УДК 004.633.2

Категоризация новостных публикаций с использованием нейронной сети

М. А. Сайфуллин [1], А. М. Сулейманова [2]

[1] mirat1618@gmail.com, [2] suleymanova.ufa@gmail.com

ФГБОУ ВО «Уфимский государственный авиационный технический университет» (УГАТУ)

Аннотация. В статье описан процесс программной реализации способа, позволяющего осуществлять отбор релевантных новостных публикаций для бухгалтерского персонала. Основа данного способа – искусственная многослойная нейронная сеть, созданная и обученная с помощью методов библиотеки «ruby-fann» на языке программирования Ruby. В заключении работы приводится сравнение полученных показателей качества классификации новостных публикаций четырех способов: обученных нейронных сетей (FANN и сети, обученной с помощью метода опорных векторов – SVM) и наивного байесовского классификатора (собственной разработки и программного расширения «nbayes»).

Ключевые слова: метод опорных векторов; FANN; SVM; категоризация; релевантность; обработка естественного языка; новости; бухгалтерия; Ruby; регулярное выражение.

ВВЕДЕНИЕ

Информированность о возможных, а также о происходящих изменениях в окружающей нас среде – это необходимость, продиктованная настоящим временем. Будучи невосприимчивыми к возможным скачкам кривой рыночного спроса, изменениям в предпочтениях потребителей, новым трендам в области производства и так далее – коммерческие предприятия снижают свою способность адаптации к новым условиям ведения бизнеса, что в дальнейшем может привести к снижению продаж (соответственно, снижению прибыли), к репутационному урону, утрате лидерства в своей сфере, оттоку клиентов – словом, к экономическим потерям.

При выполнении своих должностных обязанностей работники бухгалтерских отделов руководствуются множеством федеральных законов, нормативно-правовых актов, кодексов, регулирующих ведение бухгалтерского учета на предприятии, определяющих различные параметры – порядок исчисления налогов (сборов, страховых взносов, дивидендов), сроки уплат, порядок проведения проверок, формы отчетности, права и обязанности юридических и физических лиц. Однако, перечисленные параметры не статичны во времени: они подвергаются пересмотру с различной периодичностью законодательными органами – поэтому руководители финансовых отделов вынуждены проводить мониторинг законодательных инициатив.

Пресса – один из главных информационных ресурсов, которым пользуются руководители для получения свежих новостей: это различные газеты (печатные и электронные), сводки (дайджесты), специализированные журналы для профессионалов. Учитывая объем и различную направленность материалов, публикуемых в новостных средствах массовой информации, а также ограниченность функционала существующих инструментов-агрегаторов, можно говорить о том, что разработка собственного способа, позволяющего генерировать релевантный новостной поток для бухгалтерского персонала, не лишена смысла.

Цель проведенной работы – определить, какая из программных реализаций двух подходов (нейронная сеть или наивный классификатор Байеса) окажется наиболее

точной в категоризации новостных публикаций для бухгалтерского персонала.

Качество классификации каждого из методов оценивалось процентным соотношением количества верно категоризированных статей (то есть истинно положительных и истинно отрицательных результатов) к их общему количеству, далее – полученные данные каждого подхода сопоставлялись сравнительным методом.

ИСХОДНЫЕ МАТЕРИАЛЫ И ОБУЧЕНИЕ НЕЙРОННОЙ СЕТИ

В качестве материалов для подготовки обучающего набора данных для нейронной сети были использованы публикации следующих новостных ресурсов:

- «Банки.ру» (https://www.banki.ru/);
- «Клерк.ру» (https://www.klerk.ru/);
- «Главбух» (https://www.glavbukh.ru/);
- «Известия» (https://iz.ru/);
- «Журнал – СКБ Контур» (https://kontur.ru/articles);

и другие (обшее количество использованных информационных ресурсов: 14).

При ознакомлении с новостными статьями перечисленных веб-сайтов, с помощью специалистов в области бухгалтерского учета, была проведена сортировка публикаций на две группы:

1) Релевантные статьи – материалы, которые освещали возможные, планируемые или происходящие изменения (законодательного либо общего характера) в порядке проведения бухгалтерского учета в строительной сфере и представляли профессиональный интерес для специалистов в данной области;

2) Нерелевантные статьи – публикации общеэкономического характера, не имеющих какой-либо ценности (либо имеющих малую ценность) для бухгалтерского персонала в плане их профессиональной деятельности.

Общий объем собранных статей – 1 181 единица, 588 из которых отнесены к категории релевантных, 593 признаны нерелевантными . В дальнейшем, категория релевантных публикаций обозначается как «ham», категория нерелевантных – «spam». В категории «ham» большая часть статей затрагивала темы изменения норм налогового учета и рассматриваемых законодательных инициатив, в категории «spam» – темы внешней политики, курсов валют и фондовых бирж.

В табл. 1 приведен фрагмент перечня ключевых слов, сформированного ранее при разработке программной реализации наивного классификатора Байеса [1]:

Таблица 1

Фрагмент перечня ключевых слов

№	*Ключевое слово*
1	/\bизмен.{0,6}|обнов.{0,6}\b/
2	/\bиде.{0,4}|инициатив.{0,4}\b/
3	/\bфнс|ифнс|федеральн.{0,4}\ налог.{0,4}\ служб.{0,4}\b/
4	/\bгд|госдум.{0,4}|государственн.{0,4}\ дум.{0,4}\b
5	/\bпфр|пенс.{2,7}\ фонд.{0,3}\b/
6	/\bналог.{0,12}\b/
7	/\bнк|налого.{0,4}\ кодекс.{0,4}\b/
8	/\bсша|америк.*|соединенн.{1,3}\ штат.{1,3}\ америк.{1,3}\b/

Ключевые слова представлены в форме регулярных выражений с теми целями, чтобы при вычислениях: а) учитывались разные способы написания терминов и названий государственных структур («Министерство финансов», «Минфин»); б) учитывались однокоренные слова, измененные формы слов при их склонении или спряжении. Общее количество ключевых слов в полном перечне – 45.

В процессе обучения нейросети, при обработке каждой статьи, ее содержание векторизируется – для каждого ключевого слова рассчитывается количество его появления в тексте.

Графическое представление архитектуры сети и сформированных массивов данных указаны на рис. 1:

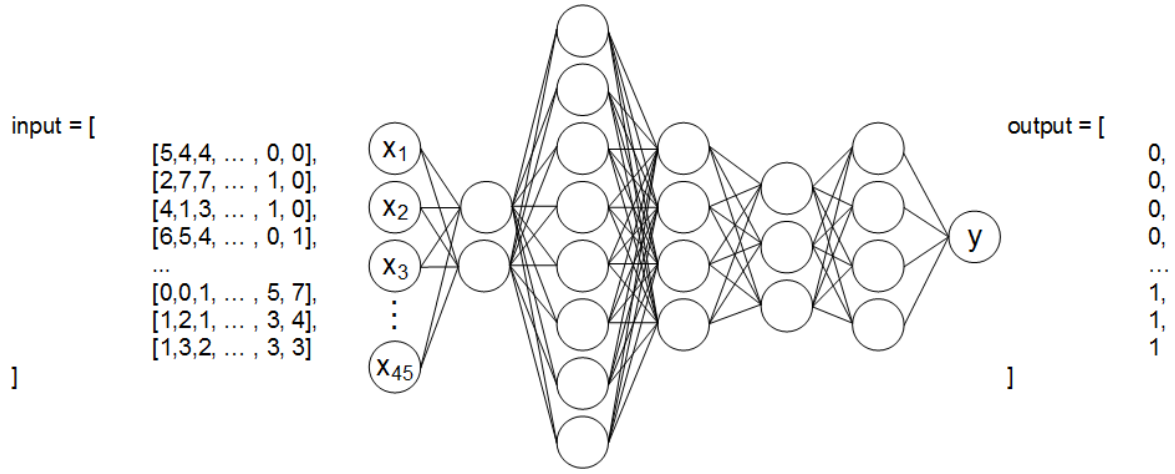

Рис. 1. Графическое представление архитектуры сети и наборов данных

Количество нейронов во входном слое – 45 (по количеству ключевых слов), в выходном слое – один нейрон, скрытых слоев – 4. Переменная input – массив массивов, содержащий количество появлений каждого ключевого слова в каждой статье, output – массив с данными о категории каждой статьи. Нулем обозначались материалы категории «ham», единицей – категория «spam». Таким образом, входными значениями для нейронной сети являлись подсчеты количества появления ключевых слов в статье, выходными – категория статьи. Первое значение (т.е. нулевой индекс) массива input соответствовало первой статье, первое значение массива output – ее категории и так далее.

СРАВНЕНИЕ ПОКАЗАТЕЛЕЙ КАЧЕСТВА КЛАССИФИКАЦИИ

Сравнение точности категоризации новостных статей, проводимых байесовским классификатором и нейронной сетью – заключительный этап проведенного исследования. Качество классификации оценивалось соотношением верно классифицированных публикаций к общему их количеству в наборе:

$$Accuracy = \frac{True\ ham + True\ spam}{Total\ number},$$

где: accuracy – точность классификации; true ham – количество верно классифицированных релевантных статей; true spam – количество верно классифицированных нерелевантных статей; total number – общее количество статей в наборе.

Далее, было сопоставлено качество работы самостоятельно разработанного наивного байесовского классификатора, наивного байесовского классификатора (расширение «Nbayes») и нейронных сетей (SVM и FANN) – обозначаются в дальнейшем как способы №1, №2, №3, №4 соответственно – на следующих данных:

– набор, состоящий из двух статей (релевантной [2] и нерелевантной [3]), не использованных ранее;

– исходный набор статей, состоящий из 1 181 статьи;

– новый набор статей, не применявшийся ранее и состоящий из 20 статей (10 публикаций на каждую из категорий).

Результаты классификации релевантной и нерелевантной статей представлены в табл. 2:

Таблица 2

Результаты категоризации релевантной и нерелевантной статей

	Способ	*Вероятность релевантности*	*Вероятность нерелевантности*	*Присвоенная категория*
Релевантная статья	№1	0.62	0.38	HAM
	№2	0.502	0.498	
	№3	–	–	
	№4	–	–	
Нерелевантная статья	№1	0.002	0.998	SPAM
	№2	0.499	0.501	
	№3	–	—	
	№4	–		

Вероятности у способов №3, №4 не указаны, так как нейронная сеть, использующая метод опорных векторов, представляет выходную информацию в бинарном виде (0 – категория «ham», 1 – категория «spam»); нейронная сеть на основе FANN представляет результат в виде десятичного числа в диапазоне от 0 до 1, которое далее округлялось до целого. Как видно из данных таблицы, все способы верно определили категории публикаций из первого набора. Результаты категоризации исходного набора статей представлены в табл. 3:

Таблица 3

Результаты категоризации исходного набора статей

	Количество статей (шт.)	*Способ*	*Признаны релевантными (шт.)*	*Признаны нерелевантными (шт.)*
Релевантные статьи	588	№1	468	102
		№2	578	10
		№3	510	78
		№4	528	60
Нерелевантные статьи	593	№1	97	496
		№2	33	560
		№3	88	505
		№4	103	490

На основе данных, представленных в таблице, можно рассчитать точность классификации каждого из подходов: 81,626%, 96,359%, 85,944% и 86,198% для способов №1, №2, №3 и №4 соответственно. Наилучший результат продемонстрирован наивным байесовским классификатором на базе готового программного дополнения «Nbayes», далее – сеть на базе FANN, сеть на основе SVM и собственная реализация байесовского классификатора.

Таблица 4

Результаты категоризации нового набора статей

	Количество статей	*Способ*	*Признаны релевантными (шт.)*	*Признаны нерелевантными (шт.)*
Релевантные статьи	10	№1	9	1
		№2	10	0
		№3	10	0
		№4	10	0

Продолжение табл. 4

	Количество статей	*Способ*	*Признаны релевантными (шт.)*	*Признаны нерелевантными (шт.)*
Нерелевантные статьи	10	№1	1	9
		№2	8	2
		№3	1	9
		№4	1	9

Показатели точности классификации на основе данных таблицы 4 следующие: 90%, 60%, 95% и 95% для способов №1, №2, №3 и №4 соответственно. При работе с новыми данными качество классификации способа №2 резко ухудшилось (8 из 10 нерелевантных статей были распознаны неверно). Искусственные нейронные сети показали наиболее высокий и стабильный уровень качества категоризации на протяжении всех экспериментов.

ЗАКЛЮЧЕНИЕ

В ходе выполнения данного исследования были поставлены и выполнены следующие задачи:

– формирование обучающего набора данных;

– подготовка используемого инструмента (обучение искусственной нейронной сети с использованием программных методов, предоставляемых библиотекой «ruby-fann»);

– проведение классификации подготовленных наборов данных с помощью обученной нейронной сети;

– сопоставление качества классификации обученной нейронной сети и ранее разработанных и испытанных способов (наивного классификатора Байеса (собственной программной реализацией и готовым дополнением), обученной нейронной сети на базе метода опорных векторов – SVM).

На основе проведенного итогового анализа качества классификации изученных подходов, было установлено, что обученные искусственные нейронные сети (на основе FANN и SVM) при проведении программных экспериментов более точно категоризировали новостные публикации по сравнению с вероятностным методом – наивным классификатором Байеса.

СПИСОК ЛИТЕРАТУРЫ

1. **Сайфуллин М.А., Сулейманова А.М**. Реализация наивного байесовского классификатора новостных публикаций в финансовой сфере на языке программирования Ruby. ИТ в управлении и экономике, 2019. С. 35–36. [M.A. Saifullin, A.M. Suleimanova, The realization of a Naïve Bayesian Classifier for financial news articles in Ruby programming language, (in Russian): Information technology in management and economics, 2019, pp. 35–36]

2. Минфин опубликовал инструкцию по 115-ФЗ для бухгалтерских фирм. СКБ Контур. [Электронный ресурс]. URL: https://kontur.ru/articles/5675 (дата об-ращения 05.12.2019). [Ministry of Finance has published an instruction on №115 federal law for accounting companies. (2019, Dec. 05). SKB Kontur. [Online]. Available: https://kontur.ru/articles/5675]

3. Объем вложений в недвижимость на 30% превысит уровень 2018 года. РБК Pro. [Электронный ресурс]. URL: https://pro.rbc.ru/demo/5dea80439a794720e195868e (дата обращения 05.12.2019). [The volume of real estate investments will exceed the level of 2018 for 30%. (2019, Dec. 05). RBK Pro. [Online]. Available: https://pro.rbc.ru/demo/5dea80439a794720e195868e]

4. **Турканов Г.И., Щепин Е.В.** Классификатор Байеса для переменного количества признаков. Труды МФТИ. 2016. № 4. С. 8. [G. I. Turkanov, E.V. Shchepin, Bayes classifier for a variable number of features, (in Russian): Trudy MFTI, 2016, vol. 3, p. 8]

5. **Шанов С.В., Чупин П.Г., Афонин А.Ю.** Применение байесовского классификатора для определения тематики текста. Моделирование, оптимизация и информационные технологии. 2018. №1. С. 133. [S.V Shanov, P.G. Chupin, A.Yu. Afonin, Application of the bayesov classifier for the definition of the thematics of the text, (in Russian): Modelirovanie optmizatsiia i informatsionnye tekhnologii, 2018, vol. 1, p. 133]

ОБ АВТОРАХ

САЙФУЛЛИН Мират Азатович, магистрант каф. АСУ.

СУЛЕЙМАНОВА Алла Маратовна, доцент каф. АСУ. Канд. техн. наук (УГАТУ, 1993).

METADATA

Title: News articles categorization using a neural network

Authors: M. A. Sayfullin [1], A. M. Suleymanova [2]

Affiliation:

Ufa State Aviation Technical University (UGATU), Russia.

Email: [1] mirat1618@gmail.com, [2] suleymanova.ufa@gmail.com

Language: Russian.

Source: Molodezhnyj Vestnik UGATU (scientific journal of Ufa State Aviation Technical University), no. 1 (22), pp. 124-128, 2020**.** ISSN 2225-9309 (Print).

Abstract: The article describes a process of a programming realization of a method enabling to select relevant news for accountants personnel. The method is based on a multilayer artificial neural network created and trained using “ruby-fann” library in Ruby programming language. At the conclusion, there is a comparison of resulting quality indicators of four examined approaches: trained neural networks (based on FANN and SVM) and naïve Bayesian classifier (an independently developed one and the other one based on "nbayes" program extension).

Key words: support vector machine; FANN; categorization; natural language processing; news publications; accounting; Ruby; regular expression.

About authors:

SAYFULLIN, Mirat Azatovich, Master’s student, Dept. of Automated Systems.

SULEYMANOVA, Alla Maratovna, Associate professor, Dept. of Automated Systems. PhD in Technique (UGATU, 1993).