Алгоритмы интеллектуального анализа текстовых данных для классификации новостных сообщений

О.В. Курбатова[2], А.В. Куприянов[1,2]

[1]*Институт систем обработки изображений РАН - филиал ФНИЦ «Кристаллография и фотоника» РАН, Молодогвардейская 151, Самара, Россия, 443001*
[2]*Самарский национальный исследовательский университет им. академика С.П. Королева, Московское шоссе 34а, Самара, Россия, 443086*

Аннотация
Классификация текстов является одной из основных задач компьютерной лингвистики, поскольку к ней сводится ряд других задач: определение тематической принадлежности текстов, эмоциональной окраски высказываний и др. В связи с ростом новостного потока, а также влияния информационного фона на принятие решений становиться актуальной задача классификации новостных сообщений. Данная статья представляет собой обзор методов классификации текстов, их механизма реализации, целями которого являются сравнение современных методов решения задачи классификации текстов, обнаружение тенденций развития данного направления.

Ключевые слова
Классификация, обработка естественного языка, Bag-of-Word, Word2Vec, BERT

1. Введение

NLP (обработка естественного языка) часто применяется для классификации текстовых данных. Классификация текста - это проблема присвоения категорий текстовым данным в соответствии с их содержанием.

На данный момент роль новостных лент и агрегатор существенно возросла. Многим компаниям необходимо анализировать информационных фон для оценки отношения пользователей к своим продуктам и своей репутации. Собрав и кластеризировав текстовые данные из новостных агрегаторов, можно определить основные темы и события, выявить связи между компаниями и корпоративными действиями.

В этой статье сравнивается хорошо известный Bag-of-Words (используемый с простым алгоритмом машинного обучения), популярная модель встраивания слов (используется с нейронной сетью глубокого обучения) и современные языковые модели (используемые при передаче обучение у трансформера), которые полностью изменили ландшафт NLP.

2. Модель Bag-of-Words

Модель Bag-of-Words проста: она строит словарь из корпуса документов и подсчитывает, сколько раз слова встречаются в каждом документе. Такой подход вызывает серьезную проблему размерности: чем больше у вас документов, тем больше словарный запас, поэтому матрица характеристик будет огромной разреженной матрицей. Следовательно, модели Bag-of-Words предшествует важная предварительная обработка (очистка слов, удаление стоп-слов, выделение корней / лемматизация), направленная на уменьшение проблемы размерности.

Частота использования терминов не обязательно является лучшим представлением текста. Фактически, можно найти в корпусе общие слова с наибольшей частотой, но с небольшой предсказательной силой по целевой переменной. Для решения этой проблемы существует расширенный вариант пакета слов, в котором вместо простого подсчета используется термин "частота - обратная частота документа" (или Tf – Idf). По сути, значение слова увеличивается пропорционально количеству, но оно обратно пропорционально частоте слова в корпусе.

3. Модель встраивания слов

Встраивание слов - это собирательное название методов изучения функций, при котором слова из словаря сопоставляются с векторами действительных чисел. Эти векторы вычисляются из распределения вероятностей для каждого слова, появляющегося до или после другого. Популярными моделями встраивания слов являются Word2Vec от Google (2013 г.), Stanford GloVe (2014) и FastText от Facebook (2016).

Word2Vec создает векторное пространство, обычно состоящее из нескольких сотен измерений, с каждым уникальным словом в корпусе, так что слова, которые имеют общий контекст в корпусе, располагаются в пространстве рядом друг с другом. Это можно сделать, используя 2 разных подхода: начиная с одного слова, чтобы предсказать его контекст (Skip-gram), или начиная с контекста, чтобы предсказать слово (Continuous Bag-of-Words).

4. Языковые модели вложения слов

Языковые модели или динамические вложения слов преодолевают самое большое ограничение классического подхода к встраиванию слов: устранение многозначности, слово с разными значениями идентифицируется только одним вектором. Одной из первых популярных моделей была ELMO (2018), которая не применяет фиксированное вложение, но использует двунаправленный LSTM, просматривает все предложение, а затем назначает вложение каждому слову.

BERT от Google сочетает в себе встраивание контекста ELMO и несколько преобразователей, а также двунаправленность. Вектор, который BERT назначает слову, является функцией всего предложения, поэтому слово может иметь разные векторы в зависимости от контекстов.

Чтобы выполнить задачу классификации текста, можно использовать BERT тремя различными способами: обучить все с нуля и использовать как классификатор; извлечь вложения слов и использовать их в слое встраивания или тонко настроить предварительно обученные модели (трансферное обучение).

5. Заключение

Вопросы, связанные с дальнейшей классификацией текстовых данных, являются актуальными в связи с колоссальным новостным потоком. Подходы и методы, представленные в статье, планируются к апробации над текстовыми данными, собираемыми из новостных в российском сегменте. Планируется развить данную тему в направлении оптимизации техники извлечения ключевых слов, которая использует языковую модель BERT.

6. Литература

[1] Коваленко, Т.В. Разработка библиотеки построения векторной модели текста на основе морфемного разбора слов / Т.В. Коваленко, Р.Б. Галинский, Ю.В. Яковлева, И.В. Никифоров // Неделя науки СПбПУ. Институт компьютерных наук и технологий – СПб.: Изд-во Политехн. ун-та, 2017. – 518 с.

[2] Mikolov, T. Efficient Estimation of Word Representations in Vector Space. / T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean [Electronic resource]. – Access mode: https://arxiv.org/abs/1301.3781 (21.11.2020).

[3] Pennington, J. Global Vectors for Word Representation / J. Pennington, R. Socher, C.D. Manning [Electronic resource]. – Access mode: https://nlp.stanford.edu/projects/glove/ (25.11.2020).

[4] Vaswani, A. Attention Is All You Need / A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones [Electronic resource]. – Access mode: https://arxiv.org/abs/1706.03762 (27.11.2020).