Механизм предварительной обработки текста перед анализом настроений

М. М. Аббаси
Кафедра теоретических основ информатики
Удмуртский государственный университет
Ижевск, Россия
e-mail: mohsinmanshad@gmail.com

А. П. Бельтюков
Кафедра теоретических основ информатики
Удмуртский государственный университет
Ижевск, Россия
e-mail: belt.udsu@mail.ru

Аннотация[1]

В этой статье исследуются способы предварительной обработки текста перед применением методов идентификации и анализа эмоций, выраженных в этом тексте. Замечено, что во всех методах машинного обучения предварительная обработка данных является очень важной фазой. Однако в анализе чувств этому не уделялось должного внимания. В этой статье мы анализируем работу разных исследователей, в области предварительной обработки текста. Мы покажем важность такой обработки, необходимые для неё факторы или механизмы. Для различных типов текстов покажем, как тип и назначение текста влияет на процесс его предварительной обработки. В разделе выводов мы обобщаем проанализированный нами материал и делаем некоторые заключения о предварительной обработке текста при анализе эмоций.

1. Введение

Анализ настроений – это тема растущего интереса. Одна из его основных задач – выявить эмоциональную полярность предложений и выяснить, выражают ли они положительные, отрицательные или нейтральные эмоции [1]. Со временем исследователи начали не только интересоваться полярностью, но и стали идентифицировать эмоциональный статус предложения, такой как гнев, грусть, волнение и т. д. в соответствии с различными классификациями [2,3]. Это процесс использования текстовой аналитики для получения полезной информации, связанной с мнениями. Часто анализ настроений проводится по данным, которые собираются из Интернета и с различных платформ социальных сетей. Для того, чтобы текст можно было анализировать, данные должны быть в формате, который делает обработку эффективной.

В основном процесс анализа чувств делится на три фазы. Первый этап – предварительная обработка текста с последующим применением методов анализа и алгоритмов анализа на подготовленном тексте. Эта фаза включает в себя идентификацию, классификацию, анализ чувств и т. д. Результаты, полученные на этом этапе, далее обрабатываются в понятной форме, прежде чем представлять их, как показано на рисунок 1.

Рисунок 1. Показывает процесс предварительной обработки текста

Ан
качества и способа их организации. Алгоритмы машинного обучения обучаются на данных. Очень важно поставлять им нужные данные для решения

Труды Шестой всероссийской научной конференции "Информационные технологии интеллектуальной поддержки принятия решений", 28-31 мая, Уфа-Ставрополь, Россия, 2018

проблемы. Даже если данные пригодны, то необходимо еще раз убедиться в том, что они находятся в нужном масштабе, формате, а также в том, что они содержат значимые части. Данные всегда «загрязняются», и их всегда нужно отфильтровывать или добавлять отсутствующие значения. Важность предварительной обработки данных обсуждается в следующем разделе. Процесс предварительной обработки данных включает в себя различные этапы. Вот некоторые из них;

Выбор данных
Он заключается в выборе подмножества всех доступных данных. Всегда есть желание включить все данные, поскольку некоторые считают, что чем больше данных, тем лучше результаты. Однако это не так в большинстве случаев. Всегда есть символы или слова, которые можно отбросить из данных. Нам нужно иметь правильное понимание текста или данных и цель его анализа. Нам нужно понять формат, в котором доступны данные, отсутствующие данные и избыточность доступных данных. Все эти факторы влияют на выбор методологии.

Предварительная обработка данных
Существует три основных этапа предварительной обработки данных: форматирование, очистка и выборка.
Форматирование включает в себя изменение формата данных или текста. Текст может быть в форме, не соответствующей или не совместимой с нашей программой. Он должен быть исправлен. Например, данные могут быть в реляционной базе данных или данные могут находиться в собственном формате файла, а мы хотели бы их иметь в реляционной базе данных или текстовом файле.

Очистка включает удаление или исправление некоторых недостающих данных. Возможно, что данные могут содержать некоторые неполные экземпляры, а не все данные, необходимые для обработки. Возможно, эти экземпляры необходимо удалить. Кроме того, в некоторых атрибутах, которые необходимо полностью удалить из данных, может быть информация.
Сэмплирование подразумевает меньшую репрезентативную выборку данных, которая может быть обработана намного быстрее для изучения и прототипирования решений прежде, чем будет рассмотрен весь набор данных.

Существует несколько способов оценить важность каждого набора данных. Важным среди них является использование частоты появления набора данных в наборе документов, описываемых с помощью уравнения.

FP = FF * LOG (N / DF)

Где FF - частота появления набора данных. N указывает количество документов, а DF - количество документов, содержащих этот набор данных [4].

Преобразование данных
Это важный этап подготовки текста перед его анализом. Текст, в соответствии с его типом, обладает определенной структурой, которая в некоторых случаях не позволяет алгоритмам или машине правильно ее анализировать.

2. Важность предварительной обработки

1. Было отмечено, что текст может обладать определенными символами, которые отвлекают процессы анализа текста.

2. Для эффективного и эффективного анализа текста необходимо обратить внимание на каждое слово, его характеристики и его группу в части речи.

3. Каждый алгоритм или программа предусматривает свою собственную структуру и конструкцию для анализируемого текста. Текст должен быть подготовлен соответствующим образом.
4. В случае использования искусственного интеллекта программы также обрабатывают определенную структуру текста и анализируют ее.

5. Эффективным будет лишь анализ надлежащим образом структурированного и подготовленного текста.

6. Необходимо понять цель анализа текста и типы алгоритмов, которые будут применяться на следующем этапе, таких как классификация, SVM и т. д.

7. Алгоритмы машинного обучения обучаются на основе данных. Очень важно, чтобы им подавались правильные данные для проблемы, которую требуется решить.

Таким образом, для предварительной обработки данных необходимо следующее:

Требуется знать тип и тему текста.
Требования к оформлению текста для анализа. В случае анализа настроений полное предложение необходимо изучать вместе как существующее соотношение между словами в предложении, затем – между предложениями и так далее.

Тексты, собранные из окружающего мира, обычно непоследовательны, неполны и зашумлены. В них могут отсутствовать определенные элементы, представляющие интерес. Зашумленные тексты

содержат выбросы, ошибки, несогласованные элементы могут содержать расхождения в кодировании.

Процесс подготовки данных в случае анализа настроений отличается от категоризации текста. В случае категоризации текста цель состоит в том, чтобы определить тему документа, сравнив ее с некоторыми уже определенными темами. В анализе настроений ассоциации, зависимости и отношения внутри текста требуют идентификации с использованием различных алгоритмов и методов сопоставления с шаблонами. Например, отношение между прилагательным и отрицанием может изменить полярность настроения или его интенсивность. В случае добавления или удаления отрицания может измениться смысл предложения и текста.

3. Связанные работы

Фаза предварительной обработки преобразует исходные текстовые данные в структуру, где идентифицируются наиболее важные текстовые функции. Эффективный препроцессор представляет документ эффективно с точки зрения, как пространства (для хранения документа), так и времени (для обработки запросов на получение) и поддержания хорошей производительности поиска (точность и время реакции) [5]. Эта фаза представляет собой сложный процесс, в котором каждый документ представлен избранным набором его индексных терминов. Первым шагом во время предварительной обработки является удаление этих стоп-слов, которые не очень важны для преследуемой цели [6]. Например, G. Salton в 1989 использовал список слов SMART stop word [7]. В алгоритме Портера-Стеммера [8] слова преобразуются в их основы, причина этого – в том, что слова с одним и тем же корнем в основном описывают те же или относительно близкие понятия в тексте.

В 1988 году этим же автором вводится понятие веса термина и связывается с каждым термином как важный показатель [9]. Три основных компонента, которые влияют на важность термина в документе – это коэффициент временной частоты (TF), коэффициент обратной частоты документа (IDF) и нормализация длины документа [10]. Авторы Cristian Moral и др. [11] из Universidad Politécnica de Madrid, Испания, провели оценку алгоритмов генерации, используемых в приложениях поиска информации.

4. Обсуждение

Информация, написанная в электронных письмах, документах, блогах или других файлах данных, классифицируется как структурированные или неструктурированные данные. Структурные данные хорошо организованы и согласованы, легко подвержены анализу, например, в таблицах Excel, таблицах, базах данных, тогда как неструктурированные данные находятся в свободной форме и их неудобно анализировать. Они нуждаются в специальной обработке, чтобы преобразовать их в понятную форму, например, таковы данные с вебсайтов блогов и социальных сетей. Сегодня основным источником данных и информации является Интернет, где высокий процент данных доступен в неструктурированной форме и требует особого внимания для понимания.

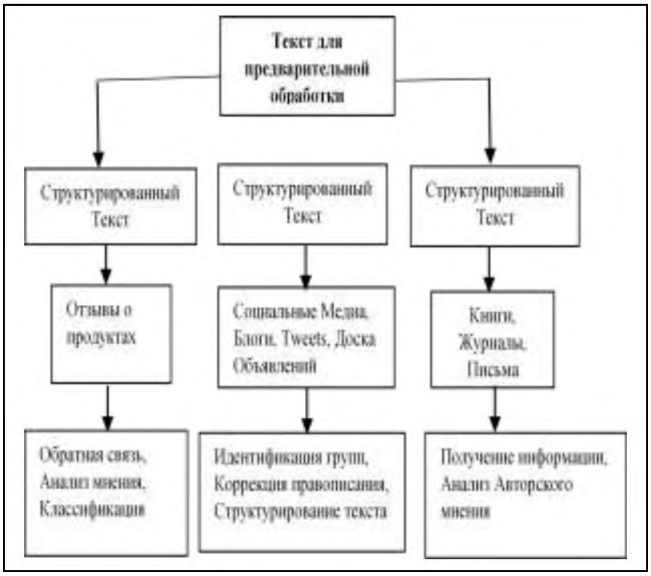

Рисунок 2 Связь между структурой текста и его обработкой

Процесс предварительной обработки текста зависит от типа текста. Каждый тип текста имеет свою структуру и цель. Например, в структурированном виде тексте получаются у пользователя в специальном формате при заполнении вопросника и даже с помощью проверки орфографии. Шансы на орфографии или другие ошибки в этом случае очень малы, тогда как в случае социальных сетей или блогов нет надлежащего формата для написания текста. Пользователи обычно пишут на неформальном языке с дефектами орфографии, пунктуациями, орфографическими ошибками, сленгами, URL-адресами, тегами HTML и многими другими особенностями. Здесь важность и процесс предварительной обработки данных сложнее. В случае полуструктурного текста, такого как книги или письмо, информация представляется в надлежащем формате с гораздо меньшими шансами на ошибки.

Существует три подхода к анализу настроений: методы машинного обучения, методы на основе лексики и лингвистический анализ. Методы машинного обучения заключаются в подготовке алгоритма, на основе известной классификации и известных функций, присутствующих в выбранных наборах текстов, а затем – в проверке на других наборах текстов того, может ли подготовленный

таким образом алгоритм обнаруживать правильные функции и давать правильную классификацию. Метод на основе лексики строится на базе предопределенного списка или совокупности слов с определенной полярностью. Построенный алгоритм ищет эти слова, подсчитывает их или оценивает их вес и измеряет общую полярность текста. Наконец, лингвистический подход использует синтаксические характеристики слов или фраз, отрицание и структуру текста для определения ориентации текста. Этот подход обычно сочетается с методом на основе лексики [12].

Авторы Giulio A. и др. в своем недавнем исследовании по предварительной обработке данных твитов для анализа настроений указали на важность нормализации данных, удаляя несущественные элементы, такие как хэштеги, URL и все знаки пунктуации перед использованием программного обеспечения Weka. Затем они удаляли гласные, повторяющиеся последовательно по крайней мере три раза, и выполняли некоторую замену для исправления грамматики, а затем преобразовали весь текст в нижний регистр. Они также определяли вероятности настроений и давали значения тегам. Оценивалась вероятность положительных и отрицательных конструкций. Все отрицательные конструкции, такие как «не могут», «нет», «никогда» и т. д. заменяются на «нет». Они отметили, что все эти методы обеспечили значительное улучшение характеристик классификатора. Некоторые из методов просто удаляли шум в данных, в то время как другие повысили значимость некоторых понятий, уменьшив сходные термины до их самого основного значения [13].

5. Вывод

Процесс предварительной обработки данных или текста зависит от требований и цели приложения. В некоторых приложениях достаточно провести основное форматирование и очистку данных приложений, тогда как в других приложениях требуются более сложные процессы, такие как агрегация данных и их дискретизация.

На тестовом наборе данных эксперимент показывает до 68% повышения эффективности для данных, генерируемых после предварительной обработки, по сравнению с данными, которые генерируются без нее [14].

Методы машинного обучения улучшаются в категоризации, а затем в классификации чувств. Это связано с характером упрямого текста, который требует большего понимания текста [15]. Классификаторы машинного обучения, такие как наивные байесовские системы, максимальная энтропия и опорная векторная машина (SVM), используются в [15] для классификации чувств для достижения точности, которая колеблется от 75% до 83%, по сравнению с 90% -ной точностью или выше в зависимости от темы категоризации.

Предварительная обработка текста является очень важной фазой во всех приложениях интеллектуального анализа данных и в частности в анализе настроений. Тем не менее, немногие работы были специально посвящены пониманию роли каждого из основных методов предварительной обработки, которые часто применяются к текстовым данным.

Фаза предварительной обработки зависит от характера анализируемого текста. В случае, если текст взят из социальных сетей, он часто не структурирован содержит орфографические ошибки и ошибочные символы. В противоположность этому в случаях писем, книг, официальной информации текст правильно структурирован и не требует глубокой предварительной обработки.

Использование проверки орфографии может помочь в эффективной предварительной обработке данных. Важная информация для токенизации данных может быть извлечена после удаления некоторых символов, таких как скобки, знаки пунктуации. К этому относится и разделение предложений на слова. Частоты слов можно использовать для категоризации слов по их важности.
Было замечено, что работа по предварительной обработке не очень велика. Это зависит от выбора данных или текста исследователем и цели исследования. В большинстве случаев основной целью исследования было выявление и классификация чувств и экспериментов, выбор определенного типа текста. Шансы ошибок были уменьшены, но в большинстве случаев показатели классификации составляли не более 70%. Наша будущая цель – построить полную систему для анализа текста разных типов. Предварительная обработка текста является важной фазой, и результаты всей работы зависят от данных, полученных на этом этапе.

Список используемых источников

1. Pang B. Lee L. Opinion mining and sentiment analysis// Foundations and trends in information. 2008. P. 1-135.
2. Cambria E. Olsher D. Rajagopal D. A common and common-sense knowledge base for cognition-driven sentiment analysis. // Proc. of the 28 AAAI conference on artificial intelligence. 2014.
3. Poria S. Cambria E. Winterstein G. Huang G.B. Dependency based rules for concept-level sentiment analysis. Knowledge-Based Systems. 2014. P. 45–63.

4. Emma H. Xiaohui L. Yong S. The Role of Text Pre-processing in Sentiment Analysis // Proc. of the International Conference on Information Technology and Quantitative Management. 2013. P. 26 – 32.

5. Pritam C.G. Patil L.H. Chaudhari P.M. Preprocessing Techniques in Text Categorization. /National Conference on Innovative Paradigms in Engineering & Technology (NCIPET-2013) //Proceedings published by International Journal of Computer Applications (IJCA), 2013.

6. Xue X. Zhou Z. Distributional Features for Text Categorization// IEEE Transactions on Knowledge and Data Engineering, 2019. Vol. 21 №.3. P. 428-442.

7. The Transformation, Analysis, and Retrieval of Information by Computer // G. Salton- Pennsylvania, Addison Wesley, Reading, 1989.

8. Porter M. An algorithm for suffix stripping, Program. 1980. Vol. 14 №.3. P. 130–137.

9. Salton G. Buckley C. Term weighting approaches in automatic text retrieval. Information Processing and Management .1988. Vol. 24 №.5. P. 513- 523.

10. Karbasi S. Boughanem M. Document length normalization using effective level of term frequency in large collections //Advances in Information Retrieval, Lecture Notes in Computer Science, Springer Berlin / Heidelberg. 2006. Vol. 3. P. 72-83.

11. Moral C. Antonio A. Imbert R. Ramírez J. A survey of stemming algorithms in information retrieval/ Information Research. 19(1) paper 605. 2014.

12. Blinov P. D. Klekovkina M. V. Kotelnikov E. V. Pestov O. A. Research of lexical approach and machine learning methods for sentiment analysis. Vyatka State Humanities University, Kirov, Russia, 2013.

13. Giulio A. Laura F. Tomaso F. Paolo F. Eleonora I. Federico M. Stefano M. A Comparison between Preprocessing Techniques for Sentiment Analysis in Twitter. Dipartimento di Ingegneria dell'Informazione Parco Area delle Scienze 181/A, 43124 Parma, Italy, 2017.

14. Arjun S. N. Ananthu P. K. Naveen C. Dr. Balasubramani. Survey on Pre-Processing Techniques for Text Mining. International Journal of Engineering and Computer Science. ISSN: 2319-7242. June 2016. Vol. 5. P. 16875-16879.

15. Abbasi M.M. Beltiukov A.P. Analysis of sentiment and emotion from text written in Russian language. 5th All Russian Conference on Information technology for intelligent decision making support, Ufa, Russian Federation, May 16-19, 2017.