Определение классов

Классы:

1. Автомобили.
2. Технологии.
3. Политика.
4. Трейлер.
5. Игры.
6. Экономика.
7. Общество
8. Аниме.
9. Манга.
10. Интернет.
11. Наука.
12. Происшествия.
13. Спорт.
14. Культура.
15. Прочее.

Выбранные методы векторизации

Для обучения и сравнения были выбраны следующие 2 метода векторизации:

- Bag of Words
- Word2Vec

Вектор признаков для bag of words

Пусть: A – множество уникальных слов на русском языке (кроме служебных частей речи, слов на латинице и неизвестных слов) в рамках обучаемой выборки. Тогда:

$$A = \{a_1, a_2, \dots, a_n\},$$

где a_i – i-ое уникальное слово в нормальной форме;

n – размер множества A.

Пусть: V – вектор признаков. Тогда:

$$V = \{v_1, v_2, \dots, v_n\},$$

где v_i – частота встречаемости слова a_i для входного текста.

Способ векторизации текстов для bag of words

Шаги:

1. Вектор V для заданного текста автоматически заполняется нулями.
2. Из текста извлекаются нормальные формы слов $B = \{b_1, b_2, \dots, b_m\}$, где m – количество слов, которые не принадлежат служебным частям речи, написаны не на латинице и найдены в словаре морфологического анализатора.
3. Проверяется принадлежность каждого b_j множеству A.
4. Если $b_j \in A$ и $b_j = a_i, i \in 1..n$, то $v_i = v_i + 1$.

Способ векторизации текстов для Word2Vec

Для векторизации текстов в Word2Vec используется модель, предоставляемая библиотекой gensim. Модель обучается на основе поступаемого списка текстов, а затем используется для векторизации текстов перед классификацией.

Выбранные модели для обучения и их сравнения

Для обучения были использованы следующие модели:

- LogisticRegression
- LinearSVC
- GaussianNB
- MultinomialNB (только Bag of Words)
- RandomForestClassifier
- LogisticRegressionCV

Оценки результатов изображены на таблицах 1-6 для Bag of Words и 7-11 для Word2Vec.

Для обучения был использован алгоритм MultiOutputClassifier, который позволяет обученной модели иметь несколько выходов. В случае нашей обучаемой модели – это массив $C^{pred} = \{c_1^{pred}, c_2^{pred}, \dots, c_m^{pred}\}$, где m – кол-во классов (категорий), $c_j^{pred}, j = 1..m$ – значение 1 или 0 – принадлежит или не принадлежит статья к данной категории.

Таблица 1 – Оценка Bag of Words, модель LogisticRegression

Accuracy	0.559
Precision	0.883
Recall	0.636
F1 score	0.740
Hamming loss	0.035

Таблица 2 – Оценка Bag of Words, модель LinearSVC

Accuracy	0.601
Precision	0.891
Recall	0.669
F1 score	0.764
Hamming loss	0.032

Таблица 3 – Оценка Bag of Words, модель GaussianNB

Accuracy	0.286
Precision	0.898
Recall	0.321
F1 score	0.473
Hamming loss	0.056

Таблица 4 – Оценка Bag of Words, модель MultinomialNB

Accuracy	0.334
Precision	0.916
Recall	0.378
F1 score	0.535
Hamming loss	0.051

Таблица 5 – Оценка Bag of Words, модель RandomForestClassifier

Accuracy	0.296
Precision	0.944
Recall	0.346
F1 score	0.507
Hamming loss	0.052

Таблица 6 – Оценка Bag of Words, модель LogisticRegressionCV

Accuracy	0.569
Precision	0.883
Recall	0.646
F1 score	0.746
Hamming loss	0.034

Таблица 7 – Оценка Word2Vec, модель LogisticRegression

Accuracy	0.525
Precision	0.847
Recall	0.625
F1 score	0.719
Hamming loss	0.038

Таблица 8– Оценка Word2Vec, модель LinearSVC

Accuracy	0.548
Precision	0.861
Recall	0.644
F1 score	0.737
Hamming loss	0.036

Таблица 9 – Оценка Word2Vec, модель GaussianNB

Accuracy	0.043
Precision	0.290
Recall	0.931
F1 score	0.443
Hamming loss	0.182

Таблица 10 – Оценка Word2Vec, модель RandomForestClassifier

Accuracy	0.601
Precision	0.862
Recall	0.675
F1 score	0.757
Hamming loss	0.033

Таблица 11 – Оценка Word2Vec, модель LogisticRegressionCV

Accuracy	0.665
Precision	0.871
Recall	0.757
F1 score	0.810
Hamming loss	0.027

На основании полученных данных был выбран метод Word2Vec с моделью LogisticRegressionCV.