НАУЧНЫЙ ОТЧЕТ

(отчет о достижении установленных при предоставлении гранта целевых показателей)

за 2019 год

по гранту Президента Российской Федерации

для государственной поддержки

молодых российских учёных

за счёт средств федерального бюджета

МД-6259.2018.6

1. Номер гранта:

МД-6259.2018.6

2. Фамилия, имя, отчество:

Бодрунова Светлана Сергеевна

3. Тема научного исследования:

Кросс-культурный тональный анализ пользовательских текстов в сети Интернет

4. Полученные за отчетный период научные (научно-технические) результаты:

Рабочей группе НИР удалось добиться следующих научных и научно-практических результатов.

1. Выполнена основная задача исследования, а именно – создан программный комплекс для мультиязычного (кросс-культурного) анализа тональности пользовательских сообщений в сети Интернет. Этот комплекс настроен на работу с короткими зашумленными текстами, но может выполнять (с более высокими показателями эффективности) также тональный анализ текстов большей длины. Метрики качества выше 0,72 были достигнуты для трех языков: английского, немецкого, русского, для французского показатели превышают 0,6, продолжается работа по достижению показателей на «шумных» текстах на других языках (испанском и итальянском). Показатели качества на лучше структурированных выборках (например, пользовательских отзывах или рекомендациях) достигают 0,9.

2. Программный комплекс осуществляет тональный анализ не только по трем привычным категориям, но также по отдельным эмоциям (с показателями качества выше 0,7 – 0,8), иронии/сарказму, проявлениям коммуникативной агрессии. Пользователям также доступен объект-ориентированный тональный анализ (по отношению к заданному объекту).

3. Разработанный комплекс осуществляет несколько дополнительных процедур, в том числе графовую визуализацию данных, отбор и ранжирование пользователей по тональности и др.

4. Получено патентное свидетельство на программный комплекс по сбору данных (Свидетельство о государственной регистрации программы для ЭВМ №2019661692 от 05.09.2019 – комплекс «Crawlector»), в процессе получения находится патентное свидетельство на программный комплекс «MultiSentimentLab» (заявка №2019664459 от 14.11.2019).

5. При этом показано, что достигнут «потолок метода» при комбинации предложенного словарного подхода и

метода опорных векторов. Это проявляется в том, что дополнительное ручное кодирование тестовых выборок не дает существенного прироста показателей качества выявления сентимента текстов, а также может приводить к переобучению машины. С помощью отказа от элементов препроцессинга данных, изменения гиперпараметров и применения инструментов регуляризации можно лишь незначительно улучшить качество метода.

6. Предложены альтернативы используемой комбинации методов, а именно – метод наложения лексической маски (прошел стадию моделирования результата) и метод, комбинирующий подходы в сфере глубокого обучения (deep learning). На момент сдачи отчета в СПбГУ проходит стадию развертывания нейронная сеть, которая будет использоваться для глубокого обучения и улучшенного выявления тональности высказываний пользователей.

7. Выявление отдельных эмоций пользователей показало, что возможно выявить роль отдельных эмоций в нарастании/угасании дискуссии. Показано, что, в отличие от ожидаемого, положительные эмоции (сочувствие) обладают более широким кросс-культурным потенциалом для наращивания сетевой дискуссии, тогда как злость/ярость такого результата не демонстрируют. Исследования будут продолжены в области иных эмоций, а также в области кросс-культурного «перетекания» (emotional spillover) между разными языками в рамках глобальных дискуссий.

8. Параллельно основной ветви исследования рабочая группа выявляла связь между тематикой дискуссии и тональностью тем. Для описания тематики дискуссии был задействован метод тематического моделирования (topic modelling) в вариациях классических алгоритмов (LDA) и алгоритмов, приспособленных для работы с короткими текстами (WNTM и BTM). По автоматическим метрикам качества (topic coherence – Umass и NPMI) был выбран алгоритм BTM для дальнейших экспериментов. Однако эксперименты с ручным кодированием для всех трех алгоритмов показали, что: 1) автоматические метрики качества моделирования не соответствуют взгляду кодировщиков; 2) мнение кодировщиков также не является достоверным показателем качества моделирования; 3) короткие тексты высокой зашумленности в принципе не дают высокого числа интерпретируемых тем.

9. В связи с этим было предложено пересмотреть в целом саму задачу тематического моделирования на коротких текстах, с тем чтобы выявлять не наибольшее количество интерпретируемых тем, а «идеальные топики». Предложены метрики качества отдельной темы (а не модели в целом), которые могут быть связаны между собой и, таким образом, опишут «идеальную тему»: интерпретируемость (может быть оценена вручную или автоматически), выпуклость темы на «тепловой карте» и устойчивость темы, то есть ее распределенность по достаточному числу ключевых слов.

10. Несмотря на обоснованные сомнения в качестве выделения тем, был проведен эксперимент по оценке связи интерпретируемости тем и негативизма в их описании (30 топ-слов). Была показана корреляционная связь между объемом негативизма в теме и ее интерпретируемостью. Сделан вывод о необходимости более подробного изучения процесса интерпретации темы, чтобы отделить «вчитывание» смыслов кодировщиками от действительной связи между негативизмом и качеством дискуссии в теме.

5. Ожидаемые направления дальнейшего использования полученных за отчетный период результатов:

Ожидаемые направления дальнейшего использования полученных результатов включают:

Продолжение научных исследований:

1. Разработка новых методов кросс-культурного тонального анализа на основе улучшенных методов сентимент-анализа (применение лексической маски и агломеративной кластеризации текстов), а также глубокого обучения на основе комбинированной архитектуры BERT/XLNet.

2. Разработка новой идеологии для тематического моделирования на коротких текстах, основанной на новых метриках качества на уровне отдельной темы (интерпретируемость, выпуклость, устойчивость темы).

3. Доработка инструментов анализа связи между тематикой и тональностью дискуссии.

4. Дальнейшая разработка методов выявления отдельных эмоций, иронии/сарказма, коммуникативных агрессий в текстах пользователей, дальнейшее выявление связи эмоций и структуры дискуссии.

Использование в индустрии информационного поиска:

1. Участие в государственных проектах по мониторингу настроений пользователей по вопросам здравоохранения и конфликтов.

2. Участие в коммерческих и некоммерческих проектах по оценке деловой репутации, общественного мнения, конфликтогенности социальных групп.

Использование в преподавании:

1. Использование результатов для обучения на социогуманитарных программах обучения (профиль – социология интернет-исследований, коммуникативистика, лингвистика и компьютерная лингвистика).

2. Использование методов для обучения на программах обучения в области компьютерных наук.

3. Использование разработанного ПО для научно-исследовательской практики магистранстов обоих направлений.

6. Выполнение грантополучателем заданных индикаторов в отчетном году:

№	Наименование индикатора	Ед. изм.	2019 г. план	2019 г. факт
1	Количество основных научных публикаций грантополучателя (монографии, учебники, учебные пособия, статьи, тезисы докладов, другие публикации)	ед.	5	5
1.1	количество публикаций, индексируемых в международной информационно - аналитической системе научного цитирования Web of Science	ед.	1	1
1.2	количество публикаций, индексируемых в международной информационно - аналитической системе научного цитирования Scopus	ед.	2	4
1.3	количество публикаций, индексируемых в международной информационно - аналитической системе научного цитирования European Reference Index for the Humanities	ед.	1	0
1.4	количество публикаций в российских отраслевых научных изданиях, входящих в перечень ведущих рецензируемых научных журналов и изданий РИНЦ	ед.	1	0
2	Участие грантополучателя в конференциях, в том числе международных (кол - во докладов)	ед.	2	4

3	Количество курсов лекций, подготовленных и читаемых грантополучателем	ед.	1	2
4	Численность защитивших кандидатские диссертационные работы под руководством грантополучателя	ед.	1	1
5	Количество привлекаемых к НИР соисполнителей	ед.	4	4
6	Количество результатов интеллектуальной деятельности в рамках проекта	ед.	0	1

6.1. Комментарий к выполнению заданных индикаторов в отчетном периоде:

1. Публикационные индикаторы по проекту выполнены с превышением некоторых заявленных показателей. Так, планировалось создать в целом 5 публикаций; к моменту написания отчета опубликованы 3 и приняты к публикации 3 статьи. Из них - пять статей в международных трудах конференций и одна глава в международной монографии 'Digital Russia Studies' издательства Palgrave-Macmillan. Поскольку статьи публиковались во второй половине года, пока проиндексированы только две из них, но ожидается публикация и индексация всех статей в течение 2020 года. Ожидается индексация в Scopus пяти из выполненных статей, в Web of Science - двух из выполненных статей, но не исключено, что индексацию в Web of Science пройдет большее число статей.

2. Показатель по индексации в ERIH равен 0, поскольку он уже выполнен в 2018 году. Статья, индексируемая в ERIH и имеющая аффилиацию с Грантом Президента РФ («Кросс-культурный тональный анализ пользовательских текстов в Твиттере», Вестн. Моск. ун-та, серия 10 «Журналистика», 2018, №6), получила Первую премию Национальной ассоциации массмедиа-исследователей (НАММИ) в номинации «Статья» за 2018 год. По той же причине не выполнен показатель по российским научным изданиям: он уже выполнен в 2018 году. Вместо публикации в российском издании выполнена публикация в международных трудах конференции, также индексируемая в РИНЦ и доступная российским исследователям.

3. Показатель по участию в конференциях также превышен. Так, заявлено было два выступления; состоялось три, а также состоится одно выступление в 2020 году, поскольку статья, аффилированная с НИР, принята к публикации на конференции HCI International 2020 в Копенгагене, Дания (июль 2020 года). Таким образом, число одобренных докладов составляет четыре.

4. Число курсов лекций, подготовленных грантополучателем, также превышено. Так, разработано содержание курса "Big social data: анализ больших массивов текстовых данных" для новой магистерской программы СПбГУ 18.5782.1 "Медиакоммуникации". Также создан курс "Цифровая демократия" для еще одной новой магистерской программы СПбГУ - "Международная журналистика", где один из модулей посвящен анализу больших массивов данных.

5. Показатель по защищенным диссертациям формально не выполнен, поскольку, как мы указывали в отчете 2018 года, целевой аспирант не смог продолжить работу над диссертацией. Однако в 2018-2019 годах грантополучатель осуществляла научное консультирование для двух диссертаций, близких тематике НИР.

Первая диссертация – кандидатская диссертация аспиранта СПбГУ А.Ю.Максимова «Вебометрические методы исследований топологий крупных веб-сегментов», непосредственно посвященная теме сбора данных,

использовавшихся в рамках данной НИР. Именно она отражена в индикаторах НИР. Научным руководителем диссертации является соисполнитель данной НИР И.С.Блеканов, постоянным научным консультантом - С.С.Бодрунова. Общее число публикаций, представляемых автором диссертации в ВАК, - 9, из них в соавторстве с грантополучателем – 3, в том числе это публикации по НИР за 2018 год. Также аспирантом получен совместный патент с И.С.Блекановым - свидетельство о государственной регистрации программы для ЭВМ №2019661691 по сбору данных из социальных сетей, а также совместный патент с грантополучатем и И.С.Блекановым – свидетельство о государственнной регистрации программы для ЭВМ №2019661692 по вебометрическому анализу. ПО, запатентованное в рамках данных РИД, применялось для сбора данных и графового анализа в том числе в рамках данной НИР (см. приложенный файл). Защита запланирована на сентябрь 2020 года.

Вторая диссертация – докторская диссертация доцента СПбГУ К.Р.Нигматуллиной «Профессиональная журналистская культура в современной России», где анализ поведения журналистов онлайн осуществляется в том числе методами, разработанными в рамках данной НИР. Общее число совместных публикаций грантополучателя с автором диссертации, входящих в диссертационный список автора, - 4, в том числе одна - отчетная за 2019 год по данной НИР. Защита запланирована на лето 2020 года.

Обе защиты отложены на вторую половину 2020 года в силу перехода СПбГУ на систему выдачи собственных степеней кандидата и доктора наук, а также в силу смены системы формирования диссертационных советов (переход к ad hoc - советам, собираемым отдельно для каждой защиты).

6. Число соисполнителей НИР осталось прежним - 4 человека, что укладывается в требуемые показатели.

7. По проекту подана заявка на РИД - свидетельство о государственной регистрации программы для ЭВМ №2019664459 от 14.11.2019 «Программа для мультиязычного анализа тональности пользовательских сообщений в социальных сетях (MultiSentimentLab)». Этот результат не планировался, однако был получен. Поскольку НИР носила технико-методологический характер, именно получение свидетельства о регистрации программного комплекса отразило завершенное состояние проекта и его основное достижение - создание инструмента для мультилингвального тонального анализа, включающего русский и другие языки. Более подробно программный комплекс описан в прилагаемом файле. Также были получены два других патентных свидетельства – на ПО по вебометрическому анализу, №2019661691 от 05 сентября 2019 года – и по сбору данных из социальных сетей, №2019661692 от 05 сентября 2019 года (см. приложенный файл).

7. Публикации грантополучателя за отчетный период по заявленной тематике: 5

7.1. Количество публикаций по типам:

- Монографии: 0
- Учебники, учебные пособия: 0
- Статьи: 4
- Тезисы докладов: 0
- Другие публикации: 1

7.2. Количество публикаций, индексированных в WoS, Scopus, ERIH, РИНЦ,:

- количество публикаций, индексируемых в международной информационно-аналитической системе научного цитирования Web of Science: 1
- количество публикаций, индексируемых в международной информационно-аналитической системе научного цитирования Scopus: 4
- количество публикаций, индексируемых в международной информационно-аналитической системе научного цитирования European Reference Index for the Humanities: 0
- количество публикаций в российских отраслевых научных изданиях, входящих в перечень ведущих рецензируемых научных журналов и изданий РИНЦ: 0

7.3. Перечень публикаций в Web of Science:

№ п/п	Название публикации	Авторы	Название издания	Тип публикации	ISSN издания/ISBN издательства	Год публикации	Идентификатор статьи в Web of Science
1	Topic modeling for Twitter discussions: Model selection and quality assessment	Bodrunova, S.S., Blekanov, I.S., Kukarkin, M.M.	Proceedings of the 6th SGEM International Multidisciplinary Scientific Conferences on SOCIAL SCIENCES and ARTS SGEM2018, Science and Humanities	Статья	2682 - 9959		нет (Вышла, ожидает индексации)*

* прилагается справка о принятии публикации в печать

7.4. Перечень публикаций в Scopus:

№ п/п	Название публикации	Авторы	Название издания	Тип публикации	ISSN издания/ISBN издательства	Год публикации	Идентификатор статьи в Scopus
1	Topics in the Russian Twitter and Relations between their Interpretability and Sentiment	Bodrunova S.S.; Blekanov I.S.; Kukarkin M.	2019 6th International Conference on Social Networks Analysis, Management and Security, SNAMS 2019	Conference Proceeding Conference Paper		2019	2 - s2.0 - 85077818908
2	Network Presentation of Texts and Clustering of Messages	Orekhov A.V.; Kharlamov A.A.; Bodrunova S.S.	Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)	Book Series Conference Paper	03029743	2019	2 - s2.0 - 85076538996
3	Topic modelling in Russia: current approaches and issues in methodology	Bodrunova S.S.	Handbook of Digital Russia Studies (монография)				нет (Отдано в печать. Заключено соглашение с автором. Публикация - в 2020 году.)*
4	When emotions grow: Cross - cultural differences in the role of emotions in the dynamics of conflictual discussions on social media	5. Bodrunova, S.S., Nigmatullina, K.R., Blekanov, I.S., Smoliarova, A.S., Zhuravleva, N.N., Danilova, Y.S.	Proceedings of the 22nd International Conference on Human - Computer Interaction (HCI International 2020)	Статья			нет (Отдано в печать Будет издана в 2020 году, принята к публикации.)*

* прилагается справка о принятии публикации в печать

7.5. Перечень других значимых публикаций, не входящих в Web of Science и Scopus:

Нет

8. Результаты интеллектуальной деятельности за отчетный период по заявленной тематике:

Общее количество: 1

№ п/п	Наименование объекта интеллектуальной собственности	Вид объекта	Охранный документ (патент, свидетельство о регистрации)		Подана заявка (заполняется в случае, если охранный документ еще не выдан)	
			№	Дата выдачи	№	Дата выдачи
1	Заявка на свидетельство о государственной регистрации программы для ЭВМ «Программа для мультиязычного анализа тональности пользовательских сообщений в социальных сетях» (MultiSentimentLab) »	Программа для ЭВМ			2019664459	14.11.2019

9. Участие грантополучателя в отчетном году в научных конференциях и семинарах по заявленной тематике (кол-во докладов):

- международные мероприятия: 4

№ п/п	Название мероприятия	Место и время проведения	Название доклада
1	6th SGEM International Multidisciplinary Scientific Conference on SOCIAL SCIENCES and ARTS – SGEM2018	Albena, Bulgaria, 24.08.2019 - 02.09.2019	Topic modeling for Twitter discussions: Model selection and quality assessment
2	6th International Conference on Social Networks Analysis, Management and Security (SNAMS)	Granada, Spain, 22.10.2019 - 25.10.2019	Topics in the Russian Twitter and relations between their interpretability and sentiment
3	6th International Conference on Internet Science (INSCI'2019)	Perpignan, France, 02.12.2019 - 05.12.2019	Network Presentation of TextNes and Clustering of Messages
4	22nd International Conference on Human - Computer Interaction – HCI International 2020	Kopenhagen, Denmark, 19.07.2019 - 24.07.2019	2020 ГОД! When emotions grow: Cross - cultural differences in the role of emotions in the dynamics of conflictual discussions on social media

- другие мероприятия: 0

10. Научно-педагогическая деятельность грантополучателя и соисполнителей за отчетный период по заявленной тематике:

- курсы лекций, подготовленные и читаемые грантополучателем: 2

№ п/п	Наименование учебного заведения	Название курса
1	Санкт - Петербургский государственный университет	Big social data: анализ данных из социальных сетей
2	Санкт - Петербургский государственный университет	Цифровая демократия

- количество дипломных работ, подготовленных под руководством грантополучателя: 4

- кандидатские диссертации, подготовленные под руководством грантополучателя: 1

№ п/п	Специальность ВАК	Количество
1	05.13.01	1

- количество публикаций соисполнителей, подготовленных совместно или под руководством грантополучателя по заявленной тематике: 0

- участие соисполнителей в выполнении исследований по гранту за отчетный период: 4

№ п/п	Ф.И.О. соисполнителя	Статус	Краткое описание выполненной работы
1	Блеканов Иван Станиславович	кандидат наук	Выполнение основных технических параметров НИР, разработка программного обеспечения, тестирование модулей краулера и показателей тонального анализа, показателей тематического моделирования, обзор литературы по основным методам тематического моделирования.
2	Кукаркин Михаил Михайлович	студент	Участие в разработке программного обеспечения для тематического моделирования и визуализации его результатов, архивация данных, тестирование методов графовой репрезентации данных
3	Смолярова Анна Сергеевна	кандидат наук	Проведение и контроль ручного кодирования немецкоязычных датасетов, участие в разработке параметров "идеальных тем", участие в кодировании и обзоре литературы по выявлению отдельных эмоций в текстах пользователей
4	Данилова Юлия Сократовна	кандидат наук	Проведение и контроль ручного кодирования франкофонных кейсов, участие в разработке параметров "идеальных тем"

11. Участие грантополучателя в других научных исследованиях (гранты, ведомственные программы, ассигнования и др.) за отчетный период по заявленной тематике

№ п/п	Название проекта	Размер финансирования (млн. руб)	Источник финансирования	Срок выполнения проекта	Основные результы проекта
1	"Кривое зеркало" конфликта: роль сетевых дискуссий в репрезентации и динамике этнополитических конфликтов в России и за рубежом	4800000.000	бюджетные источники, в том числе из государственных фондов поддержки научной, научно - технической и инновационной деятельности	2019	НИР посвящена анализу конфликтогенности онлайн - дискуссий в разных странах. Разработана методика и изучена политическая поляризация дискуссий, оценен конфликтогенный потенциал инфлюэнсеров, применены алгоритмы графового отражения результатов исследования. Оценена возможность применения методов тонального анализа и тематического моделирования для нахождения точек бифуркации дискуссий. Комплексно оценены (инфлюэнсеры, поляризация, агрессия) 3 кейса. Выдвинута гипотеза "обратной спирали молчания".

12. Общественное признание грантополучателя за отчетный период (премии, медали, дипломы и т.п.):

Общее количество: 1

№ п/п	Название премии/награды	Кем выдана	Год получения	Достижение, за которое вручена премия/награда
1	Первая премия НАММИ в номинации "Научная статья" за 2018 год	Национальная ассоциация массмедиа - исследователей (Россия)	2019	Статья С.С.Бодруновой "Кросс - культурный тональный анализ пользовательских текстов в Твиттере", Вестник Московского университета, серия 10 "Журналистика", 2018, №6.

Грантополучатель

_____/ Бодрунова С. С. /