АНАЛИЗ КВАНТИЛЬНОГО И ГАУССОВСКОГО ПОДХОДОВ В МОДЕЛИРОВАНИИ РЕДАКЦИОННЫХ ПРОЦЕССОВ

Сьянов Д. А.
Кафедра веб-технологий и компьютерного моделирования,
Белорусский государственный университет
Минск, Республика Беларусь
E-mail: syanov@bsu.by

В работе рассматривается задача вероятностного прогнозирования длительностей процессов в редакционных рабочих потоках научных публикаций. Для решения используется модель на основе архитектуры Transformer с двумя вариантами выходных подсистем: квантильной (с логарифмическим масштабированием целевой переменной) и гауссовской (предсказывающей параметры нормального распределения). Проведено сравнительное исследование эффективности обеих модификаций на реальных данных после глубокой очистки и обогащения признаков.

I. Введение

Прогнозирование сроков прохождения рукописи через редакционные этапы является одной из задач управления научными журналами. От продолжительности каждого этапа зависят планирование редакционной нагрузки, прозрачность коммуникации с авторами и общая эффективность публикационного цикла. Однако временные интервалы между событиями в редакционном процессе характеризуются высокой вариативностью, асимметрией и зависимостью от контекста, что затрудняет применение классических точечных моделей регрессии [1,2].

Рассматривается возможность использования данных о динамике поступления и обработке рукописей в редакционном процессе как аналога интенсивности поступления и интенсивности потока обслуживания заявок в классических системах массового обслуживания. Такая постановка задачи позволяет рассматривать редакционную систему как стохастический процесс с переменной нагрузкой и конечной пропускной способностью, что делает релевантным применение вероятностных моделей.

Для обработки данных используется архитектура Adaptive Probabilistic Sequence-to-Vector Transformer (APS2V-T), предназначенная для вероятностного прогнозирования длительностей этапов редакционного процесса. Модель основана на трансформерном энкодере [3], обрабатывающем последовательности событий и контекстные признаки, и включает две стохастические подсистемы – квантильную и гауссовскую, описывающие распределения длительностей.

II. Формирование датасета

Исходным источником данных выступает система для организации рецензируемых научных изданий Open Journal Systems (OJS), содержащая данные о времени наступления всех редакционных событиях. Из «сырых» логов восстанавливались последовательности событий для каждого идентификатора рукописи, начиная с момента подачи и до публикации.

Процесс подготовки данных включал нормализацию типов событий и унификацию меток редакторских решений с восстановлением пропущенных публикационных событий на основе метаданных, дополнение выборки признаками нагрузки (числом активных материалов, количеством рецензентов и интервалами активности пользователей), а также построение унифицированных последовательностей с фиксированным набором стадий и их выравниванием по общей временной шкале.

Целевые значения рассчитывались как абсолютные длительности (в часах) от даты подачи до наступления последующих стадий. Маскирование целей исключает использование данных о будущих событиях при формировании признаков для прогноза [4]. Для обеспечения корректности моделирования применялись фильтры качества: удалены записи с отсутствующими или противоречивыми временными метками, случаи нулевой или отрицательной задержки публикации. После очистки итоговый корпус составил около 4 тыс. документов и более 30 тыс. событий. Основные статистические характеристики приведены в табл. 1.

Таблица 1 – Распределение длительностей по основным стадиям публикационного процесса

Показатель	Ср.(ч)	мин.	макс.	средн.
нач.рец.	648	4	1956	423
зав.рец	1812	288	2592	595
реш.ред	2430	375	3383	502
нач. лит.ред.	3640	1697	5943	811
публик.	4210	2454	6168	890

Распределения длительностей характеризуются выраженной асимметрией с удлинённым «хвостом» на поздних стадиях, что отражает неоднородность редакционных циклов и различия в динамике работы подразделений. Этот дисбаланс

был учтён при стратификации данных и формировании батчей в процессе обучения модели APS2V-T.

III. Результаты обучения

Обучение модели APS2V-T проводилось отдельно для двух подсистем — квантильной и гауссовской. Основной целью было сравнение способности каждой подсистемы адекватно моделировать распределение временных интервалов ключевых событий редакционного процесса.

По полученным метрикам (табл. 2, рис. 1) видно, что:

- Квантильная аппроксимация демонстрирует существенное превосходство над гауссовской моделью по метрике CRPS [5], что подтверждает эффективность использования дискретной сетки квантилей при моделировании асимметричных и гетероскедастичных временных интервалов;
- Значение MAE для медианного квантиля (q50) в квантильной модели оказывается ниже, чем для точечной оценки в гауссовской, что свидетельствует о более высокой точности воспроизведения центральной тенденции распределения;
- Гауссовская аппроксимация проявляет ограниченную способность к корректному описанию асимметричных распределений длительностей, однако сохраняет ценность в качестве компактного параметрического представления прогноза.

Таблица 2 – Сравнение по ключевым метрикам

Аппрокс.	Epoch	CRPS	NLL	MAE
Квантильная	162	527.742	948.112	24.3
Гауссовская	300	3222.190	4900.509	31.7

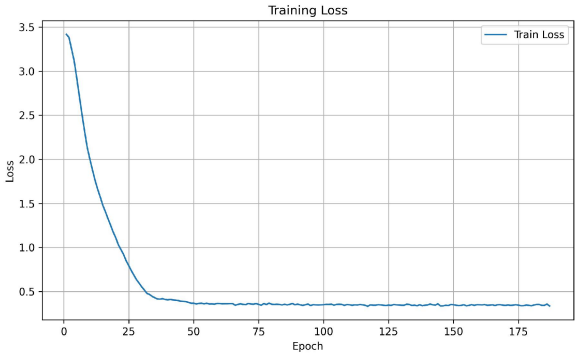

Рис. 1 – train loss квантильная

Качественный анализ предсказаний модели показывает явные различия между квантильной и гауссовской подсистемами APS2V-T. Квантильная аппроксимация корректно отражает вариативность событий на всех этапах редакционного процесса: предсказанные значения $q10$ и $q90$ дают интервал вероятных сроков, позволяя оценить как более ранние, так и более поздние сценарии наступления ключевых событий. Оверлей кривых обучения, представленный на рисунке (рис. 2), демонстрирует, что квантильная аппроксимация сходится быстрее и обеспечивает более стабильное поведение CRPS на валидационном сете, тогда как гауссовская аппроксимация демонстрирует замедленное снижение значения CRPS, что указывает на более медленную сходимость и меньшую устойчивость прогноза.

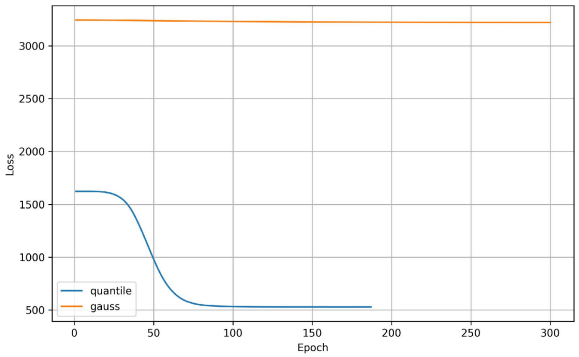

Рис. 2 – CRPS

IV. Заключение

Проверка на отдельных кандидатах, особенно для рукописей с длинными задержками, подтверждает преимущества квантильной подсистемы: она сохраняет реалистичные прогнозы для редких, но значимых длинных хвостов распределения, в то время как гауссовская подсистема систематически недооценивает длительные задержки. Таким образом, дискретная квантильная аппроксимация позволяет точнее прогнозировать медиану и разброс сроков, обеспечивая надёжное представление полной вариативности временных интервалов в редакционном процессе.

1. Wasserman, L. All of Statistics: A Concise Course in Statistical Inference [Electronic resource] : Springer NY, 2013. – Mode of access: https://doi.org/10.1007/978-0-387-21736-9. – Date of access: 23.10.2025.
2. Harrell, F. E. Jr. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis [Electronic resource] : Springer Cham, 2015. – Mode of access: https://doi.org/10.1007/978-3-319-19425-7 – Date of access: 23.10.2025.
3. Vaswani, A. et al. Attention Is All You Need [Electronic resource] : NIPS, 2017. – Mode of access: https://doi.org/10.48550/arXiv.1706.03762. – Date of access: 23.10.2025.
4. Salinas, D., Flunkert, V., Gasthaus, J. , Januschowski, T. DeepAR: Probabilistic forecasting with autoregressive recurrent networks [Electronic resource] : International Journal of Forecasting. – July–September 2020. – Volume 36. – Issue 3. – P. 1181–1191. – Mode of access: https://doi.org/10.1016/j.ijforecast.2019.07.001 – Date of access: 23.10.2025.
5. Gneiting, T., Raftery, A. E. Strictly Proper Scoring Rules, Prediction, and Estimation [Electronic resource] : Journal of the American Statistical Association. – March, 2007. – Vol.102. – No. 477. – P. 359–378. – Mode of access: https://doi.org/10.1198/016214506000001437 – Date of access: 23.10.2025.