**Литературный обзор проекта**

Sivanandham S., A. Sathish Kumar и другие в своей статье анализируют 3500 статей, опубликованных в журнале «Expert Systems with Applications» с 2009 по 2018 год. Для определения 20 наиболее популярных тем и 10 наиболее ассоциированными с ними слов используется N-Gram Latent Dirichlet Allocation (LDA). Также находится согласованность (coherence) и сходство (similarity) тем, из чего делается вывод о подтемах. Выявляются активные, мертвые и только возникшие темы. Используется VAR (Vector AutoRegression) модель для предсказания будущих популярных тем. [1]

Sun L. и Yin Y. в своей работе анализируют статьи ведущих транспортных журналов с 1990 по 2015 год. Для определения 50 ключевых тем по аннотациям статей используется модель скрытого распределения Дирихле (LDA). В результате были выявлены основные темы, освещённые в статьях; проанализировано их изменение с течением времени; с помощью метода иерархической кластеризации выявлены наиболее схожие по тематике статей журналы и страны/регионы; построена сеть наиболее часто встречающихся вместе слов [2].

Vayansky I. и Kumar S. A. P. в своей обзорной статье приводят углубленный обзор наиболее популярных методов тематического моделирования, которые могут быть применимы в различных ситуациях. Также приводится несколько обобщенных алгоритмов оптимизации, которые могут помочь при определении оптимального количества тем и других настроек моделей. Также приводится небольшое дерево решений, которое должно помочь исследователям определиться с выбором нужно метода тематического моделирования, на основе их индивидуальных потребностей [3].

Работа: <https://www.kaggle.com/vanshjatana/text-classification-from-scratch>

Задача: Классифицировать, фейковые ли новости.

Данные: текст с метками, а также заголовком, датой, и темой

Методы решения: Лог. Регрессия. Multinomial, Bernoulli Naive Bayes, Gradient Boost Classifier, XGBoost,SGDClassifier, Random Forest

Результат: Наибольшая точность у SVM и деревьев решений.

Работа: <https://www.kaggle.com/selener/multi-class-text-classification-tfidf>

(Жалобы потребителей на финансовые продукты и услуги)

Задача: Классиф. жалобы потребителей по заранее определенным категориям(категории-продукты).

Данные: После предобработки тексты жалоб и их метки

Методы решения: Смотрятся наиболее частные классы. Классифицировались также биграммы. Boxplot с точностями и отклонениями моделей.

Результат: Лучшая точность у SVM.

Работа: <https://www.kaggle.com/sagarkhambad/text-classification>

Задача: Классифицировать твит как позитивный негативный или нейтральный. Классы 'Neutral, Positive, Extremely Negative, Negative,Extremely Positive'.

Данные: Колонки ['UserName', 'ScreenName', 'Location', 'TweetAt', 'OriginalTweet', 'Sentiment']

Методы решения: 2-мерный график точности в зависимости от гиперпараметра. Для одного из классификаторов сделан трехмерный график, где на одной оси точность, а на других – 2 параметра классификатора.

Результат: Интересное заключение: Precision, Recall and F1 score for XGBoost model is descent enough than other models. So will go with XGBoost Model.

1. Sivanandham S. и др. Analysing Research Trends Using Topic Modelling and Trend Prediction // Soft Computing and Signal Processing / под ред. V. S. Reddy и др. Singapore: Springer Singapore, 2021. С. 157–166.

2. Sun L., Yin Y. Discovering themes and trends in transportation research using topic modeling // Transp. Res. Part C Emerg. Technol. 2017. Т. 77. С. 49–66.

3. Vayansky I., Kumar S. A. P. A review of topic modeling methods // Inf. Syst. 2020. Т. 94. С. 101582.