Российский Университет Дружбы Народов

Факультет физико-математических и естественных наук

Кафедра информационных технологий

Курсовая работа на тему:

Машинное обучение в задачах анализа текстов на естественном языке

Студент: Д. С. Ларионов
Научный руководитель: И. В. Смирнов

Москва, 2018

Содержание

1. Введение **2**

2. Обзор литературы **4**

1. Введение

В течении последних 10-11 лет социальные сети прочно закрепились в нашей жизнию Люди привыкли обмениваться информацией, коммуницировать друг с другом практически круглосуточно. Исключением не являются и экстраординарные ситуации, происходящие в нашей жизни. Вполне естественно наблюдать всплески трафика социальных сетей во время событий, которые так или иначе затрагивают большое количество людей., например, во время всеобщих праздников, финалов спортивных мероприятий или черезвычайных ситуаций. И если первые два типа событий происходят ожидаемо, то черезвычайные ситуации, такие как теракты, землятресения или цунами, случаются внезапно.

Притом, практически всегда пользователи социальных сетей оказываются быстрее традиционных СМИ и представителей власти по части сообщения о факте существования той или иной ЧС. Разница по скорости реакции составляет от нескольких десятков минут до нескольких часов. Мы предполагаем, что анализ данных социальных сетей с помощью алгоритмов машинного обучения позволит фиксировать факт того, что то или иное ЧС произошло наиболее оперативно. В свою очередь, оперативное обнаружение таких ситуаций позволит оперативно принимать решения как органам местной/региональной власти, так ипростым людям. В частностиии, информация такого рода может быть полезна при планировании маршрута передвижений по затронутой территории.

Помимо масштабных ЧС, так или иначе затрагивающих город ил страну целиком, существует ещё один тип ЧС - локальные. Такие ситуации, как правило, затрагивают как отдельных людей, так и небольшие группы, но в целом они не влияютна остальных жителей региона. Локальные ЧС иногда вообще не освящаются ни в СМИ, ни в официальных пресс-релизах органов власти. Что вполне логично, нет смысла рассказывать о ситуациях которые не влияют на большое количество людей. Однако информация о ЧС такого масштаба может быть полезна, например, при расчете ценовых планок аренды ил продажи жилья в том или ином районе. В данном кейсе анализ данных социальных сетей является практически единственным способом получить информацию о факте проишествия.

Согласно исследованию [1], преобладающими типами контента, который пользователи публикуют в соцсетях, являются изображения и текст. Притом доля изображений в несколько раз больше доли текста. Однако в данной работе мы сфокусируемся на работе с текстовыми данными. Такой выбор сделан, в частности, из-за большого количества исследований, касающихся анализа текстовых данных социальных сетей. В работе будет представлен сравнительный анализ существующих алгоритмов ма-

шинного обучения для задачи классификации текстов и Named Entity Recognition (извлечения дополнительной информации из текста, в частности, географического местоположения). Для тренировки будут использованы корпуса текстов из социальной сети Твиттер на русском и английском языке. Анализ будет проведен по таким характеристикам как точность, вероятность ошибки и производительность. Последняя характеристика важна, т.к мы рассматриваем алгоритмы машинного обучения для их применения в вполне практической задаче анализа текстов в реальном времени. А как известно, объем данных, генерируемый пользователями того же Твиттера за одну секунду огромен.

С целью демонстрации возможностей алгоритма, выбранного лучшим на основе анализа, мы рассмотрим создание демонстрационного приложения, анализирующего поток данных из Twitter real-time streaming API, и отрисовывающего на основе анализа полученныыых данных тепловую карту черезвычайных ситуаций.

2. Обзор литературы

На сегодня научое сообщество произвело большой объем исследований на тему извлечения информации о черезвычайных проишествиях из потоков социальных медиа. Информативность данного канала была показана в [2]. Обьективно наиболее полные сравнения различных алгоритмов машинного обучения для задачи классификации twitter-специфичного текста были представлены в работах [3, 4]. В нашей работе мы задействуем как традиционно используемые алгоритмы, представленные в данных работах, так и наиболее производительные методы на основе нейронных сетей, разработанные в течении последних нескольких лет. Эффективность таких методов, в частности сверточных нейронных сетей, в задачах бинарной классификации сообщений из Twitter показана в работе [5].

Так как одной из самых важных составляющих сообщения о черезвычайном проишествии является локация этого проишествия, на важно уметь получать из твитов их геопозицию. И вследствии крайне малого числа твитов, снабжаемых геопозицией как частью метаданных(всего таких сообщений около 2-х процентов от общего числа твитов), нам нужно извлекать максимум информации о геопозиции из самого текста. Для этого мы используем технику Named Entity Recognition. Фундаментальное описание одного из алгоритмов поиска именованных сущностей было представлено в работе исследователей из группы разработки фреймворка Stanford CoreNLP - [6]. Мы используем в том числе и эту технику для улучшения качества определения геопозиции в нашей работе.

Полноценные фреймворки обнаружения черезвычайных событий на основе данных социальных медиа и последующего применения к ним алгоритмов Named Entity Recognition, продемонстрированы в работах [7, 8]. И как было замечено в этих работах, нельзя не рассматривать классический keyword matching в качестве способа обнаружения сообщений о черезвычайных ситуациях. Наиболее полно эта тема раскрыта в исследовании [9].

В нашем случае, наша работа будет представлять собой сравнение алгоритмов классификации текста: от простого keyword based matching до сверточных и рекуррентных нейронных сетей, в кейсе задачи бинарной классификации Twitter сообщений. Для полноты эксперимента мы используем наборы данных на русском и английском языке, такие как CloudFlower10k, CrisisLex и некоторые другие. Для задачи уточнения местоположения мы проведем сравнение различных алгоритмов извлечения локации из текста, как основанные на машинном обучении, так и на статистических методах на словарях.

Список литературы

[1] Olga V. Muronets, Content of Social Networks: Trends and Patterns.

[2] Miles Osborne and Mark Dredze. 2014. Facebook, Twitter and Google Plus for Breaking News: Is There a Winner?. In ICWSM

[3] Atefeh F., and Khreich W. (2015), A Survey of Techniques for Event Detection in Twitter, Computational Intelligence, 31, 132–164, doi: 10.1111/coin.12017

[4] Alfredo Cobo, Denis Parra, Jaime Navón: Identifying Relevant Messages in a Twitter-based Citizen Channel for Natural Disaster Situations. CoRR abs/1503.05784 (2015)

[5] Tien Nguyen, Dat & Ali Al Mannai, Kamela & Joty, Shafiq & Sajjad, Hassan & Imran, Muhammad & Mitra, Prasenjit. (2016). Rapid Classification of Crisis-Related Data on Social Networks using Convolutional Neural Networks.

[6] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363-370. http://nlp.stanford.edu/ manning/papers/gibbscrf3.pdf

[7] Emadi, Noora Al et al. "QT2S: A System for Monitoring Road Traffic Via Fine Grounding of Tweets." ICWSM (2017).

[8] Devyatkin D., Shelmanov A. (2017) Text Processing Framework for Emergency Event Detection in the Arctic Zone. In: Kalinichenko L., Kuznetsov S., Manolopoulos Y. (eds) Data Analytics and Management in Data Intensive Domains. DAMDID/RCDL 2016. Communications in Computer and Information Science, vol 706. Springer, Cham

[9] To, Hien & Agrawal, Sumeet & Ho Kim, Seon & Shahabi, Cyrus. (2017). On Identifying Disaster-Related Tweets: Matching-Based or Learning-Based?. 10.1109/BigMM.2017.82.