УДК :004.91

к.ф.-м.н., доц. **Шпінарева І.М. (**ОНУ ім.І.І.Мечникова)
Геренко О.А. (ОНУ ім.І.І.Мечникова)
Морозова К.Ю. (ОНУ ім.І.І.Мечникова)

КЛАСИФІКАЦІЯ ТЕКСТІВ НА ПРИРОДНІЙ МОВІ ЗА ДОПОМОГОЮ НЕЙРОННОЇ МЕРЕЖІ

У статті були досліджені проблеми класифікації текстів на природній мові з використанням методів машинного навчання, зокрема за допомогою нейронних мереж. Визначено актуальність досліджень в направленні подання текстового документа у вигляді математичного вектора. У якості векторної моделі використовується "мішок термів". У статті розглядаються підходи побудови векторної моделі статистичними мірами TF-IDF або TF-SLF і класифікації текстів нейронними мережами прямого поширення. Проводиться порівняння ефективності класифікації для кожного з підходів при різних ознаках і обсягах вибірок. Процес класифікації текстів проходить в три етапи. На етапі передобробці в вхідному тексті видаляються стоп-слова і виконується стемінг. На етапі визначення ознак тексту обчислюються статистичні міри TF-IDF або TF-SLF. На третьому етапі класифікація виконується двошаровою нейронною мережею з прямими зв'язками і безперервною функцією активації (сигмоид). Мережа була навчена методом зворотного поширення ошібок. Зроблений аналіз і порівняння якості роботи різних методів класифікації за такими характеристиками, як точність, повнота.

Ключові слова: автоматична класифікація текстів, статистичні міри, TF-IDF, TF-SLF, нейронна мережа, навчання нейронної мережі.

Вступ. Стрімке зростання інформації в мережі інтернет природним чином спричинив проблеми пошуку і впорядкування інформації. Сьогодні в електронних сховищах по всьому світу містяться терабайти інформації, яку бажано розподіляти за тематичними каталогами. Інформаційних джерел стає все більше і отримати з цього океану потрібні знання стає все важче. У зв'язку з цими проблемами все більш актуальною стає задача класифікації інформації, аналізу її і на основі зібраних знань генерація для користувача невеликого, зручного для сприйняття тексту, що містить головну думку статті, тобто реферат.

У цій статті розглядається задача класифікації текстів на природній мові як перший крок до вирішення проблеми автоматичного реферування тексту. Крім того класифікація текстів необхідна для поділу сайтів по тематичним каталогам, боротьби зі спамом, розпізнавання емоційного забарвлення текстів та персоніфікації реклами.

Мета цієї статті − розглянути особливості застосування статистичних метрик зважування термів для класифікації текстів нейронними мережами.

Постановка задачі класифікації полягає в тому, що задано безліч документів D = {d_1, d_2, ..., $d_{|D|}$}, безліч категорій (класів) C = {c_1, ., $c_{|C|}$} і невідома цільова функція F: C ˣ D → {0,1}. Необхідно побудувати класифікатор F ', максимально близький до F.

Виклад основного матеріалу. Задачу класифікації на класи успішно вирішують з використанням різних методів машинного навчання. Для застосування алгоритмів машинного навчання текстовий документ необхідно представити у вигляді математичного вектора. В якості векторної моделі використовується "мішок термів"[1]. Вибір правильного уявлення слів в тексті мають вирішальне значення для отримання хорошої класифікації документів.

У статті [2] автор пропонує проводити зважування термів метрикою хі-квадрат, в результаті кожному об'єкту класу приписується числовий коефіцієнт, який вказує на його дискримінуючу силу. Як показують результати, за допомогою цього методу можна досягти досить високої точності для стемм, а не конкретних словоформ і отримані коефіцієнти мають великий розкид значень.

У статті [3] автори розробили модель засновану на TF-IDF, яка демонструє гарну продуктивність за класифікацією сортів мови, але не ефективна для класифікації авторів Twitter в тестових даних.

У даній роботі пропонується проводити зважування термів статистичними метриками TF-IDF або TF-SLF та класифікувати тексти за трьома тематиками: художньої, спортивної та наукової.

Процес класифікації текстів складається з 3 етапів (рис. 1).

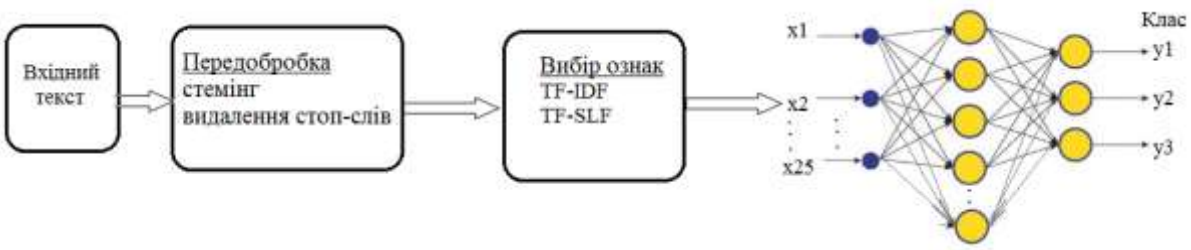

Рис. 1. Процес класифікації текстів

На першому етапі текст проходить предобработку, яка полягає в перекладі вхідного тексту в нижній регістр, видалення стоп-слів, тобто слів, які не несуть сенсу і стемінгу (відкидання змінною частини слова).

На другому етапі шляхом зважування термів з використанням статистичних мір необхідно витягти з тексту визначальні ознаки. Зважування термів означає обчислення ваг слів за допомогою мір TF-IDF або TF-SLF [4].

Міра TF-IDF визначає важливість пропозиції на основі частот слів, що входять до нього при цьому для того, щоб не враховувати слова які зустрічаються у всіх документах використовується зворотна частота документа IDF, яка дорівнює логарифму відношення кількості документів до кількості документів в яких вони зустрілись. Якщо слово часто зустрічається в даному документі, то міра TF-IDF збільшує вагу слова, і зменшує вагу слова, якщо слово часто зустрічається в багатьох документах. Вага деякого слова пропорційна кількості вживання цього слова в документі і обернено пропорційна частоті вживання слова в інших документах колекції. [5,6]

Міра TF-IDF розглядає важливість терма в рамках всього корпусу документів. При такій оцінці ігнорується важливість терма в рамках окремо взятої категорії. Для подолання даного обмеження використовується метрика TF-SLF, засновану на наступних припущеннях:

– терм є важливим в рамках категорії, якщо він зустрічається в більшості документів даної категорії;

– оцінка терма знижується, якщо він є важливим для декількох категорій.

Для розрахунку числового значення TF-IDF необхідно обчислити: число входжень *i* слова в документі, загальна кількість слів у документі, кількість документів в корпусі, кількість документів в яких зустрічається *i* слово.

Відносна частота зустрічальності *i* слова в тексті *d*:

$$TF(w;d) = \frac{n_i}{\sum_k n_k}, \quad (1)$$

де n_i– число входжень *i* слова в документ;

$\sum_k n_k$ – загальне число слів в документі.

Інверсна частота w в довільному безлічі текстів D:

$$\mathrm{DF}(w;D) = log\frac{|D|}{|(d_i|w_i)|}, \quad (2)$$

де D – кількість документів в корпусі;

$(d_i|w_i)$ – кількість документів в яких зустрічається *i* слово.

Вага TF-IDF (w; d; D) слова w є Wd тексту d в загальній тематичної колекції текстів D визначається за формулою:

$$\mathrm{TFIDF(w; d; D)} = \mathrm{TF(w; d)} * \mathrm{IDF(w; D)}. \quad (3)$$

Тепер для міри TF-SLF визначимо нормалізовану частоту зустрічальності терма t в категорії c:

$$NDF_{tc} = \mathrm{n}_{tc} / Nc\ , \quad (4)$$

де n_{tc} – число документів категорії c в яких зустрічається хоча б раз терм t;

Nc – кількість документів в категорії c;

C – безліч категорій в корпусі документів.

Оцінка $\mathrm{NDF_{tc}}$ локальна для категорії. Для отримання глобальної оцінки $\mathrm{R_t}$ в рамках всього корпусу всі $\mathrm{NDF_{tc}}$ підсумовуються:

$$R_t = \sum_{c \in C} NDF_{tc}. \quad (5)$$

Логарифм суми частот терма t обчислюється як:

$$SLF_t = log\ \frac{|C|}{R_t}\ . \quad (6)$$

SLF дозволяє усунути дисбаланс між категоріями з малим числом документів і категоріями з великим числом документів.

Оцінка TF-SLF для терма t обчислюється як:

$$TFSLF_t = TF_t * SLF_t \quad (7)$$

На рис. 2 представлені важливість терма в тексті про художника Мікеланджело, тобто результати мір TF-IDF і TF-SLF.

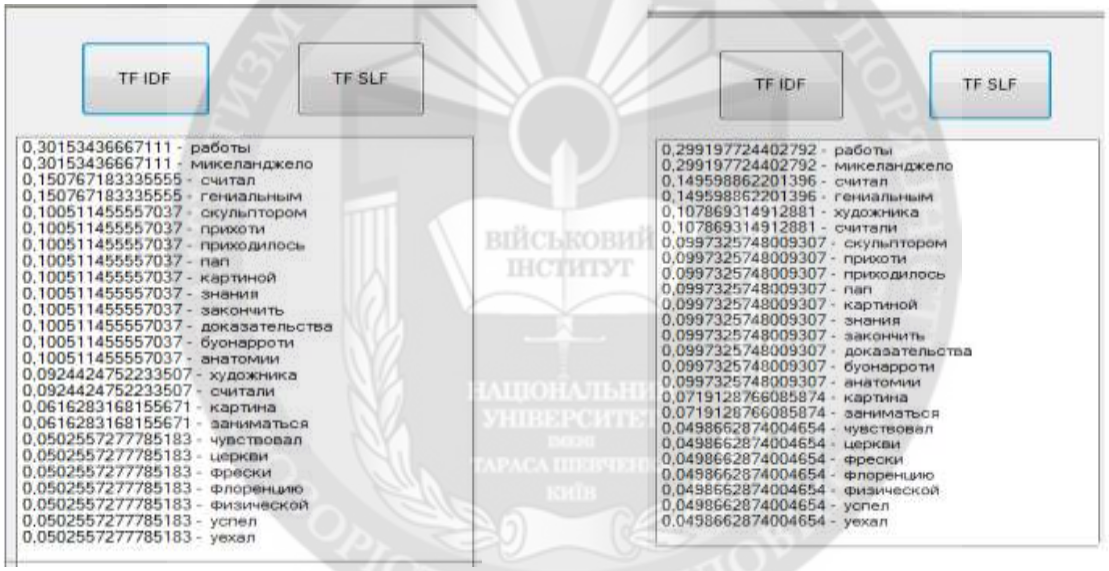

Рис. 2. Результати мір TF-IDF і TF-SLF

Після виконання другого етапу ми отримуємо для кожного слова його вагу в тексті. Отримані нормалізовані числові значення сортують в порядку убування і вибирають ключові слова з максимальною числовою вагою. Наприклад, 25 важливих ключових слів.

Третій етап класифікації полягає у виборі класифікатора. В якості класифікатора в роботі використовується нейронна мережа прямого поширення [7,8]. Нейронні мережі прямого поширення є найбільш популярними і добре зарекомендували себе в різноманітних задачах. Вони дозволяють виробляти нелінійну апроксимацію довільної функції (в нашому випадку функції приналежності до класу) по набору прикладів. Нейрони в даних мережах об'єднані в шари. Мережа складається з довільного числа шарів. Нейрони кожного шару з'єднуються з нейронами попереднього і подальшого шарів за принципом «кожен з кожним».

В роботі використовується двошарова нейронна мережа з прямими зв'язками і безперервною функцією активації (сигмоид). Вхідні дані (25 числових значень) подаються на прихований шар нейронної мережі (30 нейронів), а потім на вихідний шар нейронної мережі

(що складається з 3 нейронів), який визначає ймовірність відповідності тексту одному з трьох класів. Спочатку ваги генерувалися випадковим чином в інтервалі [-0.5,0.5].

Мережа була навчена методом зворотного поширення помилок. Значення коефіцієнта навчання прийнято 0,5. При навчанні мережа проходила 10000 епох. В якості навчальної вибірки обрані тексти художньої, спортивної та наукової тематики. Додаток реалізований в Microsoft Visual Studio 2013 на C #. Результати класифікації представлені на рис. 3.

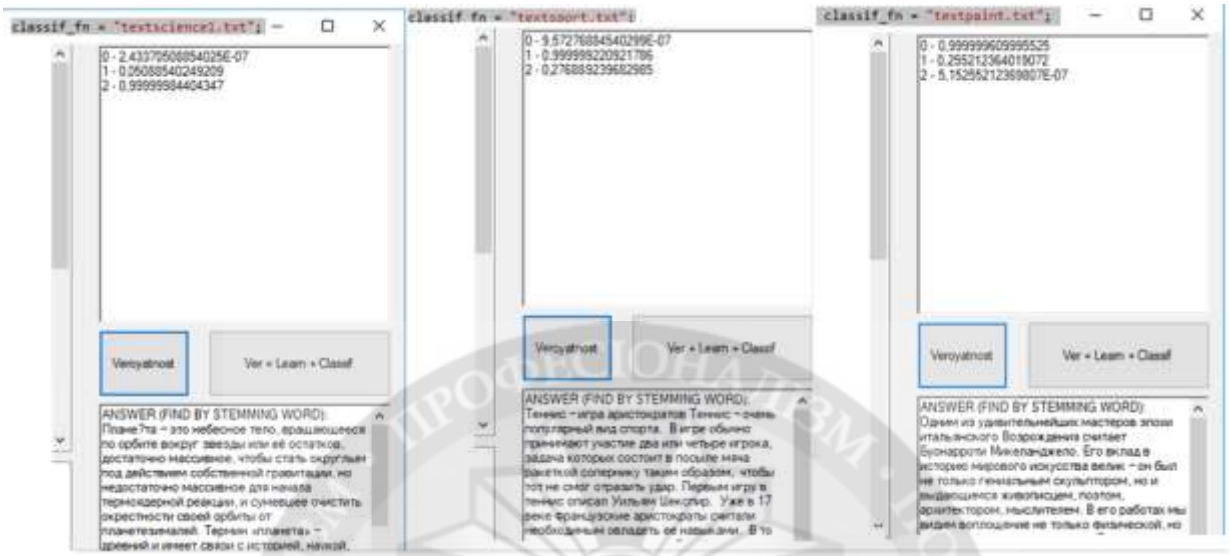

Рис. 3. Результати класифікації нейронної мережі

В якості метрик правильності класифікації текстів були обрані точність (precision) і повнота (recall) [9]. Точність в межах класу – це частка текстів, які дійсно належать даному класу, щодо всіх текстів, зарахованих класифікатором до цього класу. Повнота системи – відношення числа знайдених класифікатором текстів, що належать класу, до числа всіх текстів цього класу в тестовій колекції. Повнота і точність визначаються за наступними формулами:

$$Precision = \frac{TP}{TP + FP}, \qquad Recall = \frac{TP}{TP + FN}, \tag{5}$$

де TP – істино-позитивне рішення;
TN – істино-негативне рішення;
FP – ложно-позитивне рішення;
FN – ложно-негативне рішення.

Для 1600 текстів за 3 тематиками було зроблено 3 класу (1200 текстів – навчальна вибірка, 400 – тестова). У кожній навчальній і тестовій підвибірці містилася однакова кількість текстів про науку, мистецтво, спорт. Для кожної групи розраховувалися оцінки ефективності, а потім вираховувались їх середні значення. Таким чином, вийшли усереднені оцінки ефективності. Результати усереднених оцінок ефективності класифікації для розглянутих ознак в табл. 1.

Таблиця 1

Оцінка ефективності класифікатора

Векторна модель	TF-IDF		TF-SLF	
	Точність, %	Повнота, %	Точність, %	Повнота, %
Спорт	84,5	88,25	85,3	91,5
Мистецтво	85,2	87,4	86,2	89,4
Наука	86,3	88,5	87,4	90,6

Висновок. Отже, можна зробити висновок, що вибір правильного уявлення слів в тексті мають вирішальне значення для отримання більш якісної класифікації документів. Класифікація текстів виконувалася багатошаровою нейронною мережею на вхід, якій подавалися значення обчислені мірами TF-IDF і TF-SLF. Кращі результати показала нейронна мережа у якій вхідні значення обчислені мірою TF-SLF. Застосування нейронних мереж, по-перше, суттєво підвищує якість рішення багатьох стандартних задач класифікації текстів, по-друге, знижує трудомісткість при роботі безпосередньо з текстами.

ЛІТЕРАТУРА:

1. Text summarization by machine learning, Brno, Spring 2016 / Matej Gallo, [Електронний ресурс] − Режим доступу: https://is.muni.cz/th/ 422328/fi_b/bachelor-thesis.pdf

2. Распределение хи-квадрат и взвешивание термов / Яцко В.А. − Международный научный журнал "Символ науки" №3, 2016. [Електронний ресурс] − Режим доступу: https://cyberleninka.ru/article/v/ raspredelenie-hi-kvadrat-i-vzveshivanie -terminov

3. Using the text relationship map (trm) for automatic summarization / Kanishcheva O.V., [Електронний ресурс] − Режим доступу: http://science.lpnu.ua/sites/default/files/journal-paper/2017/ jun/3223 /13770englis harticlekanishcheva.pdf

4. Feature Extraction for Classification of Text Documents/ Abdur Rehman, Haroon A. Barbi, Mehreen Saeed, 2012 [Електронний ресурс] − Режим доступу: ibrarian.net/.../ Feature_Extraction_Algorithms_ for_Classifica.

5. Інформаційна технологія автоматизованого анотування та реферування цифрових текстів/ О.В. Бармак, О.В. Мазурець, А.В. Живілік − Вісник Хмельницького національного університету, №4, 2017 С.147-157.

6. Automatic text summarization using supervised machine learning technique for hindi langauge / N. Desai, P. Shah − International Journal of Research in Engineering and Technology, 2016 [Електронний ресурс] − Режим доступу: http://esatjournals.net/ijret/2016v05/i06/IJRET20160506065.pdf

7. Хайкин С. Нейронные сети: полный курс, 2-е издание: Пер. с англ. − М.: Издательский дом «Вильямс», 2006. – 1104 с.

8. Круглов В. В. Искусственные нейронные сети. Теория и практика. – 2-е изд./В.В. Круглов , В.В. Борисов – Горячая линия,Телеком, 2002. – 382 с.

9. Алгоритмы для интернета: Автоматическая классификация текстов / Лифшиц Ю. – 2006. [Електронний ресурс] − Режим доступу: yury.name/ internet/ 06ianote.pdf

10.Питально-відповідна довідкова система з підтримкою голосової функції / О.А.Геренко, І.М.Шпінарева, К.Ю.Морозова – Збірник наукових праць Військового інституту Київського національного університету ім. Т.Шевченка. − К.: ВІКНУ, 2017, Вип. №55 – С.119-124.

11.Модифікований метод автоматичного реферування текстів з використанням тематично зв'язаного ранжування речень / Заболотня Т.М., Федченко Н.В. – Проблеми інформаційних технологій № 19, 2016 – С.141-147, [Електронний ресурс]. − Режим доступу: http://pit.hntu.com.ua

12.Компактифицированный горизонтальный граф видимости для сети слов / Д. В. Ландэ, А. А. Снарский – Труды Международной научной конференции «Интеллектуальный анализ информации ИАИ-2013. Знания и рассуждения». – Киев: КПИ,2013. – С. 158–164.

REFERENCES:

1. Matej Gallo, Text summarization by machine learning, (2016), Brno, Spring. [Electronic resource] – Access mode: https://is.muni.cz/th/ 422328/fi_b/bachelor-thesis.pdf.

2. Yatsko VA, Distribution of chi-square and term weighting, (2016), International Scientific Journal "The Symbol of Science"No.3. [Electronic resource] – Access mode: https://cyberleninka.ru /article/v/ raspredelenie-hi-kvadrat-i-vzveshivanie -terminov.

3. Kanishcheva O.V. Using the text relationship map (trm) for automatic summarization, (2015), [Electronic resource] – Access mode: http://science.lpnu.ua/sites/default/files/journal-paper/2017/jun/3223 /13770englis harticlekanishcheva.pdf.

4. Abdur Rehman, Haroon A. Barbi, Mehreen Saeed, Feature Extraction for Classification of Text Documents, (2012), [Electronic resource] – Access mode: ibrarian.net/.../Feature_ Extraction_ Algorithms_for_Classifica.

5. O.V. Barmak, O.V. Mazurets, AV Zhivilik, Information technology of automated annotation and digital text retrieval, (2017), Bulletin of the Khmelnytsky National University, №4 − C.147-157

6. N. Desai, P. Shah, Automatic text summarization using supervised machine learning technique for hindi langauge, (2016), International Journal of Research in Engineering and Technology, [Electronic resource] – Access mode: http://esatjournals.net/ijret/2016v05/i06/IJRET20160506065.pdf

7. Khaikin Simon, Neural networks: a full course,2 nd edition, (2006), Trans. with anrl. – M.: Publishing house "Williams», – 1104 p.

8. V.V. Kruglov, V.V. Borisov, Artificial neural networks. Theory and practice, 2 nd ed., (2002), Hot line, Telecom,382 p.

9. Lifshits Yu., Algorithms for the Internet: Automatic Classification of Texts, (2006), [Electronic resource] – Access mode: yury.name/ internet/ 06ianote.pdf

10. O. A. Gerenko, I. M. Shpinareva, K. Y. Morozova, Question-relevant background systems with the support voice features, (2017), Collection of scientific works of the Military Institute of the Kyiv National University named after. T. Shevchenko, K.: The window, Vip. №55, C.119-124

11.Zabolotnya T.M., Fedchenko N.V., Modified method of automatic referencing of texts using the thematically linked ranking of sentences, (2016), Problems of Information Technology No.19, P.141-147 [Electronic resource]. – Access mode: http://pit.hntu.com.ua

12.DV Lande, AA Snarskii, Compactified horizontal graph of visibility for a network of words, (2013), Proceedings of the International Scientific Conference "Intellectual Analysis of Information IAI-2013. Knowledge and reasoning. «, Kiev: KPI, P. 158-164.

Рецензент: д.т.н., доц. Гунченко Ю.О., професор кафедри математичного забезпечення комп'ютерних систем, Одеський національний університет імені І.І. Мечникова

к.ф.-м.н., доц. Шпінарева І.М., Геренко О.А., Морозова К.Ю.

КЛАССИФИКАЦИЯ ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ С ПОМОЩЬЮ НЕЙРОННОЙ СЕТИ

В статье были исследованы проблемы классификации текстов на естественном языке с использованием методов машинного обучения, в частности с помощью нейронных сетей. Определена актуальность исследований в направлении представления текстового документа в виде математического вектора. В качестве векторной модели используется "мешок термов". В статье рассматриваются подходы построения векторной модели статистическими мерами TF-IDF или TF-SLF и классификации текстов нейронными сетями прямого распространения. Рассматривается мера TF-IDF как важность терма в рамках документа, при этом игнорируется важность терма в рамках отдельно взятой категории. Терм TF-SLF является важным в рамках категории, если он встречается в большинстве документов данной категории. Производится сравнение эффективности классификации для каждого из подходов при различных признаках и объемах выборок. Процесс классификации текстов проходит в три этапа. На этапе предобработке в входном тексте удаляются стоп-слова и выполняется стемминг. На этапе определения признаков текста вычисляются статистические меры TF-IDF или TF-SLF. Полученные нормализованные числовые значения сортируют в порядке убывания и выбирают ключевые слова с максимальным числовым весом На третьем этапе классификация выполняется двухслойной нейронной сетью с прямыми связями и непрерывной функцией активации (сигмоид). Вектор из 25 значений подается на скрытый слой нейронной сети, состоящий из 30 нейронов, а затем на выходной слой нейронной сети, который определяет вероятность соответствия текста одному из трех классов. Сеть была обучена методом обратного распространения ошибок. Произведен анализ и сравнение качества работы различных методов классификации по таким характеристикам, как точность, полнота. Результаты анализа показали, что качество классификации при одинаковых параметрах многослойной нейронной сети были лучше, когда входной вектор определен мерой TF-SLF.

Ключевые слова: классификация текстов, статистические меры, TF-IDF, TF-SLF, нейронная сеть, обучение нейронной сети.

Ph.D. Shpinareva I.M., Gerenko O.A., Morozova K.Y.

CLASSIFICATION OF TEXTS IN NATURAL LANGUAGE USING A NEURAL NETWORK

The article investigate the problems of texts classification in a natural language using the methods of machine learning, in particular with the help of neural networks. The relevance of research in the direction of the presentation of a text document in the form of a mathematical vector. A "bag of terms" is used as a vector model. The article consider the approaches to the construction of a vector model by static measures TF-IDF or TF-SLF and the texts classification by neural networks of direct propagation. TF-IDF indicates the importance of the term in the document, while the meaning of the term in the separatly taken is category is ignored. The term TF-SLF is important in the category if it is found in most documents of this category. A comparison of the classification effectiveness for each approach is performed for different characteristics and sample sizes. The process of texts classifying occurs in three stages. At the preprocessing stage in the input text stop words are deleted and stemming is performed. At the stage of determining text characteristics the statistical measures TF-IDF or TF-SLF are calculated. The obtained normalized numerical values are sorted in descending order and the keywords with the maximum numerical weight are selected. At the third stage, the classification is performed by a two-layer neural network with direct connections and a continuous activation function (sigmoid). A vector of 25 values is fed to the hidden layer of the neural network, consisting of 30 neurons, and then to the output layer of the neural network, which determines the probability of matching the text with one of the three classes. The network was trained by the method of back propagation of errors. Analyzed and comparison the work quality of various classification methods on such characteristics as accuracy, completeness. The analyze results showed that the quality of classification for the same parameters of the multilayer neural network was better when the input vector was determined by the TF-SLF measure.

Keywords: texts classification, statistical measures, TF-IDF, TF-SLF, neural network, neural network training.