Секция «Искусственный интеллект в контрольно-надзорной деятельности»

Анализ тональности русскоязычных сообщений пользователей в социальных сетях

Научный руководитель – Масленникова Юлия Сергеевна

Гимранова Камиля Ринатовна

Студент (бакалавр)

Казанский (Приволжский) федеральный университет, Институт физики, Казань, Россия

E-mail: kamilyagimranova09@gmail.com

С быстрым развитием Интернета возрастает его значение в жизни пользователей, все больше людей выражают своё мнение по тому или иному вопросу в социальных сетях, блогах, на форумах, специализированных площадках. Огромный объем информации, появляющийся в социальных сетях, регулярно обрабатывается и автоматически классифицируется для решения широкого круга задач. Одной из таких задач является классификация сообщений пользователей по тональности, что позволяет, например, оценить реакцию на какое - то событие или новость, спрогнозировать результат выборов, проанализировать отзыв о товаре или услуге, а в дальнейшем улучшить тот или иной параметр. Здесь и в дальнейшем в работе, термин “тон” обозначает эмоциональный окрас сообщения (положительное отношение к чему-либо, отрицательное отношение или нейтральное). Стоит отметить, что для английского языка разработано множество сервисов для автоматического анализа тональности текстов, в открытом доступе представлены уже размеченные базы данных, которые могут быть использованы для разработки собственных моделей машинного обучения. Однако для русского языка задача осложняется отсутствием достаточного объема размеченных данных, необходимостью трудоемкого морфологического анализа и другими специфическими особенностями русского языка.

В данной работе проведен обзор различных подходов к классификации документов по уровню тональности, в который вошли методы машинного обучения (с учителем и без) с предварительным извлечением признаков, методы основанные на семантике слова (на правилах или на словарях). Наиболее широкое распространение получили подходы, основанные на так называемых тональных словарях, где каждому слову сопоставлена тональность. Однако зачастую значение имеет не только тональность отдельного слова, а комбинация и порядок слов. Для классификации тональности текстов в работы были применены следующие методы: наивный байесовский классификатор, классификатор максимальной энтропии, метод опорных векторов, pLSA (вероятностный латентно-семантический анализ), словарный метод. Рассматривалась возможность осуществления перечисленных методов на русскому языке (с учетом его особенностей). Для анализа использовался корпус текстов русскоязычного сегмента платформы микроблогинга Twitter. Для словарного метода применялась русифицированная версия словаря WordNet-Affect, словарь RuSentiLex и словарь RuSentiment. Также в ходе обучения учитывались эмотиконы и эмодзи, все чаще используемые людьми в твитах и сообщениях. Исходя из рассмотренных методов производилась оценка качества классификации, построенная на точности и полноте, а также подтверждение выявленных исследователями плюсов и минусов того или иного метода.