

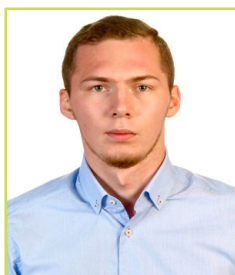
ПРОЕКТИРОВАНИЕ И ПРОГНОЗИРОВАНИЕ В СОЦИАЛЬНО-ЭКОНОМИЧЕСКОЙ СФЕРЕ

ПРОБЛЕМЫ ПРИ СОЗДАНИИ НЕЙРОННОЙ СЕТИ ДЛЯ СЕНТИМЕНТ АНАЛИЗА

PROBLEMS IN CREATING A NEURAL NETWORK FOR SENTIMENT ANALYSIS

**Зубарев
Максим Олегович**

Андроид-разработчик Gilraen Limited, Москва, Россия



**Zubarev
Maksim Olegovich**

Android developer Gilraen Limited, Moscow, Russia

**Конюхова
Галина Павловна**

*Кандидат педагогических наук, доцент кафедры
«Управление и информатика в технических системах»
Московского государственного технологического
университета «СТАНКИН»*



**Konjuhova
Galina Pavlovna**

*Candidate of Pedagogy, Associate Professor
of the Department of Management and Informatics
in Technical Systems, Moscow State Technological University
"STANKIN"*

Аннотация. Подавляющее большинство работ по анализу настроений было посвящено разработке более точных классификаторов настроений, обычно включающих контролируемые алгоритмы машинного обучения и набор функций. В этой главе мы подробно остановимся на некоторых нерешенных проблемах, вопросах, которые еще недостаточно изучены, и новых проблемах, возникающих в результате решения новых проблем анализа настроений. Мы также обсуждаем предложения по решению этих проблем. Цель этой статьи состоит в том, чтобы снабдить исследователей и практиков указателями на последние достижения в области анализа настроений и стимулировать больше работы в разнообразной области проблем, особенно в тех областях, которые относительно менее изучены.

Ключевые слова: сентимент анализ, нейронная сеть, сверточные нейронные сети, обработка естественного языка.

Abstract. A vast majority of the work in Sentiment Analysis has been on developing more accurate sentiment classifiers, usually involving supervised machine learning algorithms and a battery of features. In this chapter, we flesh out some of the challenges that still remain, questions that have not been explored sufficiently, and new issues emerging from taking on new sentiment analysis problems. We also discuss proposals to deal with these challenges. The goal of this chapter is to equip researchers and practitioners with pointers to the latest developments in sentiment analysis and encourage more work in the diverse landscape of problems, especially those areas that are relatively less explored.

Keywords: sentiment analysis, neural networks, convolutional neural networks, natural language processing.

Анализ настроений является одним из наиболее успешных и распространенных приложений в обработке естественного языка. Однако, несмотря на всю шумиху, которую он вызвал с момента своего создания, с ним все еще связано много проблем.

Анализ настроений может быть очень простым и дешевым, так как для этого существует множество бесплатных библиотек. Вот некоторые примеры: Syuzhet (для R), NLTK (python), spacy (python). Однако анализ

настроений иногда может быть очень сложным и сложным, и именно об этом я хочу здесь поговорить.

В частности, анализ настроений страдает одним серьезным недостатком: он зависит от контекста и языка. В этой статье я говорю о некоторых потенциальных проблемах, которые могут возникнуть при попытке применить анализ настроений к некоторым контекстам.

Вопросы анализа настроений

Проблема 1: Слова имеют противоположное значение в вашем контексте. Быть «агрессивным» в большинстве ситуаций считается не очень приятной чертой характера. Однако быть агрессивным, когда ты нападающий в футболе, может быть очень хорошо. Контраст еще более заметен с такими словами, как «убийца». Нападающий в футболе, который является «убийцей» или обладает «инстинктом убийцы», вероятно, является хорошим спортсменом. Однако не многие люди подумают, что «убийца» в реальной жизни — это хорошо. В этом контексте механизмы анализа настроений общего назначения будут очень запутаны.

Проблема 2: Смайлики и их использование. Иногда люди могут быть очень креативными с использованием смайликов. Например, «☐» может быть интерпретировано как отрицательное или положительное. Фактическое значение зависит от контекста. Если вы зайдете в emojiTracker, вы найдете огромное количество смайликов. Значение многих из них (например, рыбы) может быть сложным для изучения в модели машинного обучения. Это может быть очень важным вопросом для таких тем, как сарказм, который является следующей темой в нашем списке.

Проблема 3: Ирония и сарказм. В целом они относятся к числу святых граблей НЛП. Алгоритмы все еще пытаются понять сарказм и иронию. Это общеизвестно сложная проблема. В последнее время наблюдается некоторый прогресс в использовании глубоких нейронных сетей. В недавней статье был сделан вывод о том, что сарказм зависит от темы и контекста. Это означает, что алгоритм нуждается в дополнительной информации, чтобы правильно классифицировать сарказм.

Это делает его значительно более сложной проблемой, чем просто понимание того, передают ли некоторые слова положительное или отрицательное значение, и может потребовать предварительно подготовленных вложений слов и моделей личности. Решением, согласно этой статье, является использование сверточных нейронных сетей. Свертки — очень интересный метод в глубоком обучении. Если вас интересует эта тема, существуют статьи, которые хорошо объясняют, как их можно использовать в тексте.

Проблема 4: Реальные знания. Одним из наиболее сложных аспектов обработки естественного языка является понимание фактического контекста того, что говорится. В этой области проводится множество исследований. Сети памяти — это модель, которая может использовать контекстную информацию для ответа на вопросы. Сети внимания могут обрабатывать отдельные части предложения так же, как человек может обращать внимание на разные слова. Рекуррентные нейронные сети, такие как двунаправленные LSTM, могут использоваться для понимания более широкого контекста предложения. В последнее время исследования начали внедрять их в анализ настроений.

В этой статье, например, обсуждается использование сетей памяти в анализе настроений, а также в этой статье с вниманием обсуждается возможное использование двунаправленных LSTM. Оба варианта являются возможными решениями, но эти модели довольно сложны для среднего человека, плюс они требуют больших наборов данных.

Решение вопросов анализа настроений

Итак, что можно сделать для решения этих проблем? Есть несколько разных решений.

Вы можете попробовать обучить свою собственную модель для своей конкретной области, в идеале используя более продвинутый метод (например, глубокие нейронные сети). Это самое дорогое и трудоемкое решение. Однако, если у вас действительно есть ресурсы, чтобы осуществить это, вы можете обнаружить, что держите в своих руках очень ценную и уникальную интеллектуальную собственность.

Другое решение состоит в том, чтобы исследовать, насколько серьезна проблема в общей схеме вещей. Вполне вероятно, что проблема проявляется только в подмножестве данных. Возможно, вы сможете придумать несколько более сфокусированных моделей для этой части данных. Например, в приведенном выше примере, касающемся слова «агрессивный» в контексте спорта, возможное решение состоит в том, чтобы просто присвоить ему нейтральное значение. Таким образом, вы не рискуете сильно ошибиться в классификации. (например, позитивный пост помечается как негативный, или пост о ком-то, кто буквально агрессивен, помечается как позитивный). Наконец, еще одним решением является создание моделей машинного обучения для подмножеств данных.

Список литературы

1. Посевкин Р. Автоматизация сентимент-анализа естественно-языкового текста / Р. Посевкин. — LAP Lambert Academic Publishing, 2014.
2. Гудфеллоу Я. Глубокое обучение / Я. Гудфеллоу, И. Бенджио, А. Курвилль. — 2017.
3. Шолле Ф. Глубокое обучение на Python / Ф. Шолле. — 2018.
4. Жерон О. Прикладное машинное обучение с помощью Scikit-Learn и TensorFlow / О. Жерон. — 2018.
5. Саттон Р. Обучение с подкреплением / Р. Саттон, Э. Барто. — 2017.