Spot the bot: семантические траектории текстов естественного языка

Исследовательский проект

Данг Куинь Ньы, 4 курс, БПМИ182
Руководитель: Громов Василий Александрович

Москва, 2022

Постановка задачи

Задача: исследовать семантическое пространство текстов, написанных людьми и сгенерированных ботами

Цель: построение алгоритма различения текстов

Гипотезы:

- более компактная структура текстов ботов
- человеческие тексты более хаотичные

Методы:

- кластеризация
- плоскость энтропии-сложности

Существующие методы

- С использованием метаданных:
  - Сетевой анализ (Chakraborty et al., Dickerson et al.)
  - Теория информации (Chu et al.)
- С использованием текстовых данных:
  - Семантические: простые лексические и синтаксические признаки (частота букв, средняя длина слова и т.д.) (Kang et al.)
  - Тональность, стиль текста (Heidari et al., Cardaioli)

- В большинстве исследований используются алгоритмы обучения с учителем и/или исследуются конкретные боты

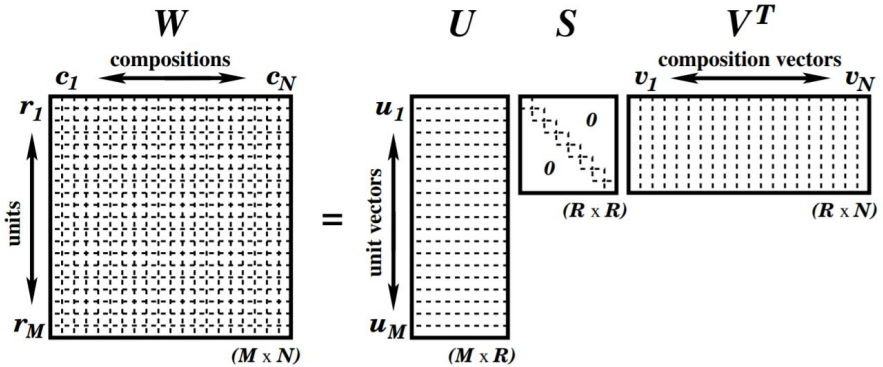
Методология

Методология

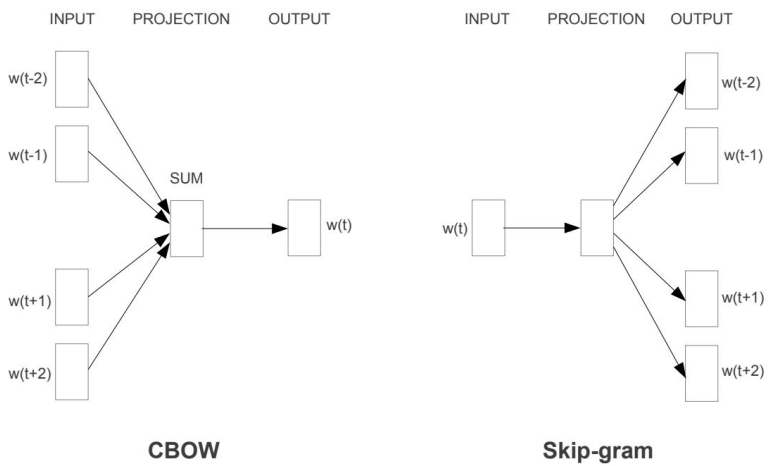
- Векторные представления:
  - SVD
  - Word2Vec (Skip-gram, CBOW)
- Кластеризация:
  - Алгоритм Уишарта
  - K-Means
- Нечеткая логика
  - Нечеткие числа
  - Нечеткая кластеризация C-Means
- Плоскость энтропии-сложности

Методология. Векторные представления

SVD



Word2Vec



Семантическое пространство: векторные представления *n*-грамм

1. Источник изображения: Bellegarda J. R.. Latent semantic mapping: principles and applications. //Synthesis Lectures on Speech and Audio Processing. Vol. 3, No. 1, pp. 1–101, 2007.
2. Источник изображения: Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781* (2013).

Методология. Кластеризация

K-Means

+ Хорошо выделяет данные со сферическими кластерами
+ Быстрый
- Чувствителен к выбору параметру числа кластеров
- Не выделяет шум/выбросы

Алгоритм Уишарта

+ Выделяет кластера произвольной формы
+ Выделяет шумовые объекты
- Плохо выделяет кластера разных плотностей
- Долго работает на данных с большой размерностью

K-Means

Разбиение на кластеры:

$$C_i = \{x : ||x - c_i||^2 \leqslant \min_{j=1,\dots,K} ||x - c_j||^2\}$$

Пересчёт центроидов:

$$c_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$$

Algorithm 1: Алгоритм Уишарта

Input: $\{x_1 \dots x_\ell\}$ — объекты, $d(x_i, x_j)$ — функция расстояния, k, h
Output: $\{y_i = y(x_i)\}_{i-1}^{\ell}$ — номера кластеров для объектов

$d_k(x_i) \leftarrow$ расстояние до k-го ближайшего соседа x_i
отсортировать объекты $d_k(x_{(1)}) \leqslant \dots \leqslant d_k(x_{(\ell)})$
for $i \leftarrow 1$ **to** ℓ **do**
  $V_i \leftarrow \{x \in \{x_{(1)} \dots x_{(i-1)}\} \mid d(x_{(i)}, x) \leqslant d_k(x_{(i)})\}$
  $C_i \leftarrow \{y(x) \mid x \in V_i\}$
  if $C_i = \varnothing$ **then**
    сгенерировать новый кластер c
    $y_{(i)} \leftarrow c$
  if $|C_i| = 1$ **then**
    пусть $C_i = \{c\}$
    if *completed(c)* **then**
      $y_{(i)} \leftarrow$ noise
    else
      $y_{(i)} \leftarrow c$
  if $|C_i| > 1$ **then**
    пусть $C_i = \{c_1, \dots, c_t\}$
    if *completed*(c_j) $\forall j$ **then**
      $y_{(i)} \leftarrow$ noise
    $S_i \leftarrow \{c \in C_i \mid c$ is significant$(h)\}$
    if $|S_i| > 1$ **then**
      completed$(c) \leftarrow True$ $\forall c \in S_i$
      $y(x) \leftarrow$ noise $\forall x \in c \in C_i \setminus S_i$
      $y_{(i)} \leftarrow$ noise
    else
      объединить все кластеры из C_i в один и обновить метки у входящих в них объектов
return y

Методология. Нечеткая логика

C-Means — нечеткая вариация алгоритма *K-Means*

$w_k(x)$ — коэффициент принадлежности объекта *x* кластеру *k*

$$c_k = \frac{\sum_x w_k^m(x)x}{\sum_x w_k^m(x)}$$

1) Пересчет центроидов

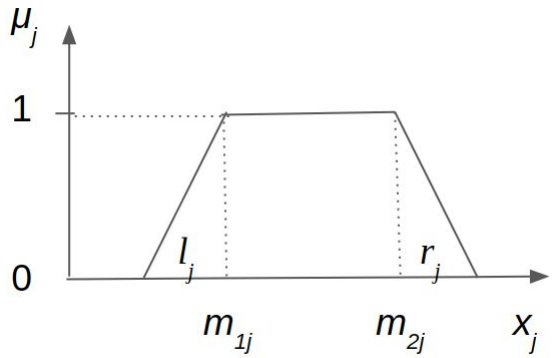
$$w_k(x) = \left(\sum_{c \in C} \left(\frac{||x - c_k||}{||x - c||} \right)^{\frac{2}{m-1}} \right)^{-1}$$

2) Пересчет коэффициентов

Методология. Нечеткая логика

Идея: нечеткое представление данных учитывает их неточность + моделирует качественные понятия

$$\mu_j(x_j) = \begin{cases} L\left(\frac{m_{1j}-x_j}{l_j}\right), & x_j \leqslant m_{1j} \\ 1, & m_{1j} \leqslant x_j \leqslant m_{2j}, \\ R\left(\frac{x_j-m_{2j}}{r_j}\right), & x_j \geqslant m_{2j} \end{cases} \quad j = 1, \ldots m$$



Методология. Нечеткая логика

$$\mu_j(x_j) = \frac{n_j}{\max\limits_j n_j}, \quad j = 1, \ldots, m$$

n_j — число вхождений *j*-ой компоненты в текст

$$\mu((x, y)) = \{\min(\mu_j(x), \mu_j(y))\}_{j=1}^{m}$$

— функция принадлежности биграммы *(x,y)*

Расстояние между нечеткими числами:

$$d(\tilde{x}_i, \tilde{x}_j) = \big(\|m_{1i} - m_{1j}\|^2 + \|m_{2i} - m_{2j}\|^2 + \|(m_{1i} - \lambda\ell_i) - (m_{1j} - \lambda\ell_j)\|^2 + \|(m_{2i} + \rho r_i) - (m_{2j} + \rho r_j)\|^2 \big)^{1/2}$$

Методология. Энтропия и сложность

Идея: тексты можно рассмотреть как временные ряды, которые разделяются на хаотические, простые детерминированные и стохастические

$$S[P] = -\sum_{j=1}^{N} p_j \ln p_j$$

$$H_S[P] = S[P]/S_{\max}$$

Энтропия

$$C_{JS}[P] = Q_J[P, P_e]H_s[P]$$

Q_J — нормированная дивергенция Дженсона-Шеннона между распределениями P, P_e = *(1/N, …, 1/N)*

Сложность

Методология. Энтропия и сложность

Ordinal pattern — перестановка индексов (0, …, *n*-1)

$$\pi = (r_0, r_1, \ldots, r_{n-1}) \qquad x_{t-r_{n-1}} \leqslant x_{t-r_{n-2}} \leqslant \cdots \leqslant x_{t-r_1} \leqslant x_{t-r_0}$$

В многомерном случае:

- Для каждой компоненты *d* получаем перестановку π_d
- Общая перестановка — $\Pi = (\pi_1, \pi_2, \ldots, \pi_m)$ (всего $(n!)^m$ вариантов)

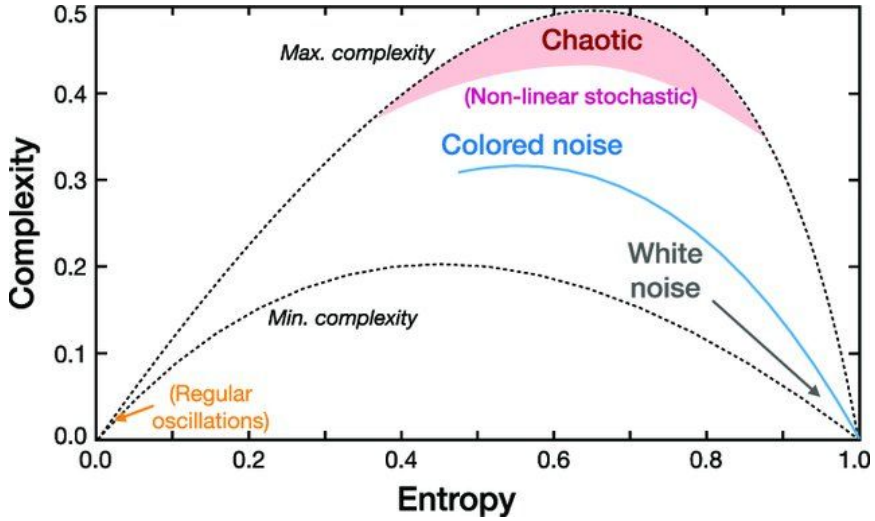
Методология. Энтропия и сложность

Рассматриваемое распределение *P* — распределение порядковых паттернов *π*

Определим распределение как частоту встречаемости паттернов.

$$p(\pi) = \frac{\{t | t \leqslant L - n + 1 : (t) \text{ соответствует перестановка } \pi\}}{L - n + 1}$$

Методология. Энтропия и сложность



- Правый нижний край: *шумовые процессы*
- Левый нижний край: *детерминированные процессы*
- Прилежащая к середине верхней границе область: *хаотические процессы*

1. Источник изображения: M. Zanin and F. Olivares, “Ordinal patterns-based methodologies for distinguishing chaos from noise in discrete time series”, Communications Physics 4, vol. 1, no. 190, August 2021. DOI: 10.1038/s42005-021-00696-z

Результаты

План работы

1. Сбор, генерация и обработка данных
2. Получение векторных представлений
3. Исследование семантического пространства
4. Построение классификаторов на основе полученных результатов анализа

Сбор и обработка данных

Данные

Корпуса художественной литературы собраны из проекта Гутенберга и других открытых источников.

Язык	**Размер корпуса**	**Средний размер текста**
Вьетнамский	1071	54532
Русский	12692	1000
Английский	11008	21000

Данные

Long short-term memory recurrent neural network (LSTM)

theo thì chiều thấy chuyện trong thuyền đến cho chỉ thì thấy thể cho chúng cho trong chàng chiều đi

он признавал от нем высоко совсем конечно в небесной в странное вообще под вам все старика, страстных сердце с стеной получила в получила в сердце на столование

and the body of the matter with the man in the way to the particularly wanting the state of the probably to the street with a word and with the beautiful ends and

Generative Pretrained Transformers (GPT-2/3)

Trong tiếng Anh còn có các nghĩa: "Thực tế, vũ trụ phát triển vượt bậc trong thời đại ngày nay đồng nghĩa với việc nó sẽ tạo ra giá trị to lớn cho sự sống trên Trái Đất".

послушайте, а что это за штука такая? — Это? — переспросил я плевательское отношение к людям, к их проблемам, к их нуждам

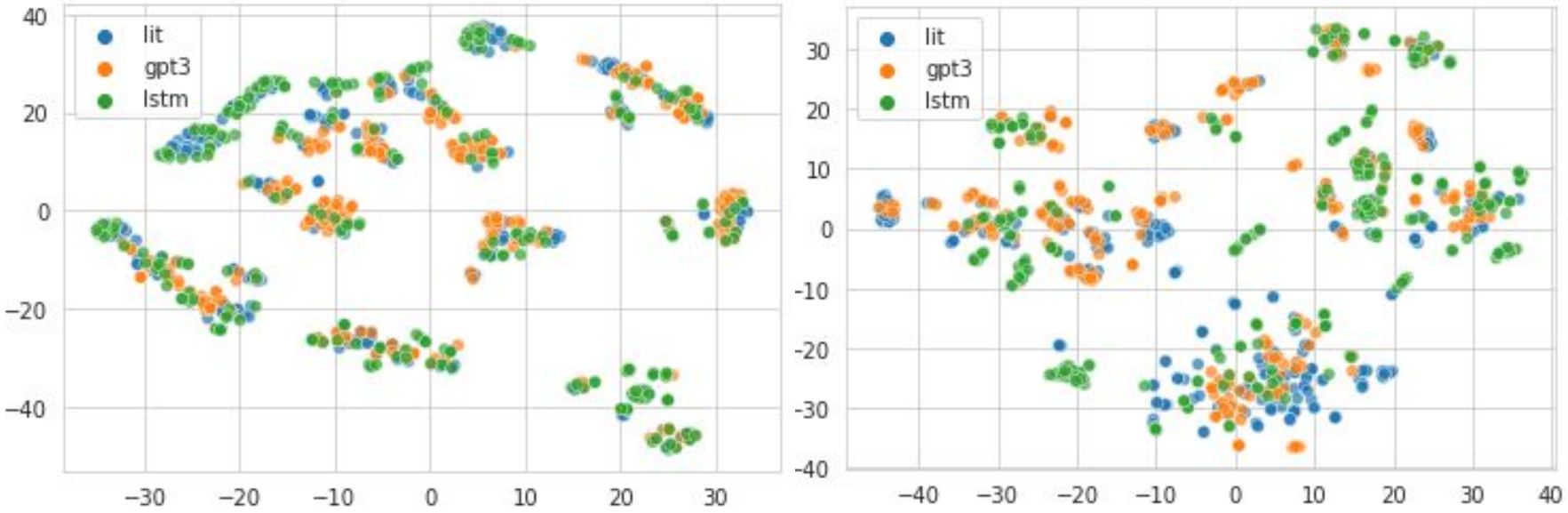
As I became more intelligent at that time it took me much longer to get myself into what I actually believed in.

Предобработка

- Токенизация
- Лемматизация
- Замена числительных, имён собственных, названий, местоимений специальными токенами

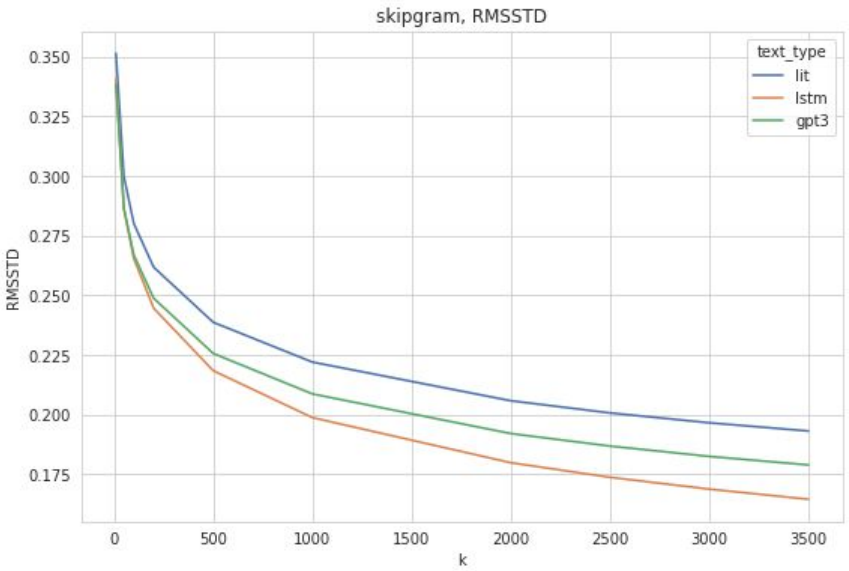
Кластеризация

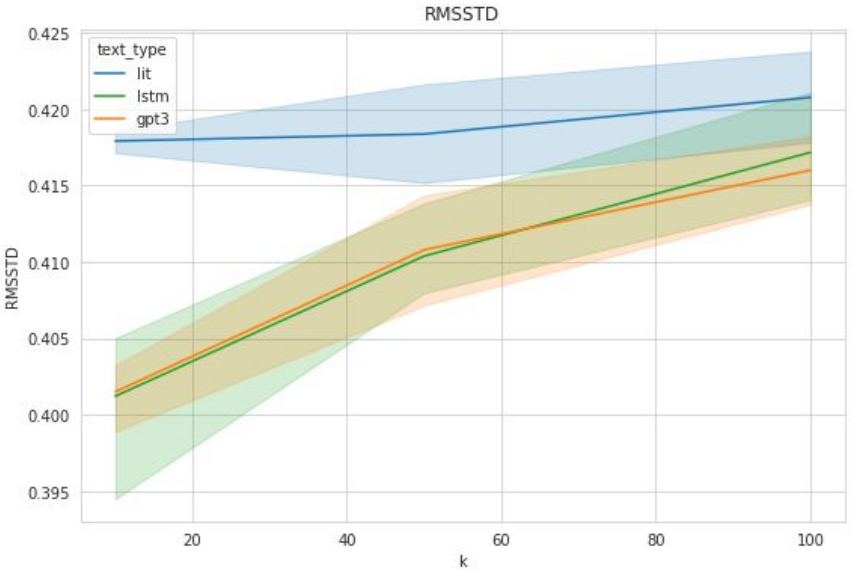
Кластеризация



TSNE-визуализация текстов разных типов для SVD представлений (слева) и Skip-gram (справа)

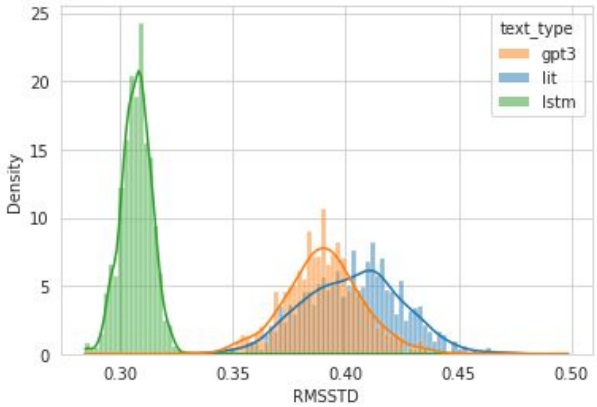
Кластеризация

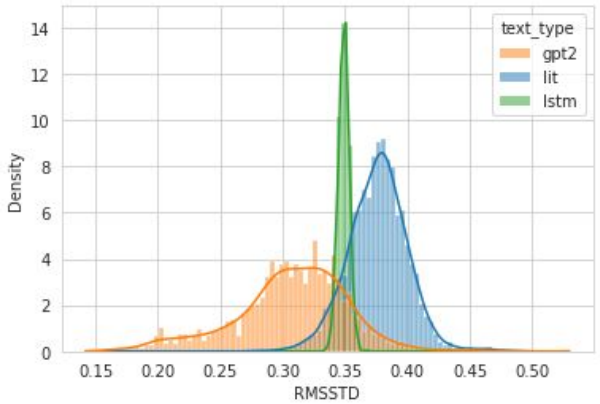


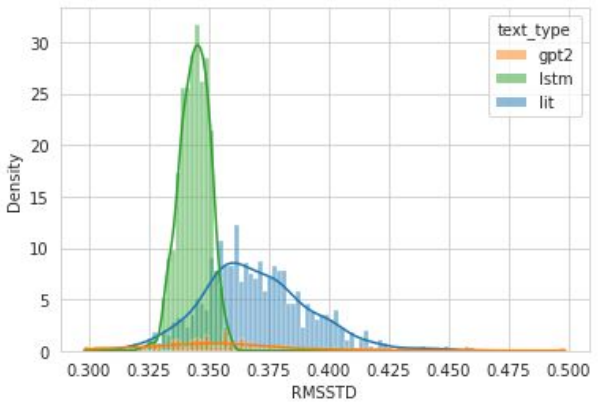


Компактность кластеров (RMSSTD) на всем корпусе вьетнамских текстов для K-Means (слева) и алгоритма Уишарта (справа) на Skip-gram векторах

Кластеризация

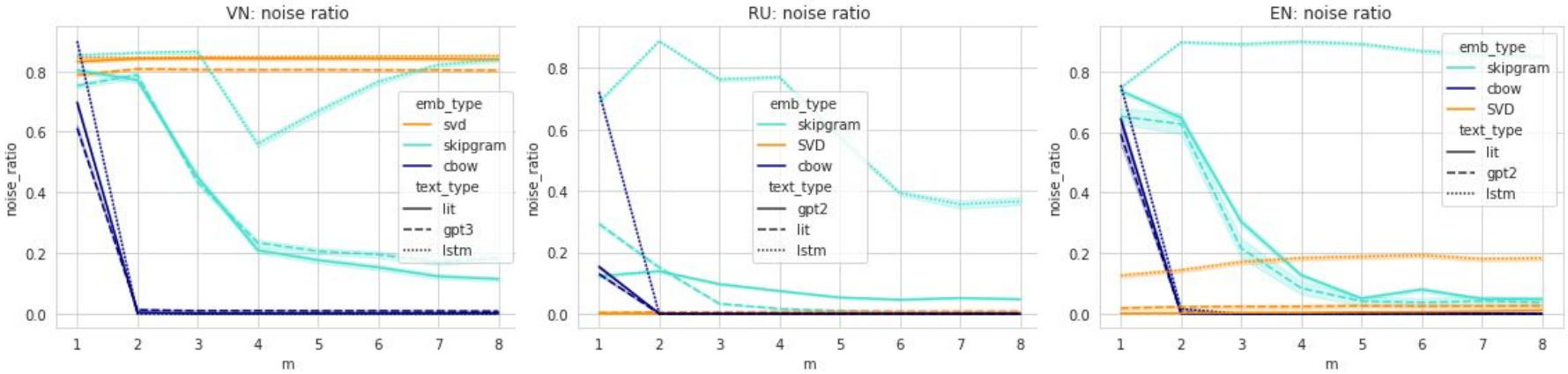






Распределение RMSSTD на корпусах вьетнамского (слева), русского (посередине) и английского (справа) языков для K-Means на Skip-gram векторах

Кластеризация



Доля шума, выделяемого кластеризацией Уишарта на корпусах вьетнамского (слева), русского (посередине) и английского (справа) языков

Кластеризация

- Тексты ботов получаются более простыми (кластера компактнее)
- Есть статистически значимое отличие распределений RMSSTD между литературными и сгенерированными ботами текстов (оценено критерием Уилкоксона при уровне значимости 5%, p-value < 0.03)
- В структуре есть явные отличия, будем использовать внутрикластерные расстояния в качестве признаков для построения классификаторов

Кластеризация

	все боты		**LSTM**		**GPT**	
Вид кластеризации	Train	Test	Train	Test	Train	Test
K-Means	0.862	0.903	1.0	1.0	0.887	0.881
Уишарт	0.902	0.896	1.0	1.0	0.893	0.900
C-Means	0.887	0.893	1.0	1.0	0.871	0.871
Уишарт на нечетких числах	**0.929**	**0.942**	1.0	1.0	**0.893**	**0.926**

Таблица 6.1: Точность классификаторов для вьетнамского языка

Кластеризация

	все боты		**LSTM**		**GPT**	
Вид кластеризации	Train	Test	Train	Test	Train	Test
K-Means	0.912	0.934	0.999	1.0	0.871	0.916
Уишарт	**0.937**	**0.954**	0.999	1.0	**0.913**	**0.944**
C-Means	0.882	0.894	0.999	1.0	0.838	0.857
Уишарт на нечетких числах	0.882	0.913	0.991	1.0	0.904	0.911

Таблица 6.2: Точность классификаторов для русского языка

Кластеризация

	все боты		**LSTM**		**GPT**	
Вид кластеризации	Train	Test	Train	Test	Train	Test
K-Means	0.947	0.975	1.0	1.0	0.903	0.881
Уишарт	**0.953**	**0.975**	1.0	1.0	0.904	0.881
C-Means	0.943	0.970	0.999	1.0	0.897	0.921
Уишарт на нечетких числах	0.945	0.947	1.0	1.0	**0.907**	**0.94**

Таблица 6.3: Точность классификаторов для английского языка

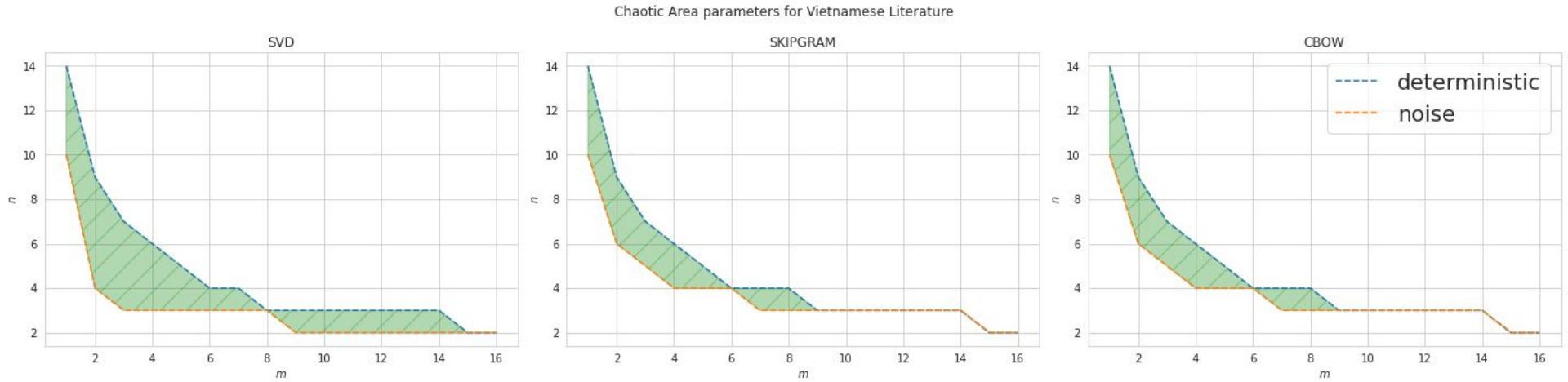
Кластеризация

- Кластера сгенерированных текстов получаются более компактными, алгоритмом Уишарта выделяется больше шума
- Выявлено статистически значимое различие между распределениями RMSSTD ботов и людей
- Классификация с использованием внутрикластерных расстояний (при кластеризации Уишарта) в качестве признаков достигает точности:
  - 94.2% — на вьетнамском языке
  - 95.4% — на русском языке
  - 97.5% — на английском языке
- Нечеткое представление данных[1] улучшает точность обнаружения сложных ботов на вьетнамском (с 90% до 92.6%) и английском (с 88.1% до 94%) языках

1. https://github.com/quynhu-d/Spot-the-Bot/tree/main/clustering

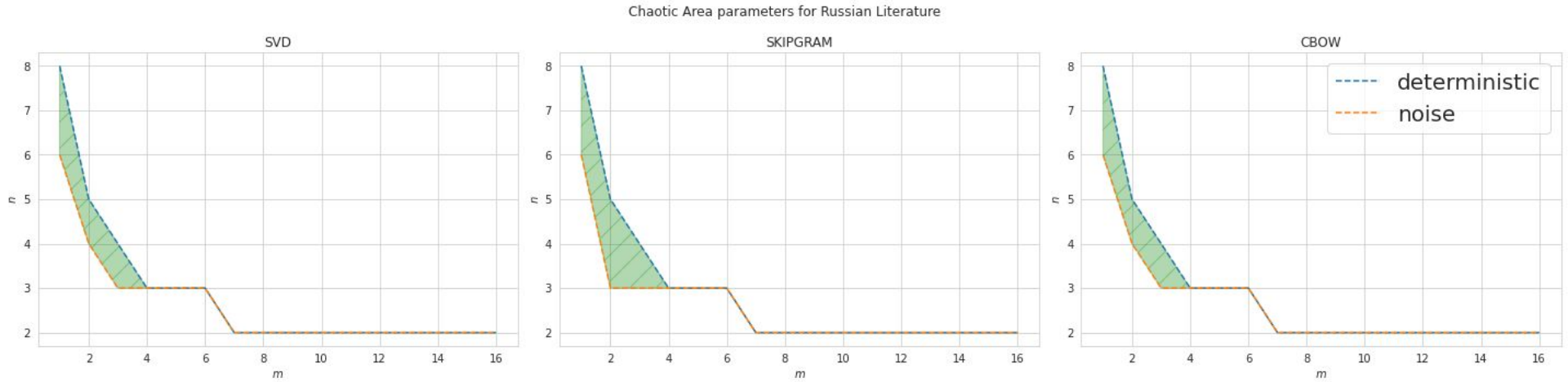
Энтропия и сложность

Энтропия и сложность



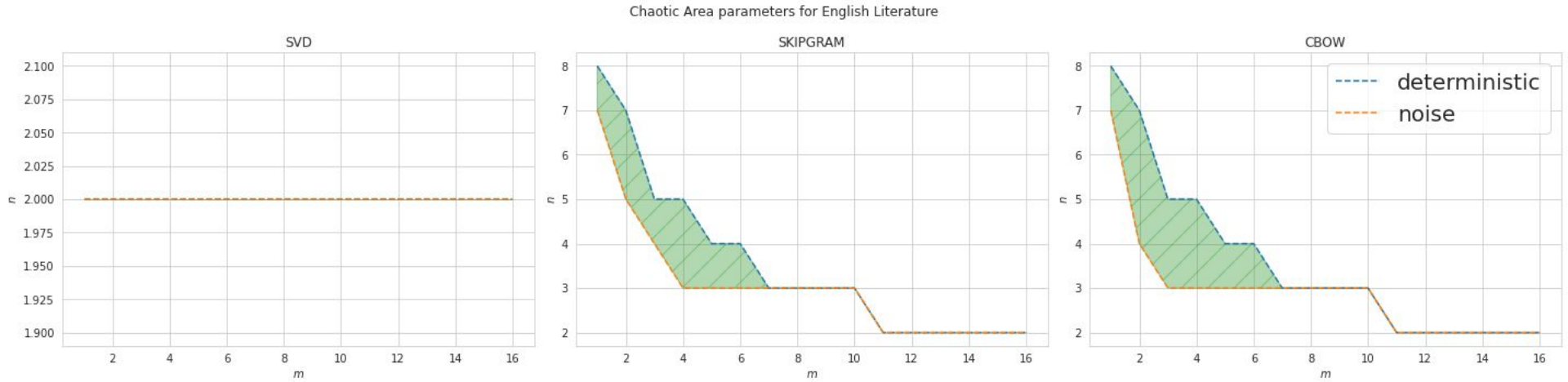
Значения *m* и *n,* при которых литературные тексты вьетнамского языка попадают в область хаотических процессов

Энтропия и сложность



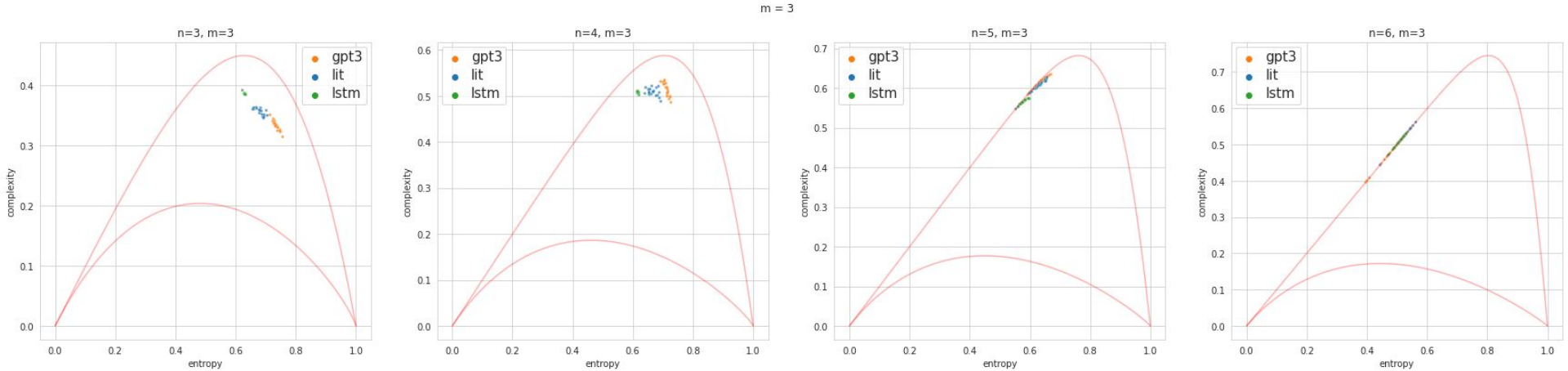
Значения *m* и *n,* при которых литературные тексты русского языка попадают в область хаотических процессов

Энтропия и сложность



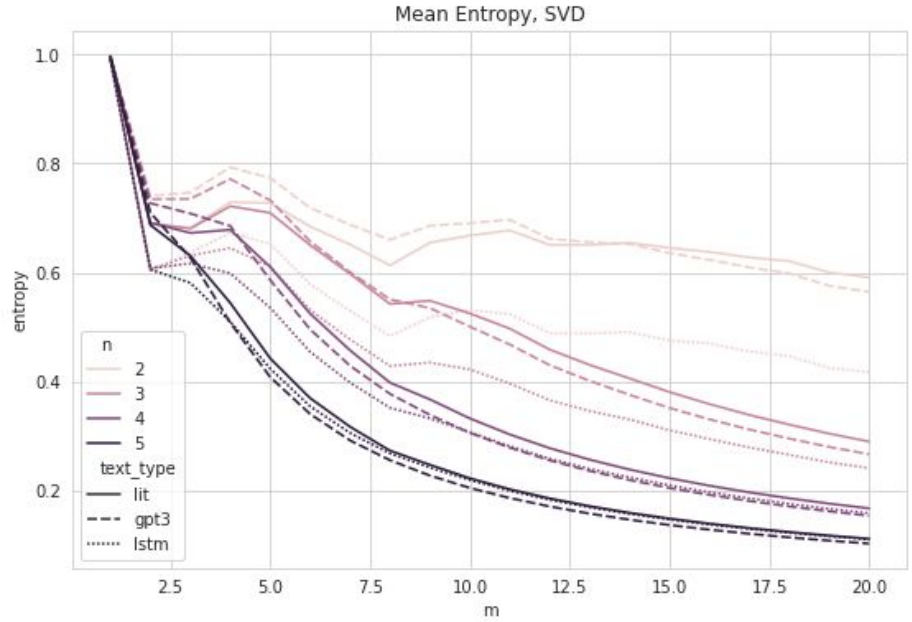
Значения *m* и *n,* при которых литературные тексты английского языка попадают в область хаотических процессов

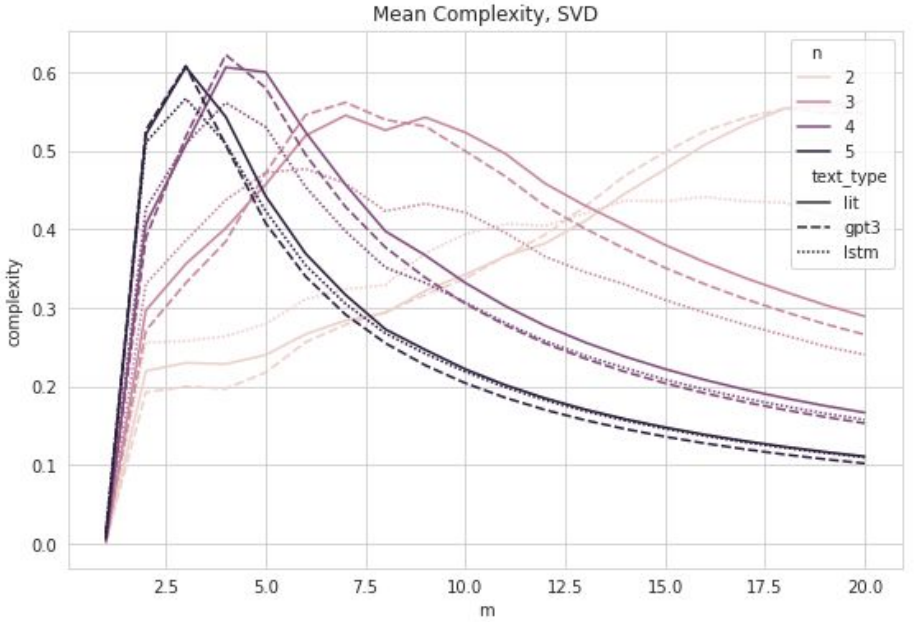
Энтропия и сложность



Плоскости энтропии-сложности для вьетнамских текстов (с SVD-векторами). $m = 3$, $n = 3, \ldots 6$

Энтропия и сложность





Зависимость энтропии (слева) и сложности (справа) от значения *m* при *n* = 2, . . . 5 на SVD-векторах для вьетнамских текстов.

Энтропия и сложность

- При $n = 2$ более сложными считаются тексты LSTM-модели
- При $n > 2$ более сложными являются литературные тексты
- Для отдельных значений m, n тексты разных типов разделимы на плоскости энтропии-сложности
- Построим классификаторы с метриками энтропии и сложности при разных m, n

Энтропия и сложность

	все боты		**LSTM**		**GPT**	
Язык	Train	Test	Train	Test	Train	Test
вьетнамский	0.981	0.989	1.0	1.0	0.991	0.995
русский	0.879	0.890	0.991	0.992	0.889	0.893
английский	0.937	0.965	0.999	1.0	0.997	1.0

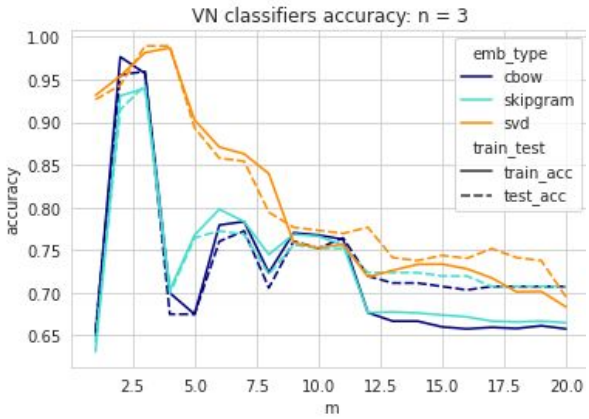
Таблица 6.4: Точность классификаторов по метрикам энтропии и сложности

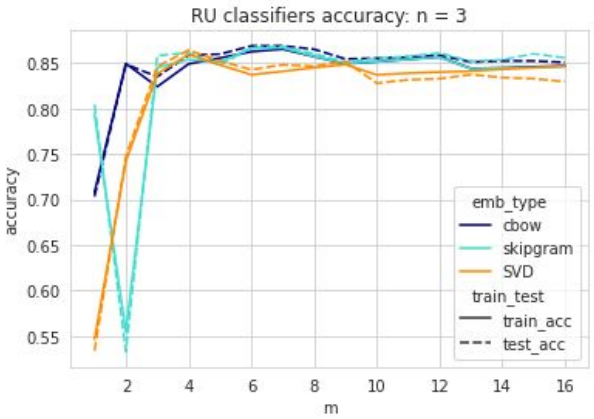
Энтропия и сложность

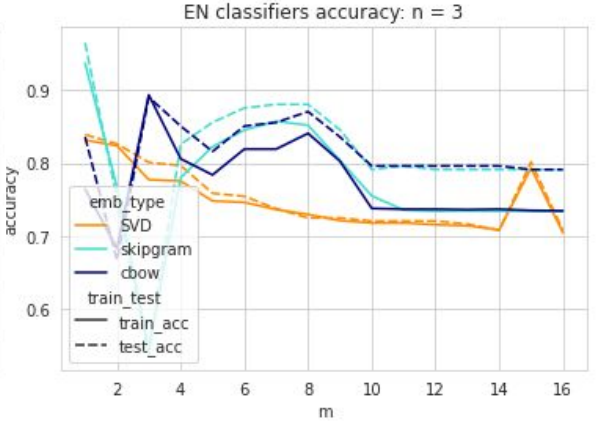
	все боты		**LSTM**		**GPT**	
Язык	Train	Test	Train	Test	Train	Test
вьетнамский	0.979	0.968	1.0	1.0	0.991	0.995
русский	0.828	0.844	0.991	0.992	0.878	0.881
английский	0.913	0.948	0.998	1.0	0.995	1.0

Таблица 6.5: F-мера классификаторов по метрикам энтропии и сложности

Энтропия и сложность







Точность классификации в зависимости от *m* на корпусах вьетнамского (слева), русского (посередине) и английского (справа) языков

Энтропия и сложность

- Наилучшая классификация[1]:
  - Вьетнамский — SVD, m = 3, n = 3
  - Русский — Skip-gram, m = 1, n = 8
  - Английский — Skip-gram, m = 1, n = 3
- Параметры *m, n,* при которых тексты попадают в хаотическую область, приводят к более точным классификаторам

1. Полные таблицы результатов для разных вариантов векторных представлений и значений m, n приведены в репозитории.

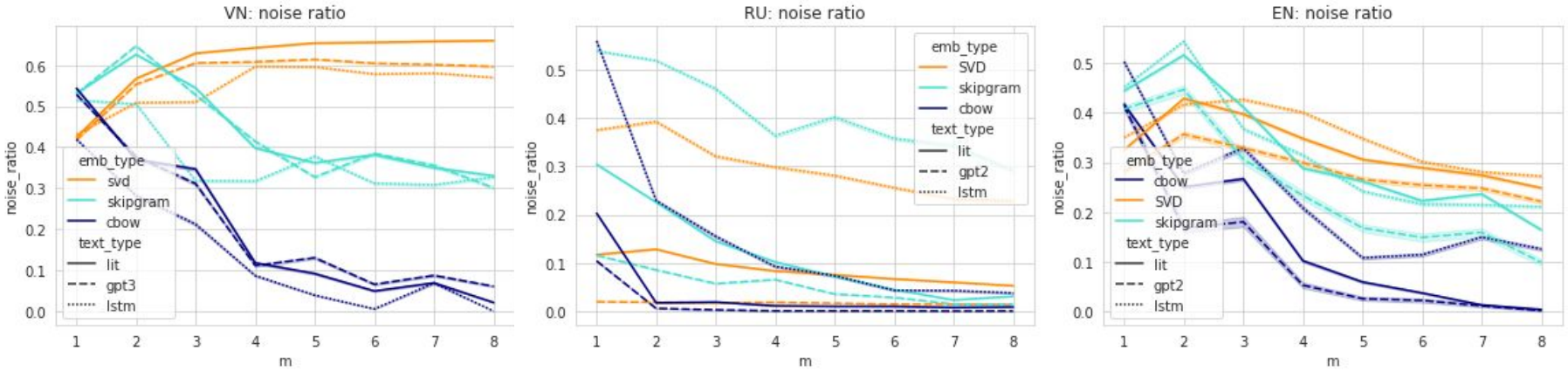
Выводы

- Классификаторы, построенные на основе методов кластеризации и построения плоскости энтропии-сложности, достигают точности 90% и выше
- В кластеризации более точная классификация на основе кластеризации алгоритмом Уишарта
- Нечеткое представление данных улучшает точность обнаружения сложных ботов

Источники

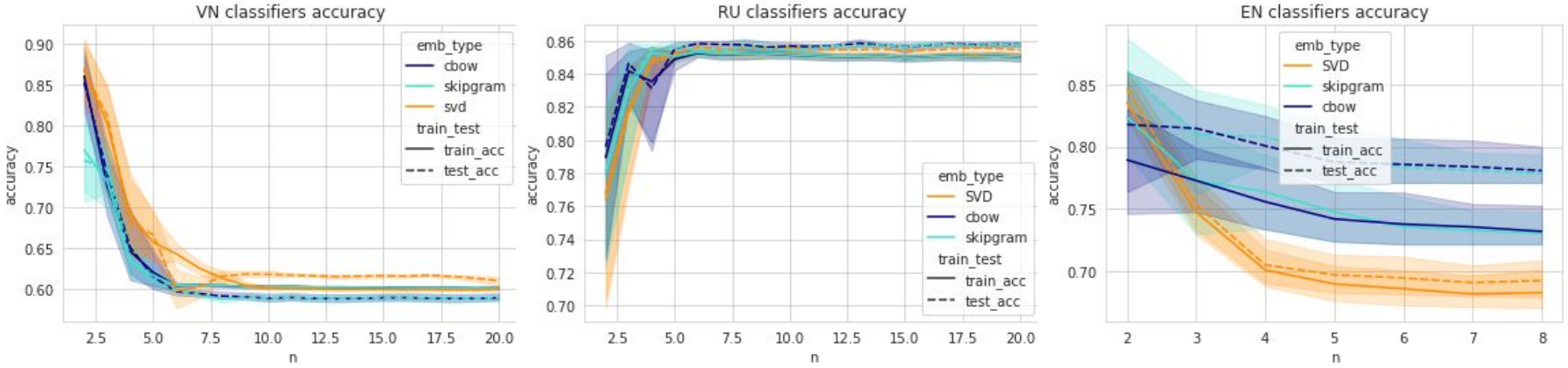
- Репозиторий: https://github.com/quynhu-d/Spot-the-Bot/
- Wishart D.. A numerical classification methods for deriving natural classes.// Nature 221, pp. 97–98, 1969.
- Novák V., Perfilieva I. and Mockor J.. Mathematical principles of fuzzy logic.// Springer Science and Business Media, vol. 517, 2012.
- Bezdek J. C., Ehrlich R. and Full W.. FCM: The fuzzy c-means clustering algorithm. // Computers and geosciences 10.2-3, pp. 191-203, 1984.
- Rosso O. A., Larrondo H. A., Martin M. T., Plastino A. and Fuentes M. A.. Distinguishing noise from chaos. // Physical review letters, vol. 99, no. 15, 154102, 2007.
- Kostenetskiy P.S., Chulkevich R.A., Kozyrev V.I. HPC Resources of the Higher School of Economics // Journal of Physics: Conference Series. 2021.Vol. 1740, No 1. P. 012050.

Кластеризация



Доля шума, выделяемого кластеризацией Уишарта на нечетких числах на корпусах вьетнамского (слева), русского (посередине) и английского (справа) языков

Энтропия и сложность



Точность классификации в зависимости от *n* на корпусах вьетнамского (слева), русского (посередине) и английского (справа) языков