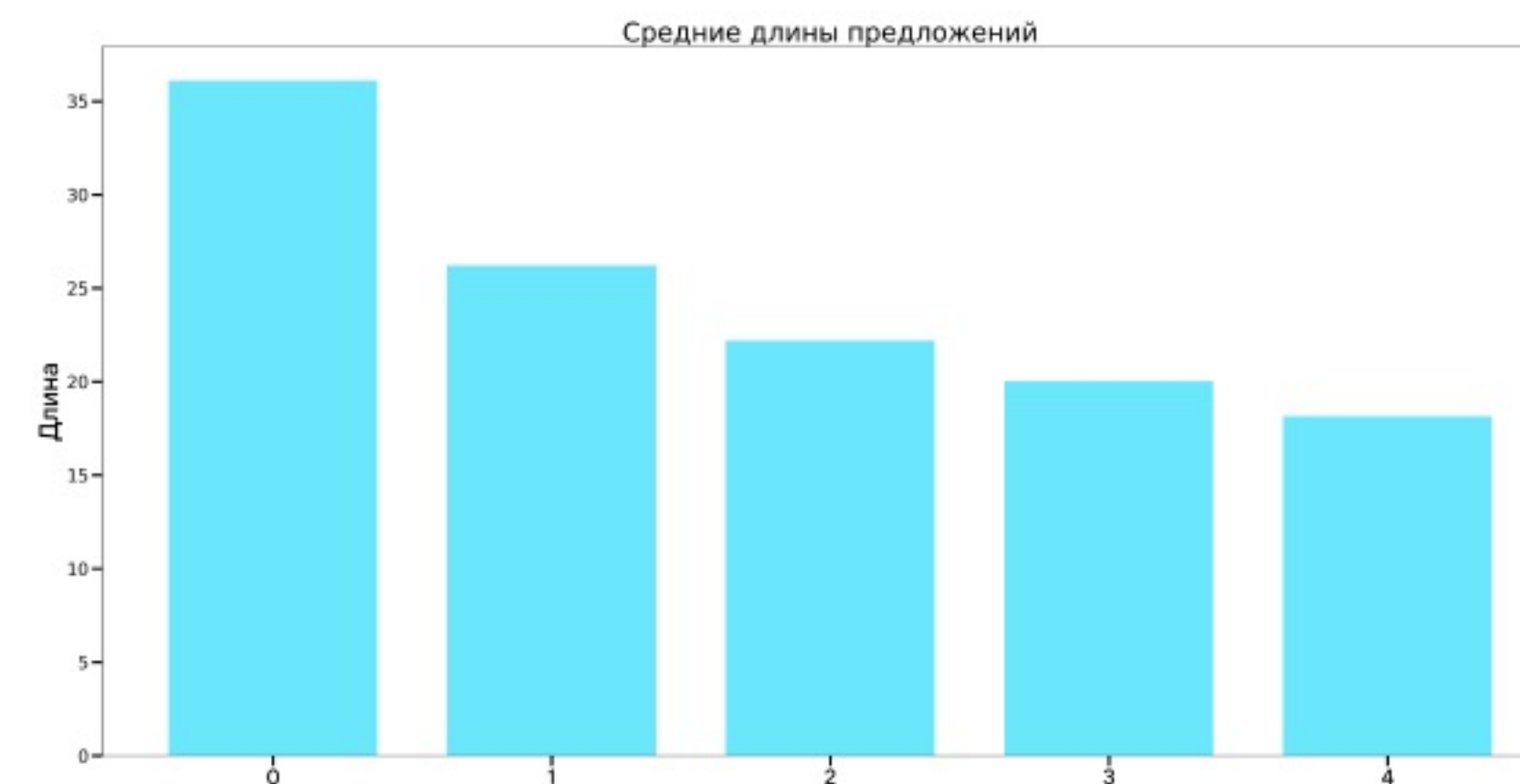
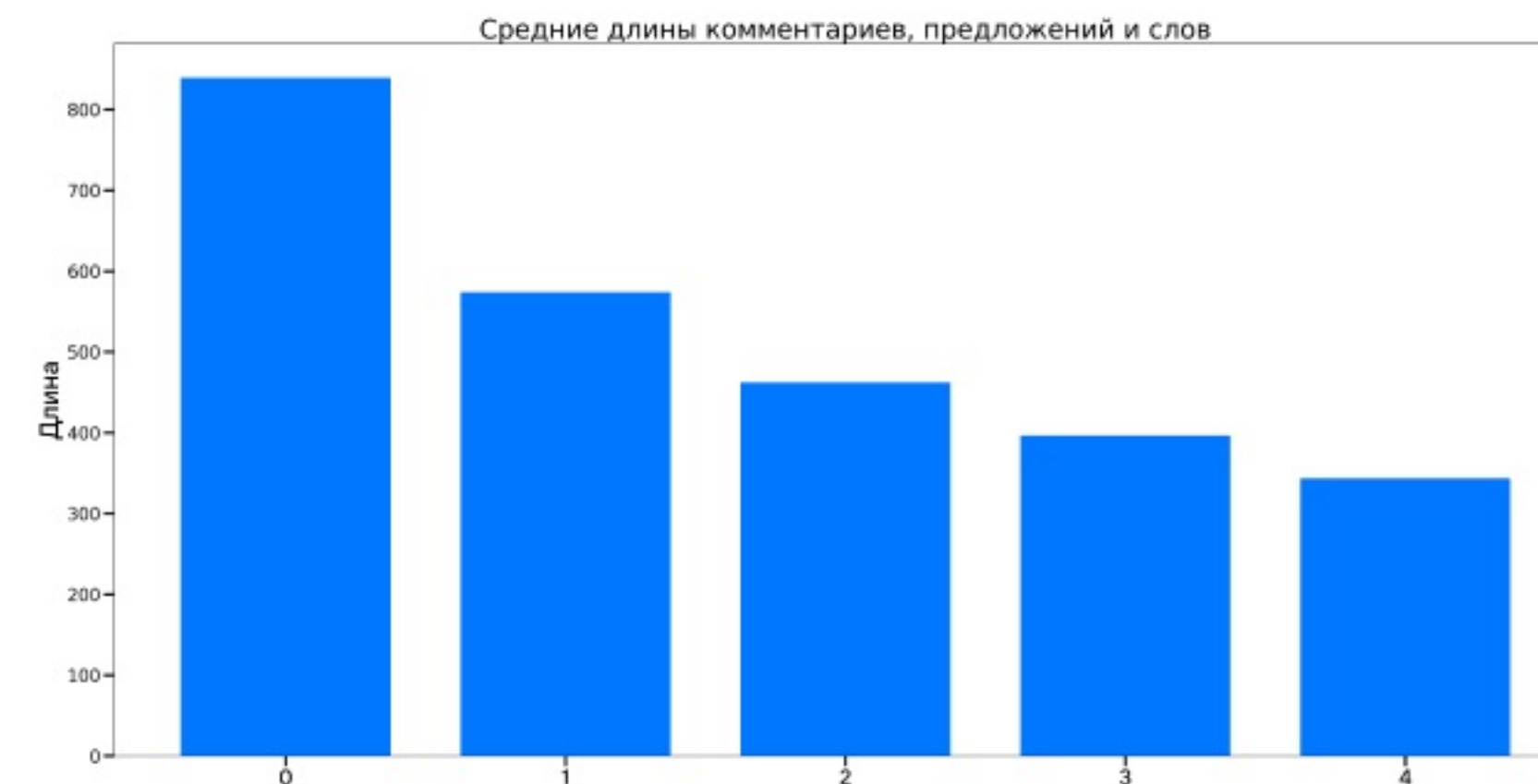
[Ответить](#) [Поделиться](#)

мне не понравился этот пост! там такой был странный датасет он ещё страшный и ну дурак!!!!!! как можно было упустить все? я считаю эти комментарии умные классные и на самом датасет их не достоин😂😂😂😂😂😂😂😂

Осталось 76 знаков



Изучение исходных данных с узкой специализацией позволяет выделить основные факторы, влияющие на оценку комментариев со стороны пользователей



При первичном анализе данных датасета было обнаружено, что для пользователей данной платформы основным критерием комментария является не только смысловая нагрузка, но и его объем. Оказалось, что длина предложений в комментариях с высокой оценкой в два раза превышает длину предложений в комментариях с низкой оценкой. Таким образом, мы сделали вывод о том, что данная платформа представляет собой форум, на котором люди ценят мнение и логику, выраженные в объемном и содержательном комментарии



Предобработка исходных данных и исследование тонких закономерностей позволяет подтвердить гипотезы и сформировать **НОВЫЕ**

Первичная обработка датасета

декодировка символов html → удаление всех ссылок и одиноко стоящих чисел

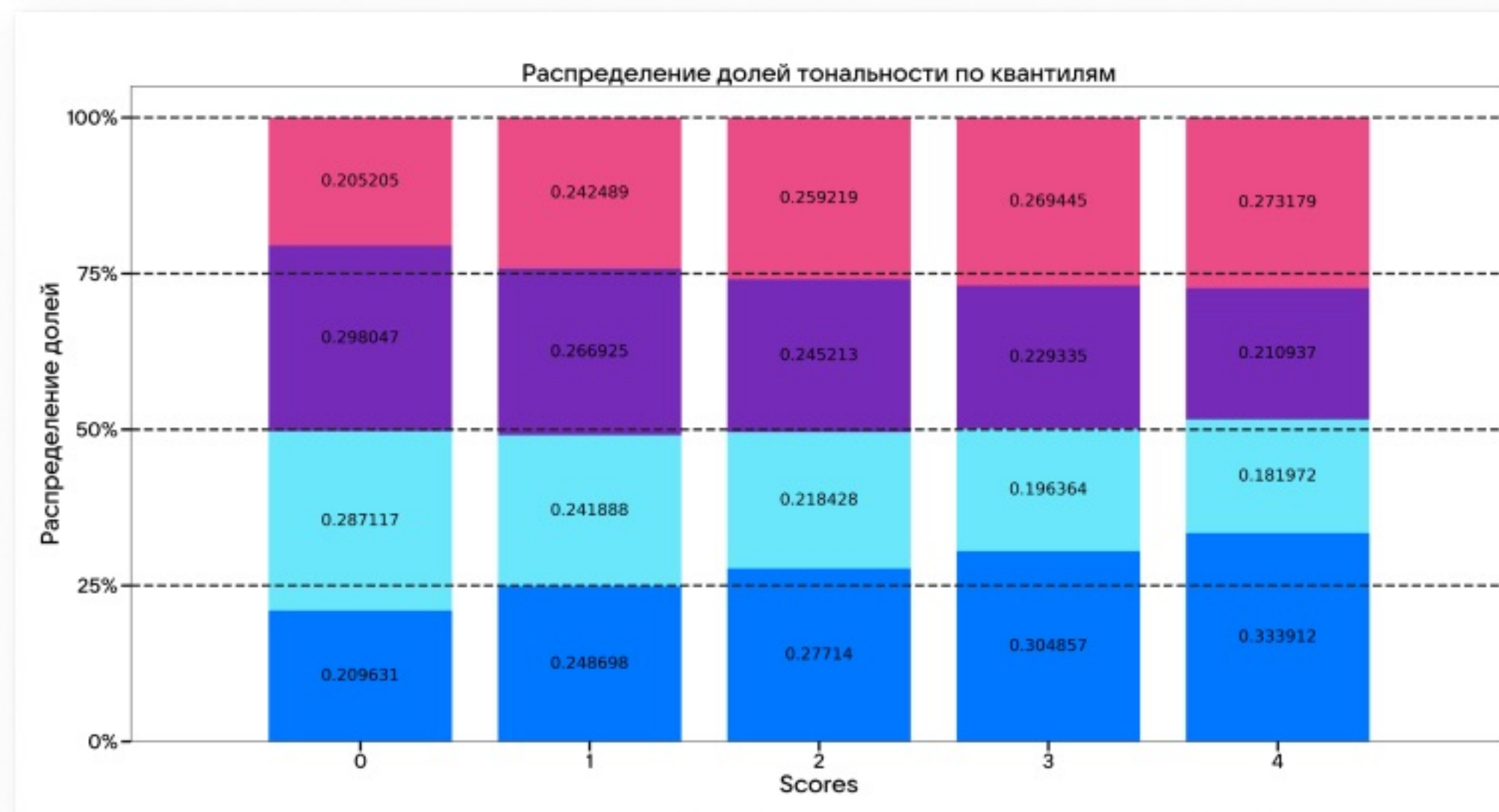
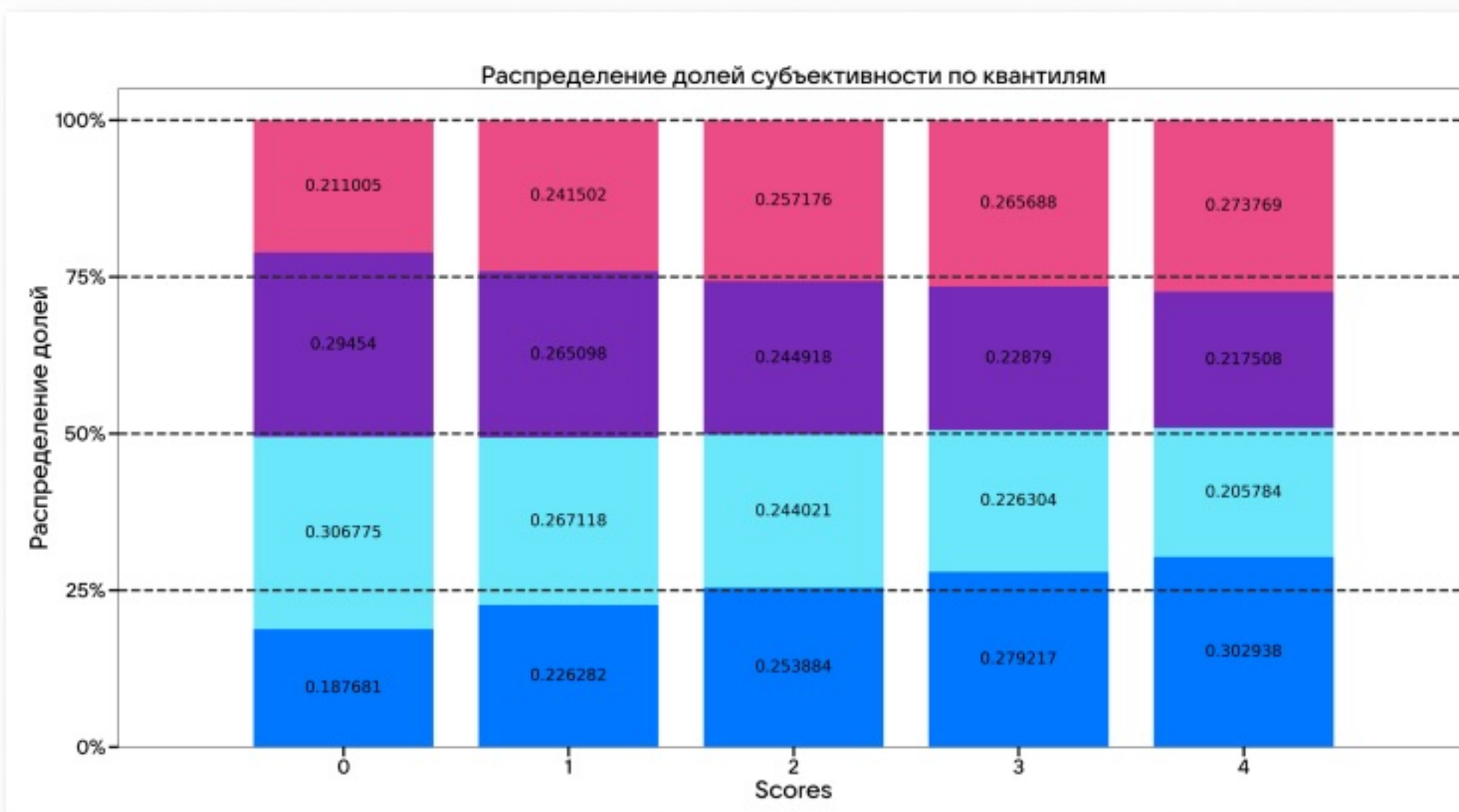
→ лемматизация слов в комментариях

Результат предобработки

были обнаружены пустые строки (ранее состоящие из чисел или ссылок), подобные комментарии были приняты за выбросы и к дальнейшему рассмотрению не привлекались

С использованием предобученных нейронных сетей была выполнена оценка тональности и субъективности каждого комментария и поста. Несмотря на отсутствие связи между этими факторами, были обнаружены интересные тенденции.

А именно: выражение эмоций в комментариях, как позитивных, так и негативных, снижает шансы автора на попадание в топ, в то время как баланс между выражением мнения и фактическими данными увеличивает вероятность успеха.





Анализ пожеланий пользователей и уникальных характеристик датасета позволяет создать полный набор факторов для обучения модели, основанный на текущих тенденциях в комментариях

Ранжирование

	точечное	парное	списочное
подход	ранг каждого комментария основывается только на его индивидуальной релевантности к посту	оценивает пары комментариев и выбирает наилучший из них	оценивает сразу все комментарии к одному посту
решение	не учитывает связь текущего комментария с контекстом остальных комментариев.	наилучший комментарий может быть не выбран в итоговом списке, если он проиграет в одной из пар.	учитывает взаимосвязь между комментариями и строит оптимальный ранжированный список

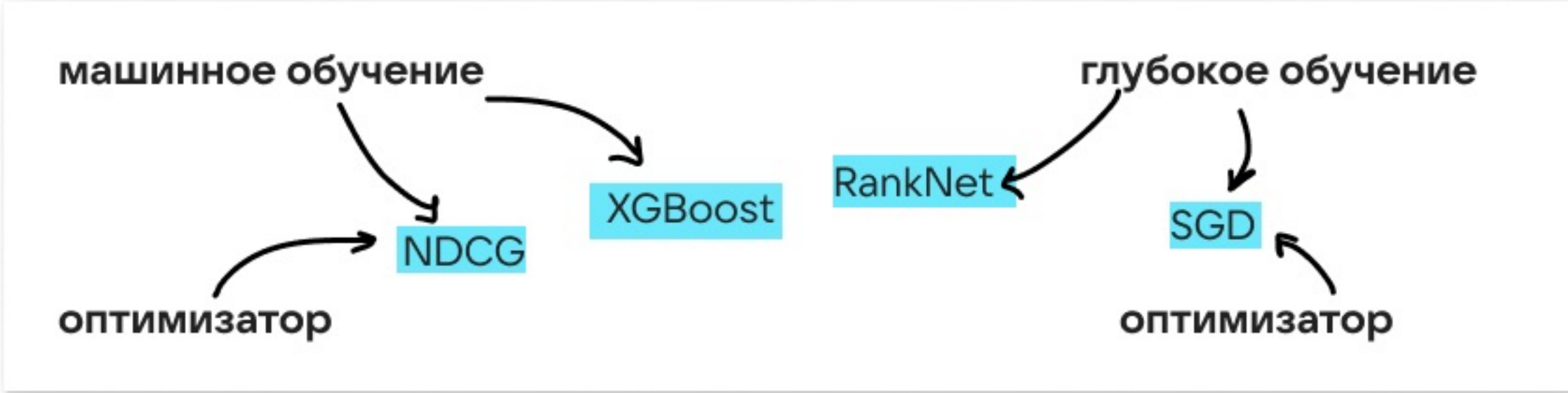
Факторы

тональность	стоп-слова	среднее число слов
субъективность	число предложений	связь текста и поста
пунктуация	знаки препинания	средняя длина слова
число символов	...	

Векторизация текста

"I love VK" + FastText → [0.1751606 , 0.20864713, -0.04136262, -0.34719867, 0.49617186, -0.37059578, -0.1569933 , -0.50440276, 0.62647736, -0.14441289, -0.10911278, -0.81821746, 0.90729654, 1.0385497 , 0.04335793]

Инструменты обучения





Executive summary

Анализ проблемы

Анализ данных

Реализация

Результат

Наша команда

Построение и обучение модели на исходном датасете позволяет с высокой точностью предсказывать поведение пользователей из представленной социальной группы и анализировать их предпочтения

Метрика NDCG

XGBoost: 81,9%

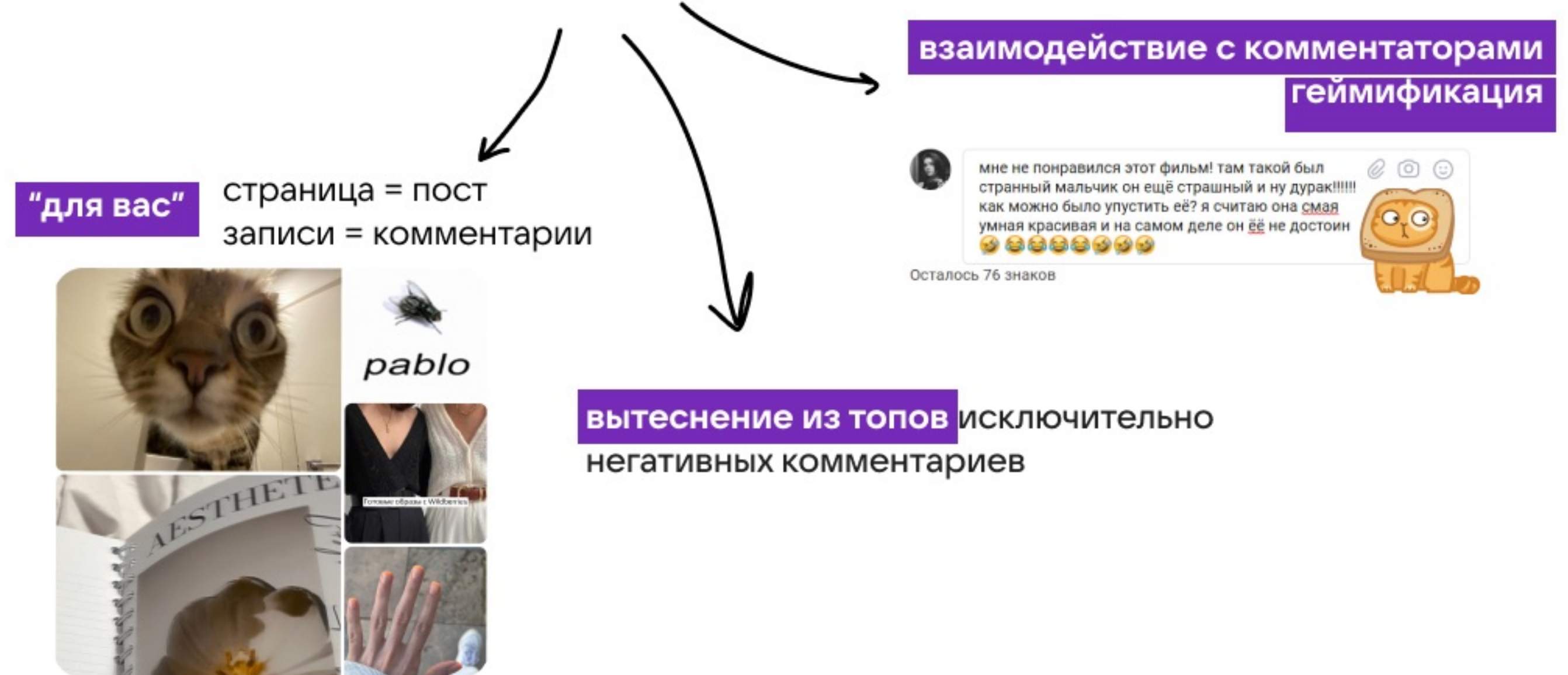
RankNet: 88,5%



Вывод: подтвердились ли наши гипотезы?

Механизм ранжирования определяется не только текстовыми факторами, однако наша модель, основываясь исключительно на тексте и находя тенденции, подтверждает важность факторов, которые мы представили для оценки популярности комментариев

Масштабирование алгоритма





Executive summary

Анализ проблемы

Анализ данных

Реализация

Результат

Наша команда

Команда **Team Target**

Студенты 3 курса Финансового университет при Правительстве РФ,
факультета Информационных технологий и анализа больших данных



Карпунь Дарья

dashapetrova01@gmail.com
+7 (965) 354-09-18



Смирнов Илья

smiril13@mail.ru
+7 (926) 764-15-62



Усенко Ксения

usenko.ks59@gmail.com
+7 (917) 126-35-83



Кудряшов Никита

nkmeowski@gmail.com
+7 (985) 977-08-98



Лим Анастасия

nastenka.lim@gmail.com
+7 (929) 900-15-56