УДК 519.248; 81`32
DOI https://doi.org/10.24144/2616-7700.2024.45(2).115-125

Є. В. Турчин[1], Ю. С. Федорченко[2]

[1] Дніпровський національний університет імені Олеся Гончара,
доцент кафедри статистики й теорії ймовірностей,
кандидат фізико-математичних наук, доцент
evgturchyn@gmail.com
ORCID: https://orcid.org/0000-0001-8947-1359

[2] Дніпровський національний університет імені Олеся Гончара,
студент 4-го курсу механіко-математичного факультету
yura.fedor4enko@gmail.com
ORCID: https://orcid.org/0009-0003-6191-4033

РОЗПОДІЛ ЧАСТИХ СЛІВ У КОРОТКИХ ТЕКСТОВИХ ПОВІДОМЛЕННЯХ

Розглядається задача про розподіл частот слів у текстовому корпусі, що складається з коротких повідомлень (акцент зроблено на частих словах). Серед декількох сімей розподілів знайдені найбільш адекватні (використовувався критерій хі-квадрат, а також порівняння за допомогою статистик AIC та BIC).

Ключові слова: розподіл частот слів, математична лінгвістика, критерій хі-квадрат, критерій AIC, критерій BIC.

1. Вступ. У математичній лінгвістиці та інтелектуальному аналізі тексту порівняно багато вивчались різні аспекти коротких текстів — це задачі класифікації, кластерного аналізу, ідентифікації автору, тематичного моделювання та інші (див, наприклад, [1–11]) Але задача про розподіл частот слів у коротких текстах явно вивчена недостатньо. Мета даною статті — заповнити (хоча б частково) цю прогалину. Ми з'ясуємо, які сім'ї дискретних розподілів краще підходять для описання розподілу частот частих слів у великій колекції коротких текстів.

2. Основні результати. Для дослідження було взято випадкові 10% від набору даних [12]. Отриманий таким чином набір даних складається з великої кількості порівняно коротких англомовних текстів (це SMS-повідомлення, Telegram-повідомлення та електронна пошта). Слід зазначити, що приблизно 40% цих текстів є спам-повідомленнями. Попередня обробка текстів з нашого набору даних включала, зокрема, видалення чисел та знаків пунктуації, видалення так званих стоп-слів (це займенники, прийменники, сполучники, різні форми допоміжних дієслів та ще декілька інших слів) і стемінг (тобто виділення основи слова). Хоча фактично ми працюємо з основами слів, але надалі заради зручності ми будемо замість фрази "основа слова" писати просто "слово".

Після попередньої обробки у текстовому корпусі залишилося приблизно 35700 унікальних слів та приблизно 480000 слів усього, документів усього — близько 4700. Кількість слів у документі корпуса знаходиться у межах від 0 до приблизно 10300, середнє число слів у документі дорівнює 100.8.

Для моделювання частот із слів, що зустрічаються в усьому текстовому корпусі як мінімум 330 разів, було відібрано 46 слів. Зазначимо, що серед обраних 46 слів є різні слова — це як "маркери" спаму (зокрема, "free", "save", "money"), так і просто часті слова.

Для кожного фіксованого слова **w** утворимо вибірку $\xi = (\xi_1, ..., \xi_n)$, де ξ_i — кількість повторень слова **w** у i-му документі нашого текстового корпуса. Що можна сказати про розподіл ξ_i?

Для всіх наших слів має місце дуже сильно виражене явище "роздутості" маси нуля (zero inflation) — для емпіричного розподілу частот маса нуля часто більше 0.9. Це природно, оскільки для великої кількості документів конкретне слово **w** зустрічатися у них не буде. Тому надалі ми моделюємо не розподіл ξ_i, а розподіл елементів перетвореної вибірки. Робимо наступне: з вибірки ξ вилучимо всі нулі, отримаємо вибірку $\eta = (\eta_1, ..., \eta_j, ...\eta_m)$, а потім зсунемо на 1 всі її елементи, тобто перейдемо до вибірки

$$\zeta = (\zeta_1, \ldots, \zeta_j, \ldots, \zeta_m),$$

де $\zeta_j = \eta_j - 1$ (і ζ_j вже набувають значення $0, 1, 2, \ldots$). І надалі нашою задачею буде моделювання розподілу ζ_j.

Вартою уваги особливістю емпіричних розподілів частот слів є те, що для багатьох слів "довжина емпіричного хвоста" є високою або навіть надзвичайно високою. Точніше,

$$\frac{\max\{\zeta\}}{q_{0.9}(\zeta)} > 10$$

для 15 з 46 слів (через q_α позначено емпіричну α-квантиль), та

$$\frac{\max\{\zeta\}}{q_{0.9}(\zeta)} \geq 5$$

для 40 з 46 слів.

Наведемо тепер розподіли, які будуть використовуватись для моделювання частот слів у перетвореній вибірці (тобто вибірці ζ).

Розподіл Зіхеля (Sichel distribution, див. [13]) з параметрами ω, κ, γ означається формулою

$$P(l; \omega, \kappa, \gamma) = \frac{(\omega/\alpha)^\gamma}{K_\gamma(\omega)} \cdot \frac{(\kappa\omega/\alpha)^l K_{\gamma+l}(\alpha)}{l!}, \quad l = 0, 1, 2, \ldots, \tag{1}$$

де $\alpha = ((\omega + \kappa)^2 - \kappa^2)^{1/2}$, $K_\nu(z)$ — модифікована функція Бесселя другого роду.

Бета-від'ємний біномний розподіл (beta-negative binomial distribution, див. [13]) з параметрами α, β, r задається наступним чином:

$$P(l; \alpha, \beta, r) = \frac{\Gamma(r + l)}{\Gamma(l + 1)\Gamma(r)} \cdot \frac{B(\alpha + r, \beta + l)}{B(\alpha, \beta)}, \quad l = 0, 1, 2, \ldots \tag{2}$$

Цей розподіл має важкий хвіст, у нього існує лише скінчена кількість моментів.

Пуассонівський логнормальний розподіл (Poisson lognormal distribution, див. [14]) з параметрами μ, σ означається формулою

$$P(l; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma l!} \int_0^\infty e^{-\lambda} \lambda^{l-1} \exp\left\{-\frac{(\ln\lambda - \mu)^2}{2\sigma^2}\right\} d\lambda, \quad l = 0, 1, 2, \ldots \tag{3}$$

Дискретний розподіл Вейбулла (див. [15]) з параметрами q, β означається так:

$$P(l; q, \beta) = q^{l^\beta} - q^{(l+1)^\beta}, \quad l = 0, 1, 2, \ldots$$

Такі добре відомі розподіли як пуассонівський та геометричний не розглядалися в якості моделей з огляду на їх недостатню гнучкість.

Для кожної сім'ї розподілів за допомогою критерію χ^2 було перевірено гіпотезу вигляду $H_0\colon F = G$, де F — розподіл частот даного слова, G — гіпотетичний розподіл (що залежить від невідомих параметрів). Невідомі параметри оцінювались за методом максимальної правдоподібності.

Також були отримані значення інформаційних критеріїв AIC та BIC. Нагадаємо, що статистики AIC та BIC означаються наступними формулами:

$$\mathrm{AIC} = -2(l - p),$$

$$\mathrm{BIC} = p\ln(N) - 2l,$$

де p — кількість параметрів, якими задається розподіл, l — значення логарифмічної функції правдоподібності, N — обсяг вибірки.

Розрахунки проводились на мові програмування `R` із використанням пакетів `DiscreteWeibull`, `fitdistrplus`, `gamlss.dist`, `sads` та `tm` (див. [16–20]).

Значення p-value критерію χ^2 наведено у табл. 1 та 2; значення статистик AIC та BIC — відповідно у табл. 3, 4 (AIC) та 5, 6 (BIC). Прочерки у деяких клітинках означають те, що відповідний розподіл підігнати не вдалося. Використані позначення BNB, DW, NB, PL відповідно для бета-від'ємного біномного розподілу, дискретного розподілу Вейбулла, від'ємного біномного розподілу та пуассонівського логнормального розподілу.

Таблиця 1.

p-value критерія χ^2

Слово	**BNB**	**DW**	**NB**	**PL**	**Sichel**
like	0.1413	$1.76 \cdot 10^{-13}$	0.0615	0.7052	0.6041
now	0.0245	$1.52 \cdot 10^{-16}$	0.0001	0.2010	0.2752
provid	0.6665	$1.37 \cdot 10^{-8}$	0.1253	0.8267	0.8839
product	0.0857	$1.60 \cdot 10^{-6}$	0.1291	0.0528	0.1457
inform	—	$9.09 \cdot 10^{-9}$	0.0412	0.0035	0.0135
number	0.0046	$6.96 \cdot 10^{-8}$	0.0156	0.0335	0.0522
offer	0.1455	$4.80 \cdot 10^{-9}$	0.0263	0.2608	0.0148
time	0.5632	$1.59 \cdot 10^{-12}$	0.0183	0.4442	0.0766
trade	0.3080	$1.30 \cdot 10^{-11}$	$3.54 \cdot 10^{-6}$	0.3956	0.8730
work	0.1711	$6.14 \cdot 10^{-10}$	0.1602	0.2266	0.0843
secur	—	$2.30 \cdot 10^{-6}$	0.0006	$3.60 \cdot 10^{-8}$	0.0003
includ	0.4074	$1.06 \cdot 10^{-9}$	0.0501	0.7641	0.5750
list	0.0444	$2.65 \cdot 10^{-12}$	0.0031	0.2059	0.2888
cash	0.0357	$2.17 \cdot 10^{-4}$	0.0409	0.2228	0.0962
opportun	—	$8.78 \cdot 10^{-7}$	0.1889	0.0729	0.0680
increas	—	$4.21 \cdot 10^{-5}$	0.0063	0.0001	0.0014
softwar	0.3699	$1.04 \cdot 10^{-3}$	0.6592	0.5677	0.4276
cost	0.3766	$1.81 \cdot 10^{-3}$	0.8146	0.3628	0.5024
world	0.0271	$3.02 \cdot 10^{-7}$	0.0195	0.0547	0.0574
approv	0.1910	$2.68 \cdot 10^{-4}$	0.3716	0.3512	0.1647
present	0.4601	$4.65 \cdot 10^{-5}$	0.7881	0.4651	0.4399
financ	0.6325	$2.54 \cdot 10^{-5}$	0.1224	0.9157	0.8742
home	0.2704	$8.57 \cdot 10^{-5}$	0.1005	0.4558	0.1277
plan	0.6598	$5.66 \cdot 10^{-4}$	0.3768	0.7614	0.4791
url	—	$9.43 \cdot 10^{-6}$	0.1488	0.1199	0.1424

Таблиця 2.

p-value критерія χ^2, продовження

Слово	**BNB**	**DW**	**NB**	**PL**	**Sichel**
credit	0.0895	$1.23 \cdot 10^{-4}$	0.0193	0.2186	0.0345
execut	0.2777	$5.91 \cdot 10^{-5}$	0.0046	0.5307	0.5989
first	0.5630	$5.79 \cdot 10^{-5}$	0.2730	0.6922	0.7425
fund	0.0135	$9.98 \cdot 10^{-6}$	0.0423	0.0558	0.0530
last	0.0365	$1.05 \cdot 10^{-7}$	0.0004	0.0704	0.0778
news	0.0110	$2.16 \cdot 10^{-5}$	0.0269	0.0033	0.0120
peopl	0.0940	$1.64 \cdot 10^{-7}$	0.0002	0.1370	0.2602
question	0.1862	$6.05 \cdot 10^{-7}$	0.0474	0.5752	0.4276
right	0.0866	$1.59 \cdot 10^{-6}$	0.0026	0.3139	0.3273
sale	0.0126	$2.04 \cdot 10^{-7}$	0.0042	0.0682	0.0877
term	0.0009	$1.48 \cdot 10^{-7}$	0.0007	0.0061	0.0022
transact	0.7511	$1.38 \cdot 10^{-2}$	0.5070	0.8782	0.6282
valu	0.0213	$1.76 \cdot 10^{-4}$	0.0209	0.4851	0.1664
move	0.0469	$2.04 \cdot 10^{-6}$	0.1273	0.8272	0.7724
buy	0.5468	$2.56 \cdot 10^{-4}$	0.0188	0.8314	0.1239
claim	—	$1.68 \cdot 10^{-4}$	0.0003	$3.74 \cdot 10^{-8}$	0.0000
high	0.0203	$3.06 \cdot 10^{-8}$	0.0154	0.1826	0.1166
save	0.0015	$2.09 \cdot 10^{-3}$	0.0228	0.0005	0.0041
billion	0.0112	$6.65 \cdot 10^{-8}$	$6.03 \cdot 10^{-6}$	0.0120	0.0777
compani	$3.43 \cdot 10^{-5}$	$4.06 \cdot 10^{-16}$	$9.76 \cdot 10^{-11}$	0.0003	0.0004
deal	0.7321	$4.40 \cdot 10^{-5}$	0.0771	0.8647	0.6723

Таблиця 3.

Значення AIC

Слово	**BNB**	**DW**	**NB**	**PL**	**Sichel**
like	1073.919	1076.704	1079.409	1071.088	1071.891
now	1034.329	1047.799	1057.33	1028.515	1025.887
provid	1024.096	1028.647	1031.029	1021.226	1022.077
product	1112.814	1111.143	1111.204	1116.548	1111.646
inform	—	1783.784	1782.552	1800.536	1784.552
number	1063.598	1062.380	1061.951	1070.928	1063.407
offer	838.844	846.588	853.525	835.881	837.663
time	1971.167	1984.697	2003.675	1970.120	1975.928
trade	923.134	945.424	963.692	920.518	917.507
work	1298.469	1299.216	1301.972	1297.612	1299.534
secur	—	1097.535	1097.276	1112.183	1099.276
includ	1015.669	1021.323	1024.785	1012.509	1013.176
list	959.341	967.788	971.867	955.248	954.776
cash	397.311	395.510	395.933	395.050	394.684
opportun	—	459.278	459.111	461.992	461.111
increas	—	402.813	401.325	411.117	403.497
softwar	490.502	488.522	488.602	490.072	490.383
cost	493.637	491.927	492.749	493.014	494.151
world	408.779	408.767	409.653	406.282	407.065
approv	436.206	434.416	434.785	434.974	436.207
present	504.251	502.283	502.396	503.353	504.370
financ	470.174	474.599	477.354	467.573	468.783

Таблиця 4.

Значення AIC, продовження

Слово	**BNB**	**DW**	**NB**	**PL**	**Sichel**
home	489.996	490.295	493.730	487.957	489.630
plan	549.016	547.823	549.321	547.455	548.178
url	—	612.554	612.641	612.907	614.166
credit	545.821	546.600	550.038	543.433	544.539
execut	414.665	418.898	423.496	412.059	412.264
first	684.814	684.033	686.041	683.076	683.351
fund	516.442	516.361	519.462	516.394	519.180
last	612.137	618.543	626.479	608.932	608.908
news	590.990	589.082	589.053	597.566	591.053
peopl	635.583	641.664	647.494	632.226	631.588
question	538.807	541.150	544.685	535.985	537.201
right	497.069	501.433	506.737	493.895	493.587
sale	538.091	539.398	540.738	535.475	535.803
term	649.873	653.966	658.369	647.858	650.215
transact	570.138	570.115	573.326	568.772	570.617
valu	482.817	482.306	483.461	480.294	480.311
move	476.753	478.836	480.744	474.167	475.512
buy	433.025	435.181	438.805	430.419	431.285
claim	—	467.600	466.533	478.757	468.533
high	505.384	509.234	511.189	502.165	502.918
save	460.922	459.095	460.481	461.264	462.213
billion	341.529	349.568	354.570	339.277	338.005
compani	2066.078	2087.486	2116.992	2060.633	2060.711
deal	1018.781	1025.508	1037.541	1016.756	1019.938

Таблиця 5.

Значення BIC

Слово	**BNB**	**DW**	**NB**	**PL**	**Sichel**
like	1087.520	1085.772	1088.476	1080.156	1085.492
now	1047.722	1056.729	1066.259	1037.445	1039.281
provid	1035.948	1036.548	1038.930	1029.128	1033.929
product	1124.673	1119.050	1119.110	1124.454	1123.506
inform	—	1792.688	1791.456	1809.440	1797.908
number	1075.697	1070.447	1070.018	1078.994	1075.506
offer	850.907	854.630	861.567	843.923	849.726
time	1985.493	1994.248	2013.226	1979.671	1990.254
trade	934.049	952.701	970.969	927.795	928.422
work	1311.714	1308.046	1310.802	1306.442	1312.779
secur	—	1104.956	1104.696	1119.604	1110.407
includ	1027.783	1029.399	1032.861	1020.585	1025.290
list	971.338	975.786	979.865	963.246	966.772
cash	406.665	401.746	402.169	401.286	404.038
opportun	—	466.119	465.952	468.833	471.373
increas	—	409.535	408.047	417.840	413.581
softwar	500.047	494.886	494.966	496.435	499.929
world	419.067	415.626	416.512	413.141	417.353
approv	446.086	441.003	441.371	441.560	446.087
present	514.526	509.133	509.246	510.203	514.644
financ	479.265	480.660	483.414	473.634	477.874

Таблиця 6.

Значення BIC, продовження

Слово	**BNB**	**DW**	**NB**	**PL**	**Sichel**
cost	503.547	498.533	499.355	499.621	504.061
home	500.802	497.500	500.934	495.161	500.437
plan	559.640	554.905	556.404	554.538	558.801
url	—	618.984	619.071	619.336	623.811
credit	555.466	553.030	556.468	549.863	554.184
execut	423.490	424.781	429.380	417.942	421.089
first	696.388	691.749	693.757	690.792	694.925
fund	525.630	522.486	525.587	522.520	528.368
last	623.685	626.242	634.177	616.631	620.456
news	600.252	595.257	595.228	603.742	600.315
peopl	646.704	649.078	654.908	639.640	642.710
question	550.620	549.025	552.560	543.860	549.013
right	508.069	508.766	514.070	501.228	504.586
sale	547.955	545.974	547.315	542.052	545.668
term	659.872	660.631	665.034	654.523	660.213
transact	579.492	576.351	579.562	575.008	579.971
valu	492.445	488.725	489.880	486.713	489.939
move	486.618	485.413	487.321	480.744	485.376
buy	442.905	441.767	445.392	437.005	441.165
claim	—	473.320	472.252	484.476	477.112
high	515.523	515.994	517.949	508.925	513.057
save	471.210	465.953	467.340	468.122	472.501
billion	348.481	354.203	359.205	343.912	344.958
compani	2078.426	2095.718	2125.224	2068.864	2073.059
deal	1029.962	1032.961	1044.995	1024.209	1031.119

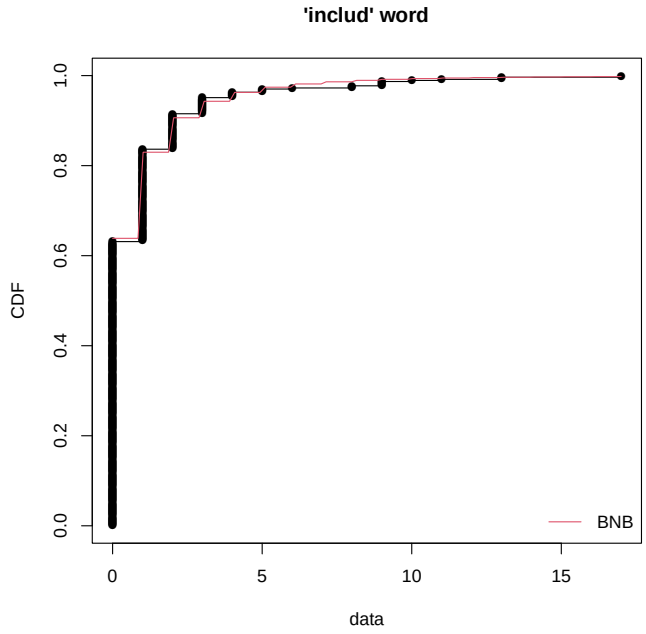

Рис. 1. Слово “includ”, бета-від’ємний біномний розподіл.

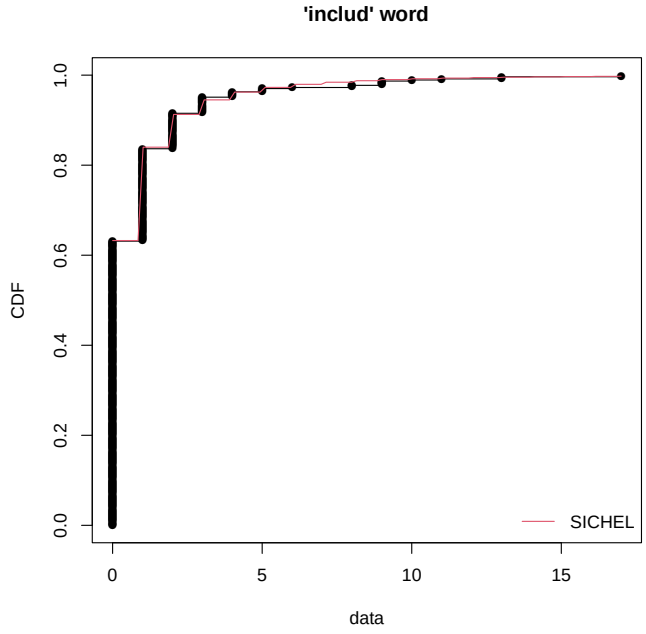

Рис. 2. Слово “includ”, розподіл Зіхеля.

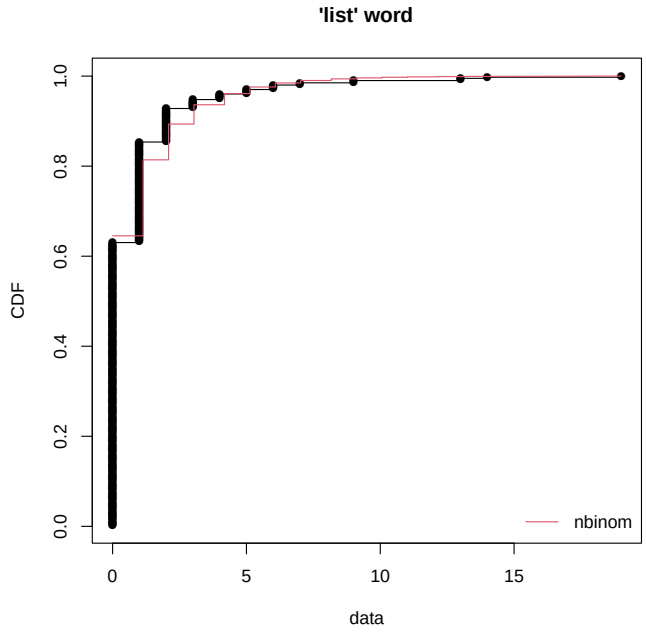

Рис. 3. Слово “list”, від’ємний біномний розподіл.

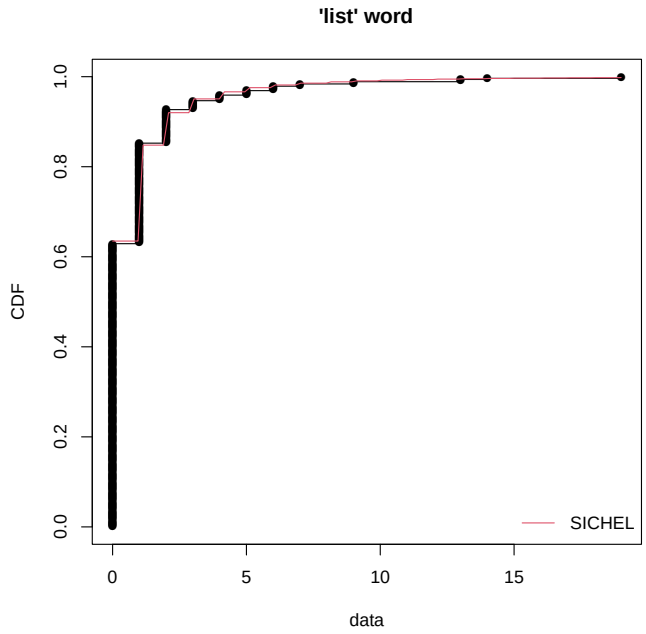

Рис. 4. Слово “list”, розподіл Зіхеля.

Графіки емпіричних функцій розподілу та функцій розподілу деяких з теоретичних розподілів для слів “includ” та “list” наведено на рис. 1, 2, 3 та 4.

Проаналізуємо отримані результати.

Дискретний розподіл Вейбулла є непридатним для моделювання частот слів — цей розподіл не є адекватним за критерієм χ^2 (p-value менше 0.05) у всіх 46 випадках.

Решта розподілів (за винятком від’ємного біномного) порівняно непогано підганяються до даних: p-value критерію χ^2 більше 0.10 для 50%, 57% і 65% з наших 46 слів відповідно для бета-від’ємного біномного розподілу, розподілу Зіхеля та пуассонівського логнормального розподілу.

Від’ємний біномний розподіл пристосований для моделювання частот слів гірше — p-value критерію χ^2 більше 0.10 лише для 33% слів. Варто зазначити, що, незважаючи на порівняно погану якість підгонки цього розподілу “у цілому”, від’ємний біномний розподіл часто добре підганяється до вибірок з порівняно коротким “емпіричним хвостом” (де $\max\{\zeta\} \leq 25$).

Що стосується критеріїв AIC та BIC, оптимальним розподілом найчастіше є пуассонівський логнормальний розподіл — для нього значення AIC та BIC є найменшими відповідно для 23 слів та 31 слова.

3. Висновки. Знайдено ймовірнісні розподіли, які є оптимальними для моделювання частот широко вживаних слів у великій колекції коротких текстів. Отримані результати можуть бути використані, зокрема, для класифікації документів та побудови регресійних моделей, де залежною змінною є частота слова.

Розподіл роботи співавторів. Є. В. Турчин: постановка задачі, методологія, написання статті, частково — обчислення. Ю. С. Федорченко: частково — обчислення.

Список використаної літератури

1. Tagg C. A corpus linguistics study of SMS text messaging. Ph. D. thesis : Birmingham, 2009. 402 p. URL: https://etheses.bham.ac.uk/id/eprint/253/ (date of access: 08.06.2024).
2. Ni X. et al. Short text clustering by finding core terms. *Knowledge and Information Systems.* 2011. Vol. 27, No. 3. P. 345–365. DOI: https://doi.org/10.1007/s10115-010-0299-7
3. Rafeeque P. C., Sendhilkumar S. A survey on short text analysis in Web. Proceedings of *2011 Third International Conference on Advanced Computing* : Chennai. India, 2011. P. 365–371. URL: https://ieeexplore.ieee.org/abstract/document/6165203/ (date of access: 08.06.2024).
4. Brocardo M. L. et al. Authorship verification for short messages using stylometry. Proceedings of *2013 International Conference on Computer, Information and Telecommunication Systems (CITS)* : Athens. Greece, 2013. P. 1–6. DOI: https://doi.org/10.1109/CITS.2013.6705711
5. Lyddy F. et al. An analysis of language in university students' text messages. *Journal of Computer-Mediated Communication.* 2014. Vol. 19, No. 3. P. 546–561. URL: https://academic.oup.com/jcmc/article-abstract/19/3/546/4067601 (date of access: 08.06.2024).
6. Xu J. et al. Self-taught convolutional neural networks for short text clustering. *Neural Networks.* 2017. Vol. 88. P. 22–31. DOI: https://doi.org/10.1016/j.neunet.2016.12.008
7. Zheng C. T., Liu C., Wong H. S. Corpus-based topic diffusion for short text clustering. *Neurocomputing.* 2018. Vol. 275. P. 2444–2458. DOI: https://doi.org/10.1016/j.neucom.2017.11.019
8. Sjarif A. N. N. et al. SMS spam message detection using term frequency-inverse document frequency and random forest algorithm. *Procedia Computer Science.* 2019. Vol. 161. P. 509–515. URL: https://www.sciencedirect.com/science/article/pii/S1877050919318617 (date of access: 08.06.2024).
9. Srinivasan L., Nalini C. An improved framework for authorship identification in online messages. *Cluster Computing.* 2019. Vol. 22. P. 12101–12110. DOI: https://doi.org/10.1007/s10586-017-1563-3
10. Albalawi R., Yeap T. H., Benyoucef M. Using topic modeling methods for short-text data: a comparative analysis. *Frontiers in Artificial Intelligence.* 2020. Vol. 3. P. 42. URL: https://www.frontiersin.org/articles/10.3389/frai.2020.00042/full (date of access: 08.06.2024).
11. Qiang J. et al. Short text topic modeling techniques, applications, and performance: a survey. *IEEE Transactions on Knowledge and Data Engineering.* 2022. Vol. 34, No. 3. P. 1427–1445. URL: https://ieeexplore.ieee.org/abstract/document/9086136/ (date of access: 08.06.2024).
12. mshenoda/spam-messages · Datasets at Hugging Face
URL: https://huggingface.co/datasets/mshenoda/spam-messages (date of access: 08.06.2024).
13. Johnson N. L., Kemp A. W., Kotz S. Univariate Discrete Distributions. Hoboken, N.J. : Wiley, 2005. 646 p.
14. Bulmer M. G. On fitting the poisson lognormal distribution to species-abundance data. *Biometrics.* 1974. Vol. 30, No. 1. P. 101–110. DOI: https://doi.org/10.2307/2529621
15. Nakagawa T., Osaki S. The discrete Weibull distribution. *IEEE Transactions on Reliability.* 1975. Vol. R-24, No. 5. P. 300–301. DOI: https://doi.org/10.1109/TR.1975.5214915
16. Delignette-Muller M. L., Dutang C. fitdistrplus: An R package for fitting distributions. *Journal of Statistical Software.* 2015. Vol. 64, No. 4. P. 1–34. DOI: https://doi.org/10.18637/jss.v064.i04
17. DiscreteWeibull: Discrete Weibull Distributions (Type 1 and 3).
URL: https://cran.r-project.org/package=DiscreteWeibull (date of access: 08.06.2024).
18. Feinerer I., Hornik K. (2024). tm: Text Mining Package. R package version 0.7-13. URL: https://CRAN.R-project.org/package=tm (date of access: 08.06.2024).
19. gamlss.dist: Distributions for Generalized Additive Models for Location Scale and Shape. URL: https://cran.r-project.org/package=gamlss.dist (date of access: 08.06.2024).
20. Prado P., Dantas Miranda M., Chalom A. sads: Maximum Likelihood Models for Species Abundance Distributions. URL: https://CRAN.R-project.org/package=sads (date of access: 08.06.2024).

Turchyn I. V., Fedorchenko Yu. S. Distribution of frequent words in short text messages.

We consider a problem of word frequency distribution in a text corpus which consists of short messages (the emphasis is put on frequent words). The most adequate distributions were found among several distribution families (the chi-square test was used, the distributions were compared using the AIC and BIC statistics).

Keywords: word frequency distribution, mathematical linguistics, chi-square test, AIC criterion, BIC criterion.

References

1. Tagg, C. (2009). A corpus linguistics study of SMS text messaging [PhD Thesis, University of Birmingham]. Retrieved from https://etheses.bham.ac.uk/id/eprint/253/
2. Ni, X., Quan, X., Lu, Z., Wenyin, L., & Hua, B. (2011). Short text clustering by finding core terms. *Knowledge and Information Systems,* 27(3), 345–365. https://doi.org/10.1007/s10115-010-0299-7
3. Rafeeque, P. C., & Sendhilkumar, S. (2011). A survey on short text analysis in web. *2011 Third International Conference on Advanced Computing.* Retrieved from https://ieeexplore.ieee.org/abstract/document/6165203/
4. Brocardo, M. L., Traore, I., Saad, S., & Woungang, I. (2013). Authorship verification for short messages using stylometry. *2013 International Conference on Computer, Information and Telecommunication Systems (CITS).* https://doi.org/10.1109/CITS.2013.6705711
5. Lyddy, F., Farina, F., Hanney, J., Farrell, L., & O'Neill, N. K. (2014). An analysis of language in university students' text messages. *Journal of Computer-Mediated Communication,* 19(3), 546–561. Retrieved from https://academic.oup.com/jcmc/article-abstract/19/3/546/4067601
6. Xu, J., Xu, B., Wang, P., Zheng, S., Tian, G., Zhao, J., & Xu, B. (2017). Self-taught convolutional neural networks for short text clustering. *Neural Networks,* 88, 22–31. https://doi.org/10.1016/j.neunet.2016.12.008
7. Zheng, C. T., Liu, C., & Wong, H. S. (2018). Corpus-based topic diffusion for short text clustering. *Neurocomputing,* 275, 2444–2458. https://doi.org/10.1016/j.neucom.2017.11.019
8. Sjarif, N. N. A., Azmi, N. F. M., Chuprat, S., Sarkan, H. M., Yahya, Y., & Sam, S. M. (2019). SMS spam message detection using term frequency-inverse document frequency and random forest algorithm. *Procedia Computer Science,* 161, 509–515. Retrieved from https://www.sciencedirect.com/science/article/pii/S1877050919318617
9. Srinivasan, L., & Nalini, C. (2019). An improved framework for authorship identification in online messages. *Cluster Computing,* 22(S5), 12101–12110. https://doi.org/10.1007/s10586-017-1563-3
10. Albalawi, R., Yeap, T. H., & Benyoucef, M. (2020). Using topic modeling methods for short-text data: A comparative analysis. *Frontiers in Artificial Intelligence,* 3, 42. Retrieved from https://www.frontiersin.org/articles/10.3389/frai.2020.00042/full
11. Qiang, J., Qian, Z., Li, Y., Yuan, Y., & Wu, X. (2020). Short text topic modeling techniques, applications, and performance: a survey. *IEEE Transactions on Knowledge and Data Engineering,* 34(3), 1427–1445. Retrieved from https://ieeexplore.ieee.org/abstract/document/9086136/
12. mshenoda/spam-messages · Datasets at Hugging Face. Retrieved from https://huggingface.co/datasets/mshenoda/spam-messages
13. Johnson, N. L., Kemp, A. W., & Kotz, S. (2005). *Univariate Discrete Distributions.* Hoboken, N.J.: Wiley.
14. Bulmer, M. G. (1974). On fitting the Poisson lognormal distribution to species-abundance data. *Biometrics,* 30(1), 101. https://doi.org/10.2307/2529621
15. Nakagawa, T., & Osaki, S. (1975). The discrete weibull distribution. *IEEE Transactions on Reliability,* R-24(5), 300–301. https://doi.org/10.1109/TR.1975.5214915
16. Delignette-Muller, M. L., & Dutang, C. (2015). fitdistrplus: An R package for fitting distributions. *Journal of Statistical Software,* 64(4), 1–34. https://doi.org/10.18637/jss.v064.i04
17. DiscreteWeibull: Discrete Weibull Distributions (Type 1 and 3). Retrieved from https://cran.r-project.org/web/packages/DiscreteWeibull/index.html

18. Feinerer I., & Hornik K. (2024). tm: Text Mining Package. R package version 0.7-13. Retrieved from https://CRAN.R-project.org/package=tm
19. gamlss.dist: Distributions for Generalized Additive Models for Location Scale and Shape. Retrieved from https://cran.r-project.org/package=gamlss.dist
20. Prado P., Dantas Miranda M., & Chalom A. sads: Maximum Likelihood Models for Species Abundance Distributions. Retrieved from https://CRAN.R-project.org/package=sads

Одержано 11.07.2024