УДК 621.316:004.85

ОБРАБОТКА ОТЧЕТОВ О ТЕХНОЛОГИЧЕСКИХ НАРУШЕНИЯХ С ИСПОЛЬЗОВАНИЕМ МАШИННОГО ОБУЧЕНИЯ

Воронин В.А., научный сотрудник
Кузбасский государственный технический университет
имени Т.Ф. Горбачева
г. Кемерово

Введение и постановка проблемы. Эксплуатация электросетевого комплекса связана с накоплением больших объемов информации как в числовой (измерения электрических величин), так и в текстовой форме (отчеты, заключения, проекты и т.д.). Статистическая обработка накопленных данных необходима для выявления факторов и закономерностей, определяющих эффективность работы всего электросетевого комплекса.

Особенную важность имеют данные, связанные с надежностью электроснабжения. На основе анализа этой информации могут быть выявлены слабые места в электрической сети, внесены корректировки в планы ремонтов и инвестиционную программу предприятия. Наибольшую сложность представляет выявление и классификация причин каждого аварийного отключения. Вся необходимая для анализа информация может быть получена из отчетов о технологических нарушениях (ТН) в электрических сетях. Однако причины ТН описаны текстом в свободной форме, что затрудняет их обработку и классификацию.

В последние годы большую популярность получили методы машинного обучения и в частности методы обработки естественного языка (NLP). Например, в работе [1] рассматривается обработка диспетчерских команд с помощью методов NLP для автоматического формирования бланков переключений. В [2] NLP используется для классификации аварийных событий в электрических сетях, а в [3] текстовая информация обрабатывается для оценки обеспечения гарантии качества на ядерных установках. Методы NLP также могут быть использованы для обработки и анализа большого объема текстов научных работ с целью выявления основных трендов развития в рассматриваемой предметной области [4, 5].

Целью настоящей работы является оценка эффективности применения методов NLP для классификации аварийных событий в электросетевом комплексе.

Методология. В качестве входных данных использован отчет о ТН в электрических сетях Кузбасской энергосистеме за один год, включающий в себя 5822 записи. В отчет входят данные по времени возникновения и завершения ТН, описание ТН, причины возникновения ТН, число обесточенных потребителей. В данной работе анализируется графа «Причины возникновения

ТН», представляющая собой текстовое описание обстоятельств ТН в свободной форме. Несколько примеров записей:

1) Попадание птицы на ошиновку ТСН-1-160 (запитан от шинного моста 6 кВ Т-1-25).
2) Пробой изолятора оп. №32.
3) На опоре №109 лопнула вязка провода.

Описания ТН часто являются избыточными для целей классификации. В связи с этим требуется их нормализация. Вследствие отсутствия установленного шаблона для описаний причин ТН, одинаковые аварии нередко описываются разными словами, что затрудняет кластеризацию массива записей.

В результате анализа описаний ТН были выделены следующие характерные категории причин ТН: обрыв провода; повреждение кабеля; повреждение изолятора; повреждение электрооборудования подстанции; погодные условия; попадание посторонних объектов на токоведущие части; пробой изоляции; действия третьих лиц; схлест проводов и др. Для целей исследования все записи отчета были промаркированы выделенными категориями вручную, что дает возможность использования методов машинного обучения с учителем.

Обработка данных и разработка моделей машинного обучения выполнена на языке программирования Python. В ходе обработки выполнена токенизация и нормализация текста с помощью библиотеки nltk. Нормализация записей включала в себя выравнивание регистра, удаление стоп-слов, знаков пунктуации, а также всех числовых данных. После чего выполнена векторизация полученных массивов текстовых данных методом TF-IDF с помощью библиотеки scikit-learn.

Для визуализации данных использован метод t-SNE (стохастическое вложение соседей с t-распределением). Для кластеризации использован метод LDA (латентное размещение Дирихле). Классификация выполнена с использованием методов логистической регрессии, стохастического градиентного спуска, наивного байесовского классификатора, опорных векторов и случайного леса. Все перечисленные методы реализованы в библиотеке scikit-learn.

Результаты и обсуждение. На рис. 1 в виде трехмерного графика представлена визуализация анализируемого массива описаний ТН, полученная с помощью метода понижения размерности t-SNE. Каждая точка на рис. 1 представляет собой одну запись в отчете ТН. Цвета точек соответствуют категориям ТН, определенным вручную.

Как следует из рис. 1, описания ТН формируют разреженные частично перекрывающиеся кластеры. Для их выделения из массива описаний ТН использован метод LDA. На рис. 2 показаны наиболее часто встречающиеся слова в выделенных кластерах (показаны только кластеры, включающие в себя не менее 25 повторяющихся слов).

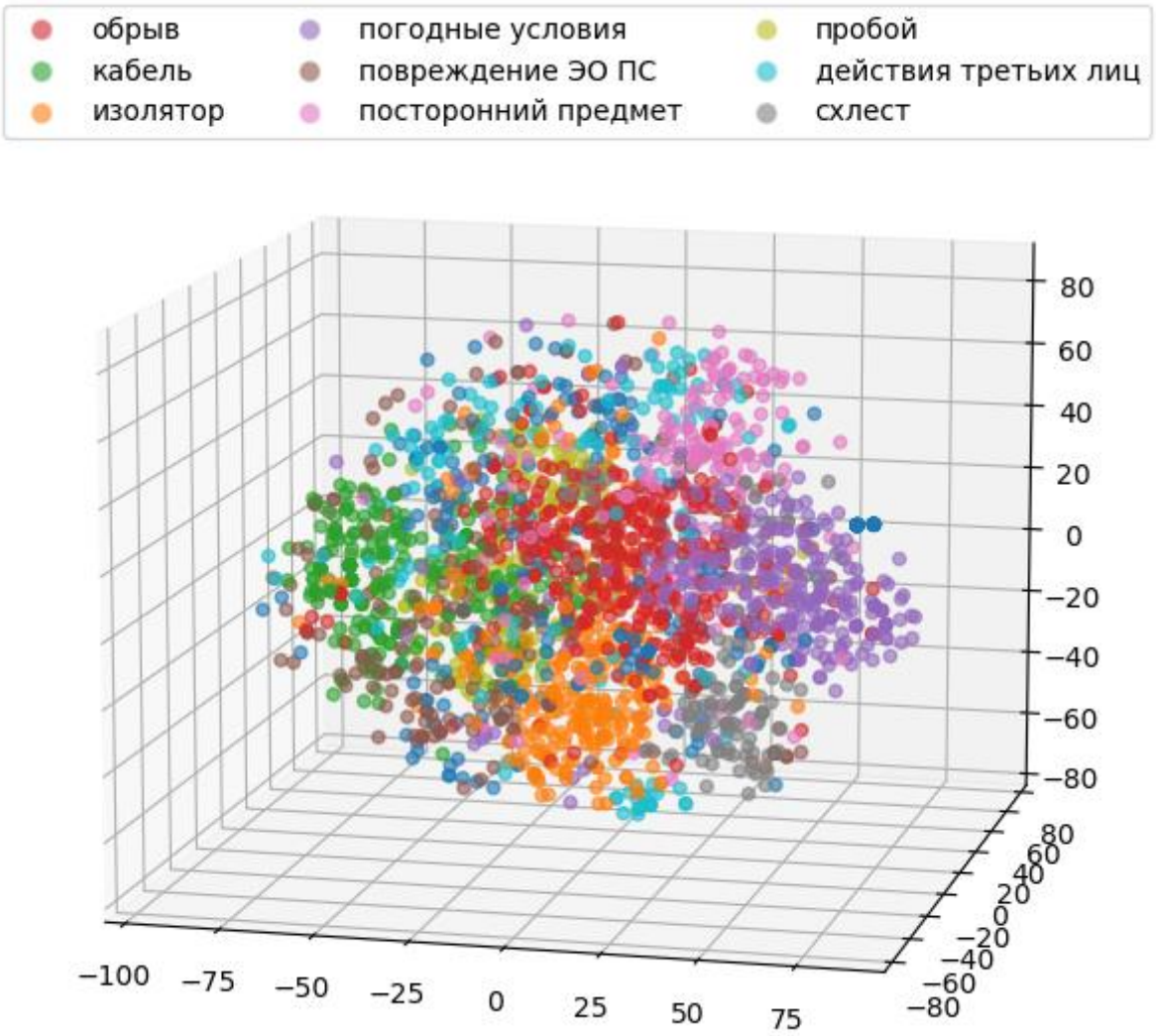

Рис. 1. Визуализация описаний ТН

Анализируя рис. 2, можно заметить, что кластер 7, 35 и 44 образуют категорию «обрыв провода», кластеры 22 и 25 – «погодные условия» и т.д. Используя результаты кластеризация выполнена маркировка записей массива описаний ТН. Оценка точности классификации определена на основе сравнения с метками категорий, предварительно присвоенных каждой записи вручную (табл. 1).

Как следует из табл. 1, результаты классификации характеризуются низким значением полноты, что говорит о том, что данные кластеры охватывают только часть записей из рассматриваемых категорий. Кроме того, не для всех категорий записей удалось выделить соответствующие кластеры.

Более точным способом классификации является использование методов машинного обучения с учителем. Однако в данном случае требуется предварительная подготовка массива маркированных данных, на котором будет проводиться обучение модели. В табл. 2 представлены средневзвешенные по всем категориям оценки результатов классификации различными методами на тестовом наборе записей.

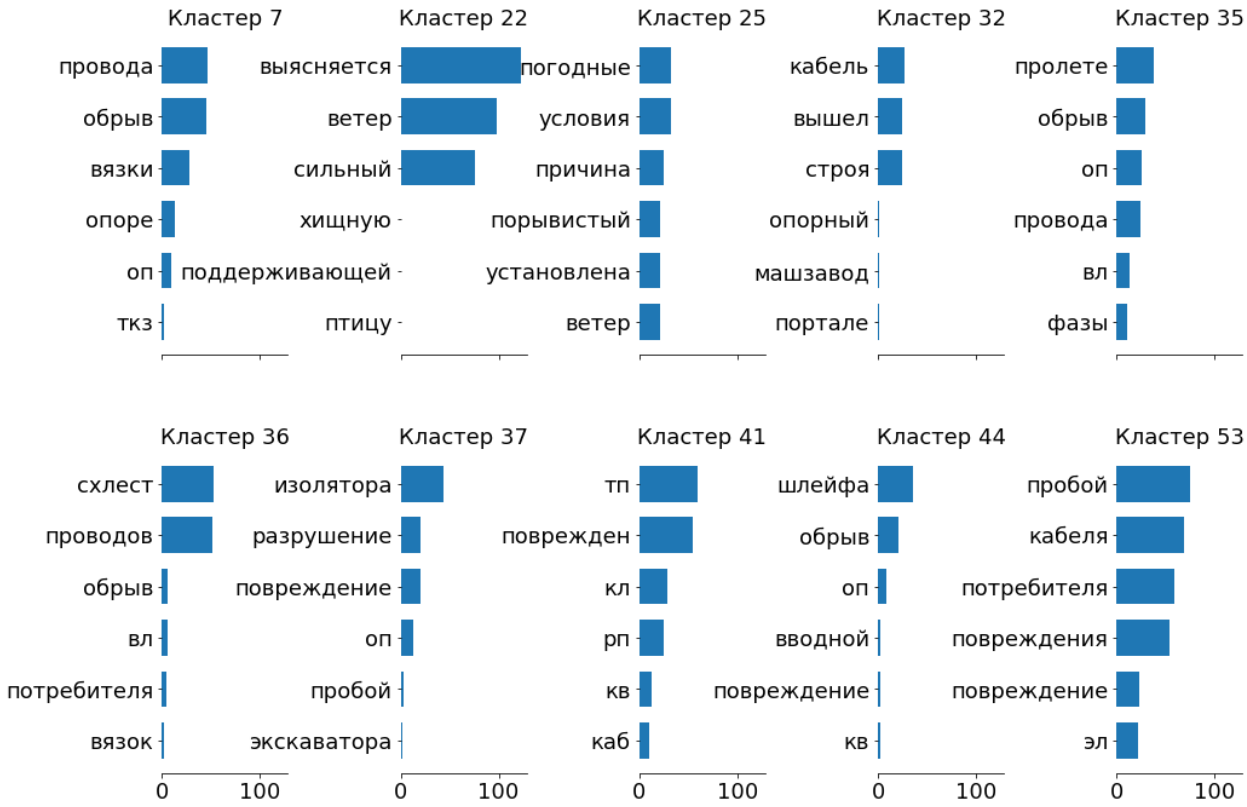

Рис. 2. Наиболее часто встречающие слова в кластерах

Таблица 1

Оценка точности классификации

Категория	Точность	Полнота	F-мера
обрыв провода	0,85	0,45	0,59
повреждение изолятора	0,80	0,19	0,31
повреждение кабеля	0,83	0,13	0,23
погодные условия	0,74	0,76	0,75
схлест проводов	0,78	0,52	0,62
пробой изоляции	0,55	0,61	0,58

Таблица 2

Оценка точности классификации

Методы	Точность	Полнота	F-мера
Опорных векторов	0,89	0,89	0,89
Случайный лес	0,88	0,89	0,88
Логистическая регрессия	0,89	0,90	0,89
Стохастический градиентный спуск	0,89	0,90	0,89
Наивный байесовский классификатор	0,78	0,82	0,78

Рассмотренные методы обучения с учителем показали схожие результаты классификации. Из сравнения табл. 1 и табл. 2 следует, что методы обучения с учителем намного более эффективны по сравнению с методами обучения без учителя для классификации описаний ТН.

Заключение. В результате выполнения данной работы выявлено, что методы машинного обучения могут быть использованы для автоматизации обработки текстовых описаний причин аварийных отключений в отчетах о ТН в электрических сетях.

Наибольшую эффективность показали методы обучения с учителем. Поэтому для создания точной модели классификации описаний ТН следует предварительно подготовить промаркированную репрезентативную выборку описаний ТН для обучения модели.

Список литературы:

1. Research on programmed operation of power grid equipment based on Natural Language Processing / J. Li [и др.] // 2020 International Conference on Electrical Engineering and Control Technologies (CEECT) 2020 International Conference on Electrical Engineering and Control Technologies (CEECT). – 2020. – С. 1-5.
2. Identification Technology of Grid Monitoring Alarm Event Based on Natural Language Processing and Deep Learning in China / Z. Bai [et al.] // Energies. – 2019. – Vol. 12. – № 17. – P. 3258.
3. Natural Language Process: A New Kind of Nuclear Quality Assurance Management Tool : International Youth Nuclear Congress 2016, IYNC2016, 24-30 July 2016, Hangzhou, China / Y. Guan [et al.] // Energy Procedia. – 2017. – Vol. 127. – P. 201-219.
4. Копайгородский А.Н. Семантический анализ Big Data в задаче прогнозирования инновационного развития энергетической инфраструктуры РФ / А.Н. Копайгородский, Е.П. Хайруллина, И.И. Хайруллин. – Автономная некоммерческая организация в области информационных технологий «Научно-исследовательский центр физико-технической информатики», 2020. – С. 199-203.
5. Михеев А.В. Возможности анализа текстовой информации научно-технологической направленности для обоснования инновационного развития энергетики / А.В. Михеев // Информационные и математические технологии в науке и управлении. – 2017. – № 4 (8). – С. 166-176.

Информация об авторах:
Воронин Вячеслав Андреевич, старший преподаватель кафедры ЭГПП, научный сотрудник НИЛ ЦТПМСК, КузГТУ, 650000, г. Кемерово, ул. Весенняя, д. 28, voroninva@kuzstu.ru