**РЕЦЕНЗИЯ**

**на выпускную квалификационную работу обучающегося СПбГУ**

**Миронюк Елизаветы Евгеньевны**

**по теме “Автоматический анализ тональности на материале сообщений о политических партиях в социальных сетях”**

Диссертация Е. Миронюк посвящена исследованию проблем автоматического анализа тональности, в работе разрабатывается и реализуется система автоматического анализа тональности сообщений о политических партиях в социальных сетях.

Тема исследования, безусловно, актуальна: с распространением социальных сервисов и стремительным ростом объемов текстовой (и не только) информации, порождаемой сотнями миллионов пользователей Сети со всего мира, автоматический анализ тональности сегодня является важным инструментом мониторинга настроений аудитории и в высшей степени востребован в социологии, маркетинге, рекламном бизнесе, PR-технологиях и многих других сферах деятельности, поэтому всестороннее изучение проблем, связанных с анализом мнений и реализацией программных продуктов, позволяющих осуществлять такой анализ, является одной из важнейших задач современной лингвистики.

Структура работы, обоснованная задачами исследования, является логичной и убедительной. Диссертация Е. Миронюк состоит из введения, трёх глав, заключения и списка используемой литературы. В первой главе чётко и непротиворечиво определяются основные термины и понятия, которые будут использованы в работе, исследуются задачи, проблемы и методы автоматического анализа тональности. Широкое использование современной литературы, включающей в себя научные работы отечественных и зарубежных специалистов в области компьютерной лингвистики и электронной коммуникации, содержащие концепции, положенные в теоретическую базу исследования, подчёркивает актуальность и научный вес работы.

Во второй главе описывается создание корпуса текстов и базы данных, содержащих материал исследования, расписаны этапы предобработки материала, связанные с его особенностями, также в этой главе содержится описание алгоритма, разработанного на основе выделенных особенностей материала и оригинального гибридного метода автоматического анализа тональности, основанного на комбинации лингвистических методов и методов машинного обучения.

В третьей главе рассматривается создание системы автоматического анализа тональности сообщений о политических партиях в социальных сетях, производится оценка её эффективности, анализируются результаты работы и определяется дальнейшее направление исследований. В процессе реализации системы автор демонстрирует владение последними достижениями науки и практики, а именно такими программными продуктами, как: Twitter API, Microsoft Visual Studio 2017, MyStem, а также веб-технологиями.

Оценка эффективности созданной в рамках диссертационного исследования системы Sentimentor показала, что гибридный метод, разработанный в исследовании, повышает эффективность системы по сравнению с классификацией на основе Наивного байесовского классификатора. При этом средняя точность для обоих классов увеличивается на 12,91%, полнота — на 16,11%, а значение F1-меры вырастает на 18,75%. Достоверность полученных результатов не вызывает сомнений.

Особенно хочется отметить интересное и, как нам кажется, очень удачное решение автора работы – разработку веб-интерфейса для дальнейшего взаимодействия эксперта с системой. Активное привлечение экспертов на созданный веб-ресурс действительно может повысить эффективность работы классификатора, более того, этот проект при должном управлении может разрастись и стать популярным среди специалистов инструментом анализа твитов политического содержания или площадкой для сотрудничества таких специалистов.

Таким образом, содержание выпускной квалификационной работы полностью соответствует заявленной в названии теме. В работе отражены актуальные проблемы автоматического анализа тональности не только теоретического, но и практического характера. Работа имеет высокую практическую значимость: её результаты, безусловно, могут быть востребованы в дальнейших исследованиях проблем автоматического анализа тональности, а также при проведении социологических и политических исследований на материале сообщений Твиттера.

Диссертацию отличает высокий научный уровень, а автора — методичность, хорошая подготовка в области компьютерной лингвистики, умение работать с лингвистическим материалом, анализировать его и обобщать полученные наблюдения.

В процессе ознакомления с работой, однако, возник ряд замечаний и вопросов, не влияющих на общую высокую оценку исследования.

Отмечая доступность работы читателям с точки зрения языка, стиля и структуры, всё же необходимо указать, что в тексте присутствуют опечатки, самая существенная из которых замечена в формуле на стр. 17: в знаменателе вместо *P(Sentiment)* должно быть *P(Sentence)*. Очевидно, что это является лишь опечаткой, а не ошибкой в расчётах, поскольку далее в работе на стр. 41 автор приводит исходную верную формулу.

Не умаляя наглядности графического материала, размещённого в тексте, следует заметить, что включение глоссария предметной области и приложения с дополнительными иллюстрациями проделанной работы, сделало бы диссертацию более завершённой с точки зрения оформления.

Также необходимо высказать замечание по оформлению цитирования. Некоторые утверждения и определения, а также графические материалы не сопровождены ссылками на источник. Например, такие утверждения как:

- “Для решения подобных задач существует набор автоматизированных методов, объединённых названием «Анализ тональности текста» (англ. «Sentiment Analysis»)” (стр. 3);

- “Самым простым и типичным решением представляется применение в таком случае аддитивного сглаживания (сглаживания Лапласа, add-one smoothing)” (стр. 42);

- “Социальная сеть— онлайн-сервис, созданный на платформе Web 2.0, структура которого представляет собой граф: в узлах (вершинах) графа находятся пользователи, каждый из которых имеет профиль с определённой информацией, а рёбрами являются связи между пользователями” (стр. 4);

- “Традиционно для оценки алгоритмов сентимент-анализа используются такие метрики, как precision(точность)**,** recall(полнота) и F1-мера” (стр. 51).

В таких случаях рекомендуется оформлять косвенное цитирование, поскольку очевидно, что данные утверждения основаны на прочитанной автором литературе. Кроме того, таблица 3 на стр. 52 не сопровождена ссылкой. При заимствовании графических материалов, в том числе таблиц, также необходимо указывать ссылку на источник информации.

В процессе ознакомления с работой возникли следующие вопросы к автору:

1. В разделе 2.4. “Предварительная обработка текстов на основе их особенностей” в пункте 6 на стр. 37 указано: “Количество повторяющихся гласных *(«ну неееееет»*) сокращается до двух («ну неет»). Предполагается, что интенсивность тональности слов с пролонгированными гласными выше, чем тех же слов в привычном написании, поэтому не производится преобразования к словарной форме”. Затем в пункте 7 указано, что такому же сокращению подвергаются восклицательные знаки. Почему на этом этапе предобработки выбран именно такой способ сокращения, если интенсивность тональности не учитывается при анализе? Как программа Mystem справилась с обработкой таких слов?
2. На стр. 65 в заключении утверждается: “В исследуемом материале длина твита редко превышает одно предложение, а потому задача определения тональности твита может быть приравнена к задаче определения тональности предложения. Из этого следует, что предложенная методика может быть применена и к текстам из других социальных сетей.” Формулировка “редко превышает” неубедительна, в данном случае требуется подкрепление утверждения объективными количественными данными. Также, в связи с этим возникает вопрос: каковы в целом различия в анализе тональности длинных и коротких текстов? Каким образом на основе анализа тональности отдельных предложений будет делаться вывод о тональности всего текста сообщения в любом другом социальном сервисе?
3. В список дальнейших направлений исследования включено решение задачи распознавания сарказма и иронии в тексте. Предлагаемое автором решение, состоящее в создании “очень большой обучающей выборки, содержащей ручную разметку” выглядит безосновательно оптимистичным. Как, в таком случае, планируется избежать проблемы переобучения? Какие методы автоматического распознавания сарказма, описанные в современной литературе, наиболее эффективны по мнению автора?
4. В рамках работы создан веб-интерфейс для дальнейшего взаимодействия эксперта с системой, что представляется весьма перспективным шагом для совершенствования классификатора. Как планируется привлекать экспертов к работе с системой?

Следует подчеркнуть, что высказанные замечания и вопросы являются дискуссионными и не влияют на итоговую положительную оценку диссертации, которая вносит существенный вклад в развитие современной научной мысли в целом и в разработку проблем автоматического анализа тональности.

Заявленная в названии тема полностью раскрыта, дано развернутое обоснование выводов.

На основании вышесказанного очевидно, что диссертационное исследование Е. Миронюк “Автоматический анализ тональности на материале сообщений о политических партиях в социальных сетях” заслуживает высокой оценки.

«25» мая 2018 г. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ к.ф.н. Карнуп Екатерина Владимировна