ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ КОРОТКИХ СООБЩЕНИЙ ДЛЯ УПРАВЛЕНИЯ ДОРОЖНО-ТРАНСПОРТНОЙ ИНФРАСТРУКТУРОЙ*

INTELLIGENT ANALYSIS OF SHORT MESSAGES FOR ROAD INFRASTRUCTURE MANAGEMENT

A. Lyapin

Summary. In the article analyzes the problems of using methods of data mining by specialists who do not have high qualifications in the field of information technology. This aspect creates difficulties in training classifiers, the selection of trained data sets and their subsequent use, so the development of self-learning methods is an urgent task. Purpose. The purpose of the work is to develop an ensemble of algorithms for intelligent data processing, which allow to classify text messages without prior training of the classifier. Novelty. The model assumes a deep integration of road transport and information technologies on the basis of a single network platform that includes a traffic management system, an alert system for emergency situations, an intelligent system for analyzing geotagged user messages, a Web interface, etc. Practical relevance. The intellectual system represents a mechanism for the classification of short text messages for the timely notification of road users about emergency situations and support for decision-making by service operators.

Keywords: data mining, Smart city, k-nearest neighbors, abnormal road situation.

Ляпин Артур Мансурович
Аспирант, Пензенский государственный университет
lyapinartur@gmail.com

Аннотация. в статье анализируются проблемы использования методов интеллектуального анализа данных специалистами, не имеющими высокой квалификации в сфере информационных технологий. Данный аспект создаёт трудности в обучении классификаторов, подборе обучаемых наборов данных и последующем их использовании, поэтому разработка самообучаемых методов является актуальной задачей. Целью работы является разработка ансамбля алгоритмов интеллектуальной обработки данных, позволяющих классифицировать текстовые сообщения без предварительного обучения классификатора Новизна: модель предполагает глубокую интеграцию дорожно-транспортных и информационных технологий на базе единой сетевой платформы, которая включает систему управления дорожным движением, систему оповещения о внештатных ситуациях, интеллектуальную систему анализа геотегированных сообщений пользователей, Web интерфейс и т. д. Практическая значимость: интеллектуальная система представляет механизм классификации коротких текстовых сообщений для своевременного оповещения участников дорожного движения о внештатных ситуациях и поддержки принятия решения операторов обслуживающих служб.

Ключевые слова: интеллектуальная обработка данных, Smart city, метод k-ближайших соседей, нештатная дорожная ситуация.

Введение

Технологии создания компонент интеллектуальной и безопасной городской среды «Smart & Save City» [1] активно развиваются и внедряется практически во все сферы жизнедеятельности человека, связанные со здоровьем, безопасностью, отдыхом, работой и транспортом. Для оценки и анализа проблем безопасности целесообразно использовать информацию, сгенерированную самими людьми в социальных сетях и медиапространстве сети Интернет. Именно здесь многочисленные информационные источники являются наиболее доступными, открытыми и актуальными.

В статье рассматривается предложенный подход для интеллектуального анализа данных из социальных сетей Twitter и «Вконтакте». Twitter представляет собой сеть микроблогов, где пользователи обмениваются короткими сообщениями — твитами. В Twitter зарегистрировано более 600 миллионов пользователей по всему миру и котором ежедневно размещается примерно 65 миллионов твитов.[1] Сети «Вконтакте» — социальная сеть российского Интернет сегмента, крупнейшая в Европе. Среднесуточная аудитория составляет более 80 миллионов пользователей, а зарегистрировано более 450 миллионов.[2]

Рассмотрим предложенный метод анализа сообщений, на примере сообщений, связанных с безопасностью дорожного движения, которые имеют временные и геотеггированные метки с информацией о координатах местоположения смартфона пользователя, отправившего сообщение, времени отправки. Сообщения, как правило, содержат информацию о дорожном инциденте или описание проблем с дорожно-транспортной инфраструктурой, результатом которых становятся пробки, изменения дорожного трафика, маршрутов движения и другие события.

* Результаты работы получены при финансовой поддержке РФФИ в рамках гранта № 18–07–00975.

Таблица 1. Характеристика классов сообщений

Класс	Описание
Позитивный	Сообщения о хороших дорожных условиях. Указывают на то, что транспортные средства могут передвигаться свободно.
Затор	Сообщения о заторах и пробках.
Препятствие	Сообщения о событиях которые препятствуют нормальному движению транспортных средств.
Инцидент	Сообщения о событиях, в которых люди или транспортные средства получили повреждения.
Прочие	Сообщение не удалось классифицировать. Недостаточно информации, для определения класса события.

Рис. 1. Этапы преобразования данных

Особенностью предлагаемой методологии является использование метода машинного обучения для классификации и кластеризации получаемых данных. В отличие от других подходов, в которых используются ручные классификаторы, процесс анализа полностью автоматизирован. Так как сообщения являются геотегированными и содержат сгенерированный пользователем контент с координатами и временем события, то результаты кластеризации визуализируются на цифровой картографической основе с целью определения текущего состояния дорожной обстановки и краткосрочного прогнозирования изменения ситуации в некоторой области рядом с инцидентом. Это позволит наглядно представить критические области, причину и оценить масштабность происшествия.

Метод классификации событий по информации из текстовых сообщений

Предложенный подход использует открытую информацию для классификации событий дорожного движения, которые влияют на степень загруженности транспортной системы в области мониторинга. Результаты необходимы для оптимизации дорожного трафика и улучшения дорожной обстановки в городе. Методика включает четыре этапа преобразования собираемых коротких сообщений в информацию о дорожной ситуации (Рис. 1):

1. Установка контакта с социальными сетями и сбор данных,
2. Очистка данных,
3. Выбор и классификация событий,
4. Визуализация событий на карте.

На первом этапе в процессе сбора данных устанавливается контакт с источниками. Извлеченные данные сохраняются в необработанном виде. Вместе с сообщениями собирается временная и геопространственная информация, а именно особенности рельефа местности, о которой идет речь в сообщении, географические координаты для фиксации границы участка. Сообщения о дорожном движении и транспортной ситуации в районе исследования сохраняются в базе данных.

Сообщения собираются из открытых сообществ социальных сетей согласно учетным записям пользователям, а также с новостных сайтов. Существует несколько интерфейсов прикладных программ, которые позволяют извлекать сообщения в соответствии с запросами или путем настройки PUSH уведомлений о появлении новых сообщений. Каждый раз, когда размещается новое сообщение в интересующем аккаунте пользователя или на новостном сайте, то оно сохраняется в базы данных.

На этапе обработки сообщения классифицируются по ключевым словам, найденным в тексте сообщения. Определено пять классов дорожных событий: позитивные, заторы, препятствия, инциденты и прочие (Таблица 1).

Подготовка данных к классификации происходит по следующей методике.

1. Специальный скрипт проводит лексический разбор текста, разделяет его на слова и удаляет знаки препинания.
2. Далее удаляются специфичные «стоп-слова». Словарь таких слов составлен с помощью сервиса Ranks NL (https://www.ranks.nl/stopwords/russian).
3. Устанавливается предварительный класс сообщения, чтобы применить метод «K-ближайших соседей». Здесь необходимо иметь набор слов, идентифицирующий классы «Bag of words», который составлен с помощью сервиса RusVectōrēs (http://rusvectores.org/ru).
4. Подсчитывается количество совпадений слова в каждом наборе класса. Класс с наибольшим количеством совпадений присваивается сообщению.

В результате каждое сообщение получает метку класса, а также на выходе получается текст без специальных символов и стоп-слов, т.е. выполняется его «очистка».

На этапе классификации данных по событиям применяется алгоритм машинного обучения «K-ближайших соседей» (KNN), описание которого приведено ниже:

Пусть задана обучающая выборка пар «объект-ответ» $X^m = \{(x_1, y_1), \dots, (x_m, y_m)\}$.

Пусть на множестве объектов задана функция расстояния $p(x, x^{\square})$. Эта функция должна быть достаточно адекватной *моделью сходства* объектов. Чем больше значение этой функции, тем менее схожими являются два объекта $x, x^{\square}$

Для произвольного объекта u расположим объекты обучающей выборки x_i в порядке возрастания расстояний доu:

$$p(u, x_{1,u}) \le p(u, x_{2,u}) \le \dots \le p(u, x_{m,u})$$

где через $x_{i,u}$ обозначается тот объект обучающей выборки, который является i-м соседом объекта u. Аналогичное обозначение введём и для ответа на i-м соседе: $y_{i,u}$. Таким образом, произвольный объект u порождает свою перенумерацию выборки. В наиболее общем виде алгоритм ближайших соседей есть:

$$a(u) = \underset{yY}{argmax} \sum_{i=1}^{m} [x_{i,u} = y] w(i,u)$$

где $w(i,u)$ — заданная *весовая функция*, которая оценивает степень важности i-го соседа для классификации объекта u.[3]

Перед использованием метода KNN сообщения нормализуются и приводятся к числовому виду. Для этого применяется векторная модель частоты слов в сообщениях. Строится словарь, который содержит слова, найденные в выбранных сообщениях, с информацией о частоте слов в каждом. Составление словаря происходит по следующему алгоритму:

1. Задается пустой словарь.

2. Для всех сообщений класса выполняется процесс:
a. Сообщение разделяется на слова,
b. Аналогичное слово ищется в словаре,
c. если слово присутствует в словаре, то его вес увеличивается на 1,
d. если слово не существует, то оно добавляется в словарь с весом 1.

3. Как только в словаре будут собраны все возможные ключевые слова и их частотные значения (частота повторений в сообщениях класса), то выполняется переход к этапу нормализации сообщений.

На этапе нормализации строится векторное представление сообщения по следующему алгоритму:

1. Создается вектор с числом позиций равным длине составленного словаря слов.
2. В каждую позицию вектора записывается количество повторений слова в сообщениях.
3. Если слово из словаря встречается n раз в сообщение, то в векторе на позиции слова сохраняется значение n.
4. Если слово из словаря не встречается в сообщении, то его значение в векторе устанавливается равным 0.

Полученные вектора сообщений далее используются при обучении и тестировании алгоритма KNN. Алгоритм работает с ключами, которыми являются сгенерированные векторы. Алгоритм классифицирует сообщение, об-

рабатывая большое количество образцов и в процессе кластеризации использует Евклидово пространство для вычисления расстояния между экземплярами.

Обучение и тестирование разработанного метода проводится посредством выполнения следующих действий.

1. Обрабатываются сообщения из обучающего набора данных.
2. Вычисляется классификация каждого вектора используя алгоритм KNN.
3. Применяется «десятикратная перекрестная проверка». Основная идея: образцы делятся на десять наборов и каждый раз, когда один набор классифицируется, девять оставшихся наборов считаются обучающими.
4. Сообщению присваивается класс.
5. Присвоенный класс сравнивается с изначально назначенной классификационной меткой.

На этапах обучения и тестирования использовалась выборка из 500 сообщений. Набор сообщений был разделен на два массива:

1) обучающий, который состоит из 300 сообщений,
2) тестовый, который состоит из 200 сообщений.

Обучающая выборка имеет изначально установленную классификационную метку, присвоенную на втором этапе. Эта метка требуется, так как модель KNN учитывает отношение соседних сообщений для классификации тестового образца.

После классификации рассчитываются метрики для оценки эффективности. В качестве показателей эффективности выбраны: точность(precision), полнота(recall) и F-метрика. Показатель точность(precision) можно интерпретировать как долю объектов, названных классификатором положительными и при этом действительно являющимися положительными, а показатель полнота(recall) показывает, какая доля объектов положительного класса из всех объектов положительного класса была найдена.[4] Существует несколько различных способов объединить точность(precision) и полноту(recall) в агрегированный критерий качества. F-мера (в общем случае F_β) является одним из этих способов, и представляет среднее гармоническое precision и recall:

$$F_\beta = \frac{(1+\beta^2) * precision * recall}{(\beta^2 * precision) + recall}$$

β в данном случае определяет вес точности в метрике, и при β=1 это среднее гармоническое (с множителем 2, чтобы в случае precision = 1 и recall = 1 иметь F1=1)

F-мера достигает максимума при полноте и точности, равными единице, и близка к нулю, если один из аргументов близок к нулю.[5]

Как только данные метрики достигают удовлетворительных результатов, модель становится готовой для классификации тестового набора. Метрики также подсчитываются при обработки тестового набора для целей аналитики и статистики.

Визуализация результатов кластеризации

Целью предложенного подхода является выделение и цветовая дифференциация областей на карте с проблемным дорожным трафиком, а также определение причин. Для этого координаты источника сообщения привязываются к цифровой картографической основе и в результате сообщение представляется меткой на карте. Для демонстрации работы метода выбран период наиболее высокой активности дорожного движения, а именно время утреннего часа пик. Исследуемой областью является центральная часть города Пензы. На рисунке 2 показана карта и обработанные сообщения в виде меток. По карте отображены не все сообщения, но можно заметить, что пользователи активно пишут о дорожной ситуации и отмечают координаты проблемных мест. Анализ показал, что проблемные места в основном находятся в центральном и южном районах города. Это обусловлено тем, что большинство офисных помещений находится в центральной части города, а в южной части дорожная сеть давно не реконструировалась и перегружена.

На карте метками в виде красного креста изображены сообщения, классифицированные как аварии и столкновения. Желтым цветом отмечены сообщения о заторах и пробках. Красным отмечены области с препятствиями и сужениями дорог. Можно заметить, что в районах с повышенным трафиком число дорожных инцидентов больше. С другой стороны, если около дорожного инцидента появляются пробки, то повышается плотность траффика.

Зелёным цветом отмечены сообщения, классифицированные как отзывы о благоприятных дорожных условиях. Таких сообщений меньше, чем сообщений о плотном трафике или инцидентах, и располагаются они вне центральной части города, как правило, на дорогах с новым асфальтовым покрытием и низким трафиком. Это объясняется тем, что для экспериментального исследования метода был выбран наиболее напряженный в плане дорожной обстановки временной промежуток, когда поток машин сконцентрирован в центральной части города. Кроме того, люди больше заинтересованы сообщать о проблемных местах, чтобы выразить возмущение и беспокойство, чем сообщать о нормальной дорожной обстановке.

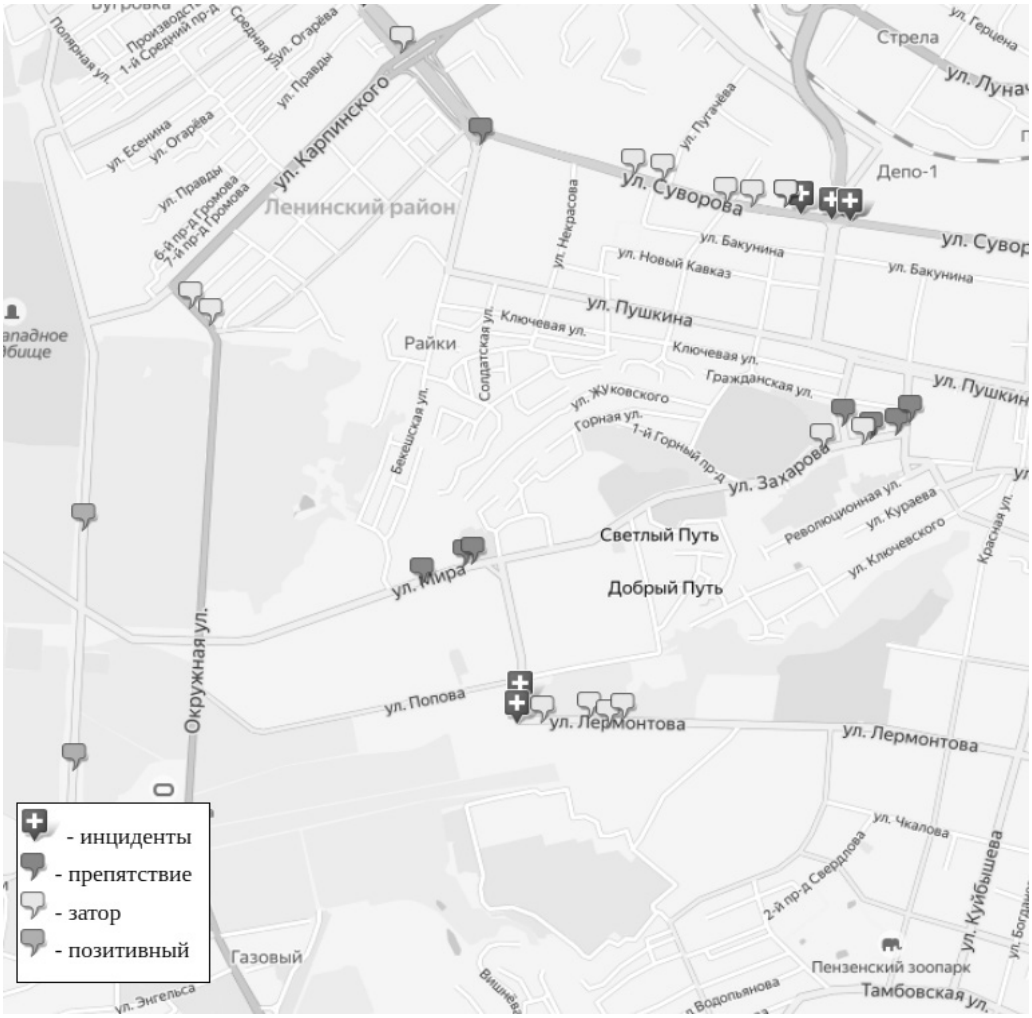

Рис. 2. Карта дорожных событий

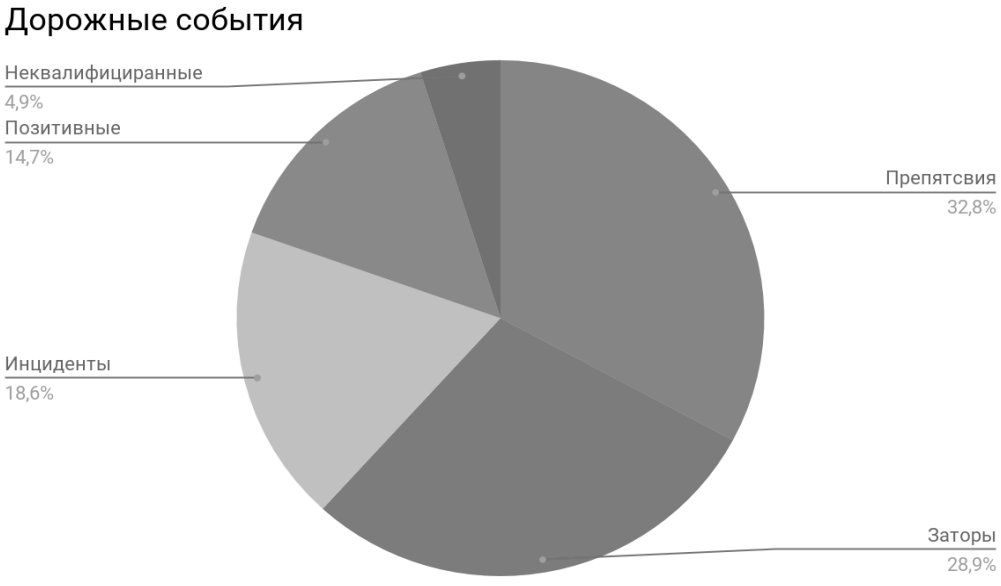

Рис. 3. Диаграмма распределения сообщений по классам

Результаты исследования метода

Для проверки работы метода был выбран набор из 500 сообщений, связанных с дорожной обстановкой в исследуемой области, которые были структурированы и преобразованы в числовые вектора. Каждый из них был первоначально классифицирован путем анализа, присутствующих в нем слов, с помощью процедуры интеллектуального анализа. Для проверки метода создан тестовый набор из 200 сообщений: 38 из них являются «инциденты», 64 — «препятствия», 30 — «позитивные», 58 — «заторы» и 10 — прочие. На рисунке 3 показана диаграмма распределения сообщений по классам.

Оценка эффективности показала, что показатель precision равен 0.92, показатель recall — 0.86, а F-метрика — 0.69.

В целях сравнения результатов классификации и оценочных метрик был выбран алгоритм Naive Bayes (NB), который работал на том же тестовом наборе. Алгоритм NB выбран в качестве сравнения, поскольку является одним из наиболее популярных для классификации коротких сообщений электронной почты и SMS сообщений, например, при идентификации спама. Главным преимуществом алгоритма NB перед KNN, заключается в том, что NB генерирует собственный набор слов и вычисляет распределение слов до классификации данных. Несмотря на преимущества в нашем случае алгоритм NB показал худшие результаты. После тестирования показатель precision равен 0.75, recall — 0.69, F-метрика — 0.7.

Заключение

В результате исследований были получены следующие выводы:

1. Метод обрабатывает тексты, полученные из социальных сетей, и позволяет классифицировать их с помощью алгоритма машинного обучения.
2. Метод был апробирован для анализа сообщений о событиях в дорожном движении. Однако его можно использовать для анализа и других сообщений, которые связаны с различными аспектами городской жизни, например, экология, реакция граждан на события и т.д.
3. Для анализа использовался набор сообщений из открытых групп их социальных сетей. Информация из этих источников является достаточно актуальной и помимо контента содержит координатную информацию.
4. Набор классифицированных сообщений отражает дорожные события и показывает реакцию жителей, что позволяет использовать его для принятия решений по улучшению транспортной обстановки.

Поскольку муниципальная и научная сферы обеспокоены улучшением жизни людей, важность исследования процессов в городской среде возрастает с каждым годом. Дорожное движение в городе является одним из факторов, влияющим на городскую среду, как например пробки. Данные ситуации могут быть изучены с целью выявления причин и особенностей, а также оповещения аварийных и медицинских служб. Социальные сети являются важным источником информации, поскольку сообщения в них написаны людьми, описывающими события в непосредственной близости от них, причем в контенте хранится реакция участников дорожного движения.

ЛИТЕРАТУРА

1. Свободная энциклопедия «Википедия» [Электронный ресурс].— Режим доступа https://en.wikipedia.org/wiki/Main_Page.— (Дата обращения: 14.01.2018)
2. Сайт социальной сети «Вконтакте» [Электронный ресурс].— Режим доступа https://vk.com.— (Дата обращения: 10.01.2018)
3. Профессиональный информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных « MachineLearning.ru» [Электронный ресурс].— Режим доступа http://www.machinelearning.ru.— (Дата обращения: 16.01.2018)
4. А. Б. Мерков «Введение в методы статистического обучения»: Москва: Едиториал УРСС, 2011.— 254 с
5. Блог компании «Open Data Science» [Электронный ресурс].— Режим доступа https://habrahabr.ru/company/ods/.— (Дата обращения: 17.01.2018)
6. Харламбос Марманис, Дмитрий Бабенко «Алгоритмы интелектуального Интернета. Передовые методики сбора, анализа и обработки данных», Символ-Плюс, 2011.— 480 с.

Журнал «Современная наука: актуальные проблемы теории и практики»