# АНСАМБЛЕВЫЙ МЕТОД В ЗАДАЧЕ ИДЕНТИФИКАЦИИ АВТОРА ТЕКСТА

Труханович И. А., Парамонов А. И.
Кафедра программного обеспечения информационных технологий, кафедра информационных систем и технологий Института информационных технологий, Белорусский государственный университет информатики и радиоэлектроники
Минск, Республика Беларусь
E-mail: ilya.trukhanovich@gmail.com, a.paramonov@bsuir.by

*На текущий момент, в связи с постоянным ростом контента в интернете, область исследования установления автора текста становится важнее с каждым днём. Она заключается в статистическом исследовании лингвистических и вычислительных характеристик текстовых документов, позволяющем установить истинного автора текста. В данной работе описывается обзор различных методов анализа авторства и приводится описание ансамблевого метода, агрегирующего преимущества уже существующих. Также приводятся результаты экспериментов с дальнейшим анализом результатов и потенциальных модификаций, которые могут позволить улучшить имеющийся результат.*

## Введение

Рост важности задачи идентификации текста в основном связан с широким использованием мессенджеров, возрастающим значением электронной почты для корпоративной переписки, популярностью блогов и форумов. Пользователи могут оставлять свои сообщения и без регистрации. Сама регистрация часто носит чисто символический характер. Анонимность интернет-сообщений все больше привлекает киберпреступников [1].

Также методы идентификации автора могут быть применены в других областях. Они могут быть использованы в лингвистических исследованиях или для изучения признаков стиля конкретного автора [2] [3].

Для идентификации авторства применяются разные традиционные подходы, которые основаны на методах NLP (nature language processing) для анализа и классификации текста. В качестве основных используются известные модели мешка слов, n-грамм и т.д. Они позволяют преобразовывать тексты необходимым образом для дальнейшего использования в обучении моделей. Также дополнительно перед анализом используются традиционные методы предварительной обработки текста. Однако каждый отдельно известный подход на сегодня не решает в целом задачу идентификации автора в силу ряда ограничений и сложностей применения.

## I. Постановка задачи

Исследуемая задача формулируется следующим образом.

Предположим, что у нас есть наборы текстов T авторов A. Мы знаем авторов некоторых текстов подмножества T, поэтому у нас есть соответствия авторов и текстов в качестве обучающей выборки L. Цель состоит в том, чтобы создать классификатор, который является функцией, определяющей истинного автора произвольного текста из множества T.

## II. Существующие методы

На данный момент существует определенное количество различных решений, основанных на методах машинного обучения:

- Линейные классификаторы (логистическая регрессия, наивный байесовский метод и т.д.);
- Нелинейные классификаторы (полиномиальная регрессия, сети радиально базисных функций и т.д.).

Все более становится популярным нейросетевой подход, особенно в последние годы, благодаря различным исследованиям и архитектурным усовершенствованиям. Но все же он требует значительного улучшения из-за низкой точности в разных случаях и значительных затрат ресурсов для выбора архитектуры и обучения [4].

## III. Ансамблевый метод

В работе задачу идентификации автора текста в обозначенной формулировке предлагается решать с помощью ансамблевого метода [5].

Ансамблевый метод подразумевает механизм голосования большинством, в котором каждый отдельный классификатор в обучении ансамбля предсказывает метку класса. Класс с большинством голосов (мода) назначается в качестве результирующей метки:

$$\hat{y} = mode\{C_1(x), C_2(x), ..., C_m(x)\}.$$

Ансамблевое обучение обеспечивает лучший результат, чем обычная одиночная модель.

Ансамблевый метод в текущей работе предполагает следующие этапы:

1. Извлечение исходных текстов;
2. Предварительная обработка текстов (удаление служебных слов, стемминг и т.д.);

3. Представление текстов с помощью TF–IDF;
4. Запуск двух классификаторов по отдельности и трёх классификаторов в одном ансамбле;
5. Запуск общего ансамбля из трёх предыдущих компонентов.

Общая схема метода приведена ниже на рисунке 1.

## IV. Эксперимент

Для проведения эксперимента с предложенным методом был отобран набор текстовых документов в виде произвольного пакета статей новостной ленты Reuters. Пакет включал по 8 текстов для каждого из 50 авторов.

В ходе эксперимента выполнялся мониторинг результатов классификаторов, запущенных как поодиночке, так и в составе сформированного ансамбля.

Результаты эксперимента приведены в таблице.

Таблица 1 – Результаты эксперимента

| Метод | Точность | Полнота | F1 |
|---|---|---|---|
| Линейная регресссия | 0.647 | 0.600 | 0.579 |
| kNN | 0.574 | 0.530 | 0.517 |
| Наивный байесовский | 0.584 | 0.560 | 0.539 |
| Ансамбль (линейная регресс-сия, kNN, наивный байесовский) | 0.647 | 0.590 | 0.582 |
| Случайный лес | 0.658 | 0.620 | 0.605 |
| Градиентный бустинг | 0.463 | 0.430 | 0.417 |
| Ансамбль (общий) | 0.673 | 0.640 | 0.618 |

Можно сделать вывод о том, что результатами голосования в ансамблях стало увеличение качества модели при одних и тех же условиях обучения. Минусом является то, что обучение подобной модели является более затратным.

## V. Заключение

Представленный метод на основе комбинаций имеющихся классификаторов предлагает увеличение точности предсказаний. Вместе с тем он может быть усовершенствован в дальнейшем в ряде аспектов:

– Перегруппировка комбинаций в ансамблях;
– Добавление новых линейных и нелинейных классификаторов;
– Замена метода голосования в ансамблях.

## VI. Список литературы


1. Iqbal, F. Machine Learning for Authorship Attribution and Cyber Forensics / F. Iqbal, M. Debbabi, B. Fung. – New York City : Springer, 2020. – 167 p.
2. Griffin, R. J. Anonymity and Authorship. New Literary History / R. J. Griffin. – The Johns Hopkins University Press. – 1999. – Vol. 30, № 4. – P. 877–895.
3. Types of plagiarism [Electronic resource]. – Mode of access: https://www.bradford.ac.uk/library/find-outabout/plagiarism/types-of-plagiarism/. – Date of access: 14.10.2022.
4. Wallace: Author Detection via Recurrent Neural Networks [Electronic resource]. – Mode of access: https://cs224d.stanford.edu/reports/YaoLeon.pdf/. – Date of access: 15.10.2022.
5. A Gentle Introduction to Ensemble Learning Algorithms [Electronic resource]. – Mode of access: https://machinelearningmastery.com/tour-of-ensemble-learning-algorithms/. – Date of access: 15.10.2022.


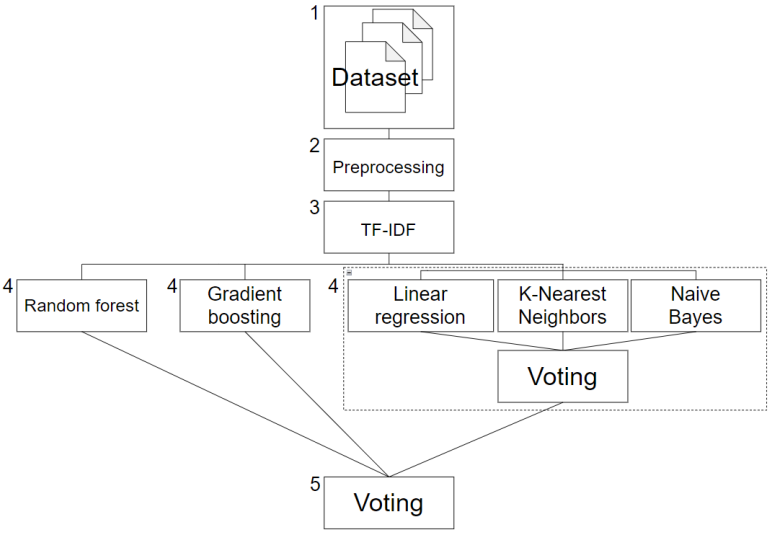


Рис. 1 – Схема ансамбля