УДК 004.82

ПРОДВИНУТЫЕ НЕЙРОСЕТЕВЫЕ МОДЕЛИ ДЛЯ РЕШЕНИЯ ЗАДАЧИ ОПРЕДЕЛЕНИЯ ТОНАЛЬНОСТИ

А. Г. Сбоев*, И. Е. Воронина, Д. В. Гудовских*, А. А. Селиванов***

**НИЦ «Курчатовский институт»*
***Воронежский Государственный Университет*

Поступила в редакцию 25.11.2016 г.

Аннотация. Представлены результаты ретроспективного исследования задач тональности SentiRuEval 2016 на базе сравнения различных сложных моделей. В работе оценивается влияние компонентов модели, таких как векторизация входных данных с помощью word2vec моделей, тип классификатора и выбор учебных корпусов. Лучшая модель демонстрирует результаты F1-микро 0,57 и F1-макро 0.61 для выборки сообщений по банкам, и F1-микро на 0.61 и F1-макро 0.74 для сообщений по телекоммуникационным компаниям.
Ключевые слова: нейронные сети, анализ тональности, рекуррентные сети, word2vec, LSTM, GRU.
Annotation. Results of retrospective investigation of tonality task of DIALOG 2016 on base of comparing different complicated models are presented. Influence of components of these models, such as word2wec vectorizing input data, the type of classifier, and the selection of training corpus, are evaluated. The best model demonstrates F1-scores with micro average of 0.57 and macro average of 0.61 for banks dataset, and F1-micro of 0.61 and F1-macro of 0.74 for telekom dataset.
Keywords: neural networks, sentiment analysis, recurrent networks, word2vec, LSTM, GRU.

ВВЕДЕНИЕ

В настоящее время применение нейронных сетей для задач анализа тональности текста получило достаточно широкое распространение. В частности, это иллюстрируется работами, представленными на крупнейших соревнованиях по автоматизированному анализу текста для английского языка SemiEval (Preslav Nakov, 2016 [1]), а также русского языка – SentiRuEval (Тарасов Д., 2015 [2]), (Трофимович Ю., 2016 [3]), (Карпов И.А., 2016 [4]) .

Одной из подзадач данных соревнований является классификация тональности сообщений из Twitter по следующему принципу: положительные, отрицательные и нейтральные.

Для русского языка соревнования содержат две тематические выборки сообщений: о банках и телекоммуникационных компаниях. Результаты соревнований SentiRuEval 2016 по классификации тональности сообщений Твиттера показали, что большинство участников использовали хорошо известные подходы в выборе классификатора модель SVM (Лукашевич Н.В., 2016 [5]) или подход на основе правил (Васильев В.Г., 2015 [6]), представление в виде набора словоформ, леммы или N-граммы и TF-IDF меры.

Среди лидеров соревнования были также и те, кто использовал более современные модели (Трофимович Ю., 2016 [3]), но в ограниченном количестве: была использована нейронная сеть глубокого обучения, состоящая из GRU слоев. Входные сообщения преобразовывались в вектор с помощью модели word2vec, обученной на данных из социальных сетей, Живого журнала и новостных сайтов. В целом представляется, что потенциал применения сложных нейросетевых моделей является недооцененным. В этой работе мы сравниваем ряд современных моделей на базе данных представленных на соревновании SentiRuEval 2016, в том числе:

1) Stacked LSTM или GRU – новый вид рекурентных нейросетевых моделей с памятью;

представляется в виде последовательного каскада LSTM или GRU слоев. Это позволяет запоминать последовательности большей длины по сравнению с обычными рекуррентными нейронными сетями.

a. LSTM – тип нейронов с ячейкой памяти.

b. GRU – упрощенный вид LSTM, в некоторых задачах дает прирост в точности по сравнению с LSTM, за счет меньшего числа параметров модели.

2) Word2vec – хорошо зарекомендовавшая себя модель векторного представления слов. Модели обучены на различных корпусах, таких как НКРЯ, веб-страниц, коротких сообщений в Twitter.

1. ОПИСАНИЕ НЕЙРОСЕТЕВЫХ МОДЕЛЕЙ

1.1. Описание данных

Работа основана на материалах соревнований по оценки тональности сообщений Твиттера SentiRuEval 2016. Сообщения разбиваются на три класса: положительные, нейтральные и отрицательные. Итоговая оценка результатов проводится с помощью F1-микро-и F1-макро оценок только для двух классов: положительные и отрицательные. Тестирование проводится на двух тематических наборах сообщений: банки и телекоммуникационные компании. Размер коллекции предоставляемой участникам соревнований для обучения равен 19,673 для Телеком и 19,586 для Банк (Лукашевич Н.В. [5]).

1.2. Описание используемых наборов признаков

Мы использовали модели word2vec для кодирования текста в последовательность векторов, которые являются входными данными для нейросетевых моделей. Для этой цели применялись уже существующие модели (RusVectōrēs) и обученная нами модель (Word2vec в Twitter).

• W2v ruscorpora – модель в открытом доступе, обучалась на полном национального корпуса русского языка (НКРЯ [8]). Корпус содержит 107 561 399 слов. Модель была обучена с использованием «Continuous Bag-of-Words» алгоритма. Размерность вектора 300, размер окна 2. Леммы, встречающиеся менее 3 раз, игнорировались.

• W2v web – модель в открытом доступе, обучалась на коллекции случайно собранных русскоязычных Интернет страниц, собранных в декабре 2014 года, объемом около 9 миллионов документов. Размер корпуса слов 660 628 738. Обучена с использованием «Scip-gram» алгоритма. Размерность вектора 500, размер окна 2. Леммы, встречающиеся менее 30 раз, игнорировались.

• W2v wikiruscorpora – модель в открытом доступе, обучалась на объединенных корпусах: национальный корпус русского языка и русская Википедия. Корпус содержит 280 187 401 слов. Модель обучена с использованием алгоритма «Continuous Bag-of-Words». Размерность вектора 500, размер окна 2. Леммы, встречающиеся менее 5 раз, игнорировались.

• W2v twitter – Модель тренировались на коллекции сообщений Twitter. Корпус содержит 12 миллионов сообщений (Рубцова Ю. В., 2015 [9]) и дополнен сообщениями из обучающего множества, представленного на соревнованиях SentiRuEval. Обучена с использованием «Scip-gram» алгоритма. Размерность вектора 200, размер окна 5. Леммы встречающиеся, менее 5 раз, были проигнорированы. Тексты были очищены от знаков препинания, имен пользователей, ссылок и нормализовались с помощью программы Mystem 3.0 (Сегалович И., 2003 [10]).

• Parent – в тексте выделялись биграммы на основе их синтаксической связи. Вектора слов в биграмме складывались последовательно в один общий вектор. Синтаксические связи устанавливались с помощью морфо-синтаксического разборщика, представленного в работе Рыбки Р. Б. [11].

• SumRank – предварительном этапе каждый тематический набор (банки и Телеком) обрабатывался с помощью энтропйно-вероятностного алгоритма, который представлен в работе Молошникова И. А. [12]. Этот алгоритм позволяет ранжировать слова по важности их представления в тематическом корпусе. Таким образом, получается дополнительный вектор признаков для сообщений из Twitter.

1.3. Топология нейронной сети

В статье рассматриваются несколько нейросетевых моделей:

1) Простая модель (none) – состоящая из одного скрытого слоя с тремя нейронами и активационной функцией softmax, которая предсказывает вероятность тональных классов.

2) Stacked Long Shot-Term Memory (LSTM) – в нее входят 2 LSTM слоя и выходной слой с тремя нейронами и активационной функцией «softmax». Каждый слой LSTM состоит из 200 нейронов и тангенсальной активационной функцией. Топология модели Stacked LSTM представлена на рис.

3) Stacked Gated Recurrent Unit (GRU) – архитектура данной нейронной сети похожа на предыдущую модель, но вместо нейронов LSTM используются нейроны GRU.

Все модели обучались с использованием функции оптимизации Adam. Использовался критерий раннего останова на основе валидационной ошибки для предотвращения переобучения модели. Тренировочная выборка была разделена на две части: 70 % на обучение и 30 % для проверки. Валидационная ошибка начинала расти, как правило, после 3–5 эпох. Модель с наилучшими параметрами (#22) мы обучали на всем тренировочном множестве за 3 эпохи, данная модель показала лучшие результаты.

2. ЭКСПЕРИМЕНТЫ

Для получения достоверных результатов оценки работы нейронной сети, обучение проходило с кросс-валидацией в 5 этапов. Каждая выборка данных делилась на 70 % для обучения и 30 % для тестирования, за исключением модели № 22, где для обучения использовалось все множество примеров. Тестирование проводилось на непересекающемся с тренировочным корпусе «золотой стандарт», на котором проходила оценка результатов соревнования SentiRuEval. Точность моделей оценивалось по показателям f1-макро и f1-микро (Лукашевич Н.В., 2016) для двух классов (положительные, отрицательные).

В табл. представлены результаты тестирования модели на различных наборах данных.

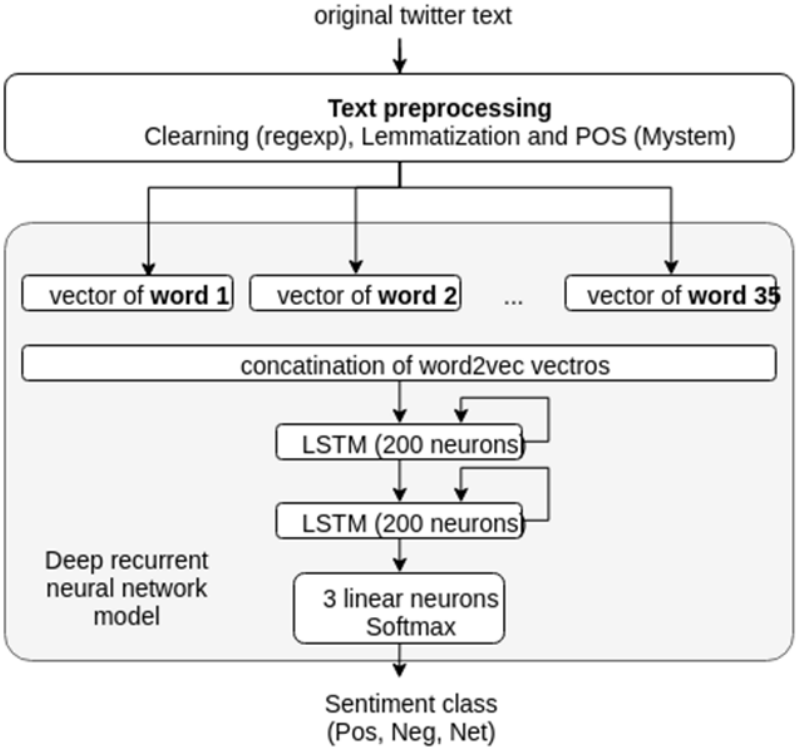

Рис. Топология модели Stacked LSTM

Результаты экспериментов

N	Hiddenlayer	Word2vecmodels				Syntacticparent	SumRank	Banks		Telekoms	
		Ruscorpora	WikiRuscorpora	Web	Twitter			F1Macro	F1Micro	F1Macro	F1Micro
1	None	+						0.33	0.41	0.45	0.59
2	None		+					0.37	0.45	0.45	0.6
3	None			+				0.38	0.46	0.5	0.64
4	None				+			0.4	0.49	0.48	0.66
5	None				+	+		0.44	0.5	0.5	0.66
6	None	+		+		+		0.44	0.5	0.51	0.65
7	None			+		+		0.44	0.5	0.51	0.65
8	None	+		+				0.4	0.47	0.51	0.64
9	LSTM	+						0.5	0.52	0.53	0.66
10	LSTM	+				+		0.5	0.52	0.54	0.66
11	LSTM		+					0.5	0.52	0.52	0.62
12	LSTM		+			+		0.5	0.52	0.51	0.63
13	LSTM			+				0.56	0.57	0.56	0.68
14	LSTM			+		+		0.5	0.53	0.54	0.67
15	LSTM				+			0.55	0.59	0.58	0.71
16	LSTM				+	+		0.52	0.57	0.54	0.67
17	LSTM	+		+	+	+		0.54	0.57	0.56	0.68
18	LSTM	+		+		+		0.52	0.55	0.55	0.68
19	LSTM	+		+				0.53	0.56	0.57	0.69
20	LSTM	+		+	+			0.55	0.58	0.58	0.7
21	LSTM	+			+		+	0.57	0.61	0.6	0.73
22	**LSTM**	**+**			**+**		**+**	**0.57**	**0.61**	**0.61**	**0.74**
23	Dialog2016bestresults							*0.55*	*0.59*	*0.56*	*0.68*
24	LSTM				+		+	0.56	0.6	0.59	0.71
25	LSTM				+	+	+	0.53	0.57	0.55	0.68
26	GRU	+				+		0.48	0.5	0.51	0.63
27	GRU		+			+		0.51	0.54	0.52	0.65
28	GRU	+						0.51	0.53	0.53	0.65
29	GRU			+		+		0.5	0.54	0.52	0.66
30	GRU		+					0.51	0.54	0.54	0.67
31	GRU	+		+		+		0.51	0.54	0.54	0.67
32	GRU	+		+				0.52	0.55	0.55	0.68
33	GRU				+	+		0.52	0.56	0.54	0.68
34	GRU				+			0.54	0.56	0.56	0.68
35	GRU	+		+	+	+		0.53	0.56	0.56	0.69
36	GRU			+				0.53	0.57	0.55	0.69
37	GRU	+		+	+			0.55	0.57	0.59	0.7

*модель была обучена за 3 эпохи на всем обучающем множестве, **лучший результат представленный на конференции SentiRuEval

ЗАКЛЮЧЕНИЕ

Достигаемая точность классификации существенно зависит от того, на сколько корпус, на котором обучается word2vec модель, близок по составу теме и стилю текста к анализируемому корпусу, в данном случае – твиттов. Модель на основе w2v_twitter показала большую точность по сравнению с остальными, такими как web, ruskorpora. Использование синтаксических связей дает прирост в точности около 1 %, что показали простые модели № 5,6,7. Дополнительные признаки в виде SumRank дают прирост примерно 1 % точности. Сочетание векторов из различных моделей w2v дает приросты в точности от 2 % до 5 %.

Данная работа выполнялась при поддержке гранта РФФИ №16-37-50078.

СПИСОК ЛИТЕРАТУРЫ

1. *Nakov P.* SemEval-2016 task 4: Sentiment analysis in Twitter/ Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, Veselin Stoyanov // Proceedings of the 10th international workshop on semantic evaluation (SemEval 2016), San Diego, US (forthcoming). – 2016.

2. *Тарасов Д. С.* Глубокие рекуррентные нейронные сети для аспектно-ориентированного анализа тональности отзывов пользователей на различных языках / Д. С. Тарасов // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Москва, 27–30 мая 2015 г.). Вып. 14 (21). – Москва: Изд-во РГГУ, 2015.

3. *Трофимович Ю.* Сравнение архитектур нейронных сетей в задаче анализа тональности русскоязычных твитов / Ю. Трофимович, К. Архипенко, И. Козлов, К. Скорняков, А. Гомзин, Д. Турдаков // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Москва, 1–4 июня 2016 г.). Вып. 15 (22) – Москва: Изд-во РГГУ, 2016.

4. *Карпов И. А.* Объектно-ориентированный анализ тональности при помощи синтаксических шаблонов и сверточной нейронной сети / И.А. Карпов, М. В. Кожевников, В. И. Казорин, Н. Р. Немов // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». Вып. 15 (22). – Москва: Изд-во РГГУ, 2016.

5. *Лукашевич Н. В.* SentiRuEval-2016: преодоление временных различий и разреженности данных для задачи анализа репутации по сообщениям твиттера / Н. В. Лукашевич, Ю. В. Рубцова // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». Вып. 15 (22). – Москва : Изд-во РГГУ., 2015

6. *Васильев В. Г.* Выделение аспектов и классификация тональности Твитта с помощью правил / В. Г. Васильев, А. А. Денисенко, Д. А. Соловьев // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». Вып. 14 (21). - Москва: Изд-во РГГУ, 2015.

7. RusVectōrēs. RusVectōrēs: дистрибутивные семантические модели для русского языка [online]. Получено из: http://ling.go.mail.ru/dsm/en/

8. НКРЯ. Национальный корпус русского языка [online]. Получено из: http://ruscorpora.ru/en/

9. *Рубцова Ю. В.* Построение корпуса текстов для настройки тонового классификатора / Ю.В. Рубцова // Программные продукты и системы. – 2015 – №1 (109) – С. 72–78.

10. *Сегалович И.* A Fast Morphological Algorithm with Unknown Word Guessing Induced by a Dictionary for a Web Search Engine / И. Сегалович // MLMTA. США, Лас-Вегас – 2003. – С. 273–280.

11. *Рыбка Р. Б.* Морфо-синтаксический анализ на основе нейронных сетей и корпусных данных / Р. Б. Рыбка, А. Г. Сбоев, Д. В. Гудовских, И. А. Молошников // Procedia of The AINL IESM FRUCT Conference. – Санкт-Петербург, 2015. – С. 89–95.

12. *Молошников И. А.* Вероятностно-энтропийный подход поиска тематически схожих документов с созданием контекстно-семантического графа для изучения эволюции общественного мнения / И. А. Молошников, А. Г. Сбоев, Д. В. Гудовских, Р. Б. Рыбка // Procedia Computer Science. – Москва, 2015. – Т. 66. – С. 297–306.

Сбоев Александр Георгиевич – канд. физ.-мат. наук, ведущий научный сотрудник Курчатовского комплекса НБИКС-технологий.
Тел.: +7-926-253-72-17
E-mail: sag111@mail.ru

Воронина Ирина Евгеньена – д-р техн. наук, профессор кафедры программного обеспечения и администрирования информационных систем факультета прикладной математики, информатики и механики.
Тел.: +7-903-650-44-10
E-mail: irina.voronina@gmail.com

Гудовских Дмитрий Владимирович – инженер-исследователь Курчатовского комплекса НБИКС-технологий.
Тел.: +7-977-685-53-39
E-mail: dvgudovskikh@gmail.com

Селиванов Антон Александрович – инженер-исследователь Курчатовского комплекса НБИКС-технологий.
Тел.: +7-985-437-85-64
E-mail: aaselivanov.10.03@gmail.com

Sboev Aleksandr G. – PhD of Science, Senior Research Officer of the Kurchatov complex of the NBICS-technologies.
Tel.: +7-926-253-72-17
E-mail: sag111@mail.ru

Voronina Irina E. – PhD, Dr. habil. inTechnical Sciences, Professor of Software Development and Information Systems Administration Department of Applied Mathematics, Informatics and Mechanics Faculty.
Tel.: +7-903-650-44-10
E-mail: irina.voronina@gmail.com

Gudovskikh Dmitry V. – research engineer of the Kurchatov complex of the NBICS-technologies.
Tel.: +7-977-685-53-39
E-mail: dvgudovskikh@gmail.com

Selivanov Anton A. – research engineer of the Kurchatov complex of the NBICS-technologies.
Tel.: +7-985-437-85-64
E-mail: aaselivanov.10.03@gmail.com