Анализ тональности отзывов о запрещенных веществах

Кирилл О. Конча

Национальный исследовательский университет «Высшая школа экономики», Москва, Россия, majortomblog@gmail.com

0. Дисклеймер

Употреблять наркотики смертельно опасно. Хранить их и торговать — уголовное преступление. Работа посвящена лингвистическим аспектам этой противозаконной деятельности. Автор против наркотиков, поэтому не раскрывает название площадки откуда были собраны отзывы и способы попасть туда.

1. Введение

Термин Darknet обычно используется для обозначения сайтов, которые не индексируются поисковыми системами. Несмотря на большую популярность Darknet'а (особенно в области нелегальной деятельности), очень мало известно о характеристиках используемого там языка.

Работа [Chosen et al 2019] показывает, что языки легальной и нелегальной деятельности в Darknet'е отличаются. Эти два типа текстов имеют лексические и синтаксические различия между собой. Кроме того, именованные сущности в текстах нелегальной деятельности часто не имеют соответствующих страниц в Википедии. Настоящее исследование представляет собой анализ тональности отзывов о наркотических веществах с русскоязычного тематического портала. С помощью анализа тональности в работе выявляются эмоционально-окрашенные лексические средства свойственные положительным и отрицательным отзывам.

В настоящем исследовании анализ тональности производится с машинного обучения с учителем. В качестве алгоритмов используются линейные модели, так как взаимосвязь элементов и их коэффициентов в них прямая, что повышает интерпретируемость важности лексических средств. Согласно [Chitla 2021, p.31; Nguyen et al. 2018, p.16], логистическая регрессия (LogReg) и метод опорных векторов (SVM) показывают высокую эффективность при анализе тональности. В работе не используются нейросетевые модели из-за проблем с интерпретируемостью: целью работы является не только успешное определение полярности отзывов, но и выявление и описание специфичных лексических средств, используемых в них.

2. Сбор и обработка материала

В тренировочную выборку входят 1,000 отзывов о разных видах наркотических веществ, собранных и размеченных с тематической площадки в Darknet'e. В выборку входят тексты о пяти видах наркотических веществ, по 200 отзывов о каждом (половина положительных и половина отрицательных). Отзывы собирались вручную: в качестве положительных отзывов брались отзывы с четырьмя или пяти звездами (из пяти); в качестве отрицательных брались отзывы с нулем или одной звездой. Положительные отзывы были размечены как 1, отрицательные как -1. В тестовую выборку входят 200 отзывов с того же сайта, среди которых 99 положительных и 101 отрицательный. Все отзывы были лемматизированы (PyMorphy2) и очищены от стоп-слов (NLTK). Для векторизации текстов использовались Count Vectorizer и TF-IDF.

3. Обучение моделей

Для того, чтобы выбрать оптимальную пару метода векторизации текста и модели использовалась кросс-валидация с делением тренировочных данных на пять частей.

Наилучший результат достигается, если использовать SVM и TF-IDF (табл. 1). На тестовом наборе данных accuracy этой связки составляет 0.935.

	Count Vectors	**TF-IDF**
LogReg	0.86	0.864
SVM	0.859	0.875

Таблица 1. Средние accuracy моделей при перекрестной проверке с делением данных на пять частей.

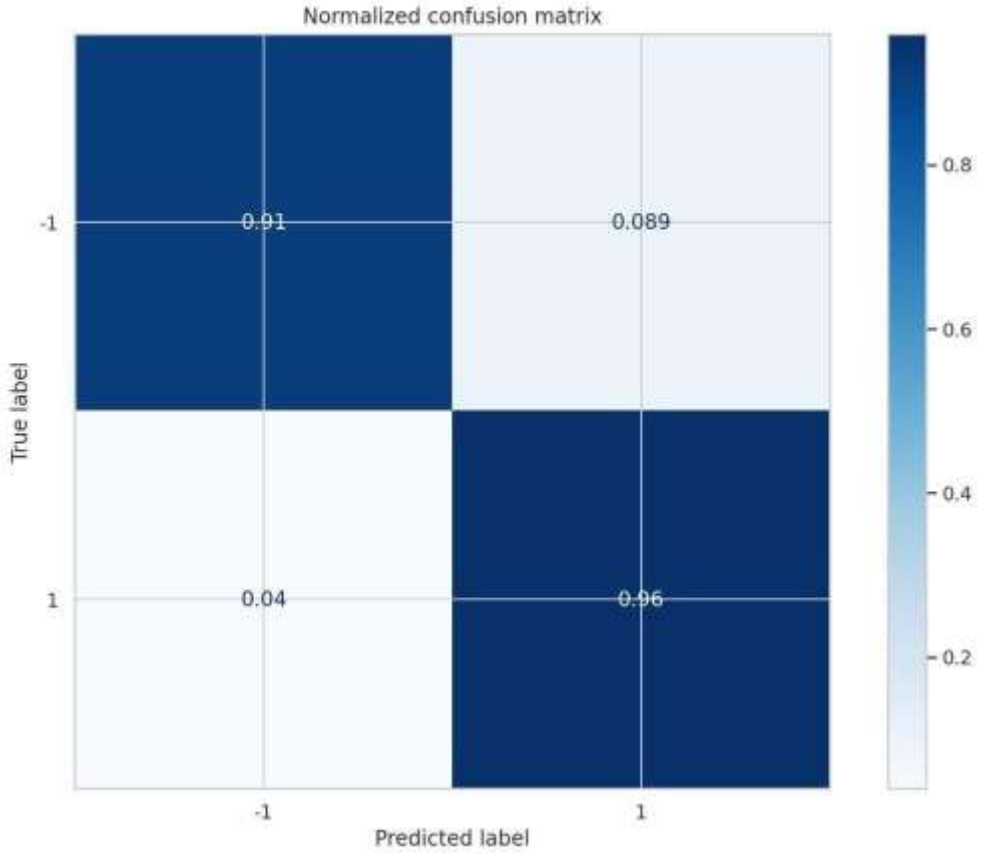

Рисунок 1. Нормализированная матрица ошибок для связи SVM и TF-IDF Vectorizer.

Кроме того, с помощью Spacy-Stanza были также размечены части речи во всех отзывах (без удаления стоп слов). При обучении SVM с TF-IDF на данных с проставленными частеречными тэгами среднее значение accuracy при кроссвалидации с разделением тренировочных данных на пять частей составляет 0.665. На тестовом наборе accuracy имеет значение 0.655. Это может свидетельствовать в пользу того, что между положительным и отрицательными отзывами помимо лексических отличий также существуют синтаксические.

Рисунок 2. Пять самых важных частей речи для положительных (красный цвет) и отрицательных отзывов (синий цвет).

4. Результаты

В результате были получены 20 самых важных слов для определения положительности или отрицательности отзыва (рис. 3). Обоими моделями наибольшие веса были приписаны моделью словам касание (для положительных отзывов) и ненаход (для отрицательных отзывов). *Ненаход* обозначает ситуацию, когда потребитель не нашел товар на месте. *Касание* же наоборот используется, когда товар был без трудностей получен. *Касание* может может употребляться как в качестве самостоятельного слова, так и с предлогом *в*, а также с глаголами *забрать, снять* и *поднять*.

(1) Худший магазин. В людном месте , ненаход уже 2 раз …

(2) Касание, всё как обычно. Купил забрал сожрал)))10.10.10

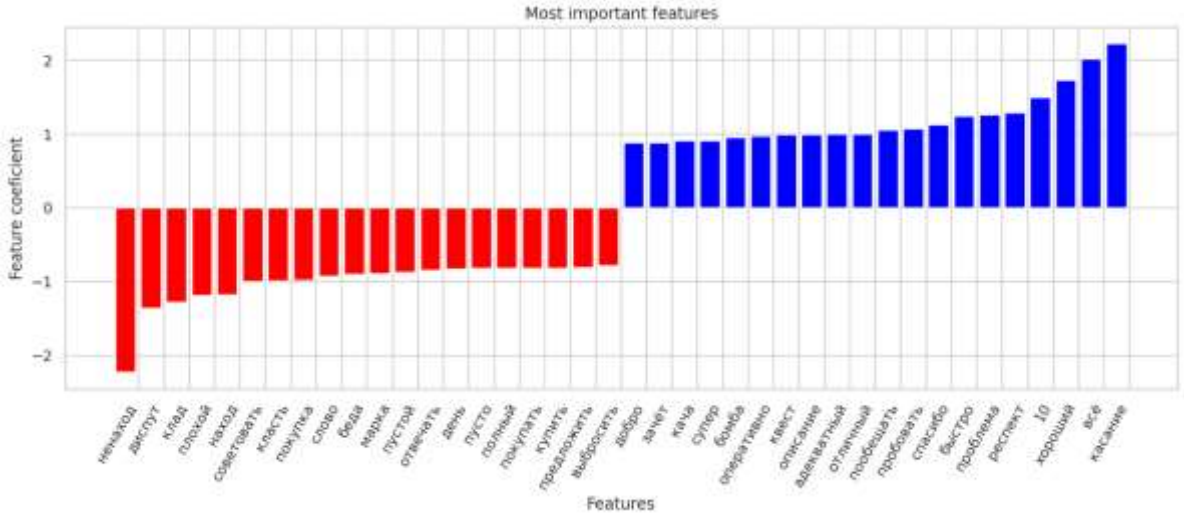

Рисунок 3. 20 самых важных слов, по которым модель оценивала отзыв как положительный (выделены синим цветом) или отрицательный (выделены красным цветом) и их коэффициенты важности.

5. Заключение

Стоит отметить, что выявленные слова относятся не к самому качеству продаваемых товаров, а к качеству их дистрибуции. В условиях, когда невозможно просто получить товар и для этого приходится проделывать ряд сложных действий, качество отходит на второй план. На первый план же выходит сам факт получения покупки и то, насколько это получение было простым.

Благодарности

Автор выражает благодарность за ценные замечания Даниилу Скоринкину и Ивану Торубарову, а также анонимным рецензентам.

Литература

Chitla 2019 – Chitla, Pravalika Ravikumar. Sentiment Analysis of Reviews / Iowa State University Capstones, Theses and Dissertations. Creative Components 721, 2019.

Chosen et al. 2019 – Choshen. L, Eldad. D, Herschcovich. D, Sulem. E, Abend. O. The Language of Legal and Illegal Activity on the Darknet Florence / Italy: Association for Computational Linguistics, 2019.PP. 4271–4279.

Nguyen et al. 2018 – Nguyen, Heidi; Veluchamy, Aravind; Diop, Mamadou; and Iqbal, Rashed. Comparative Study of Sentiment Analysis with Product Reviews Using Machine Learning and Lexicon-Based Approaches / SMU Data Science Review: Vol. 1, No. 4, Article 7, 2018.