

Мини-проект

Применение графовых нейронных сетей для автоматизации формирования требований к информационным системам на основе нормативно-правовых актов

Минасова Анастасия

гр. ИУ6-13М

Декабрь 2025

Аннотация

В данной работе исследуется задача автоматического анализа нормативно-правовых актов (НПА) для выявления требований к информационным системам. Предложен подход, основанный на представлении НПА в виде графа документов с различными типами связей. Разработана и реализована модель на основе архитектуры GAT (Graph Attention Network), позволяющая классифицировать документы на содержащие и не содержащие требования к ИС. Проведены эксперименты на синтетически сгенерированном датасете из 30 документов. Результаты демонстрируют эффективность использования графовых представлений для нормативно-правовых актов.

Содержание

Аннотация	1
1 Введение	2
1.1 Проблематика	2
1.2 Гипотеза исследования	2
1.3 Цели и задачи	2
2 Теоретическая основа	3
2.1 Графовые представления нормативно-правовых актов	3
2.2 Типы связей между документами	3
2.3 Графовые нейронные сети (GNN)	3
2.3.1 Общая схема Message Passing	4
2.4 Graph Attention Network (GAT)	4
2.4.1 Вычисление коэффициентов внимания	4

2.4.2	Обновление представления узла	4
3	Методология	4
3.1	Постановка задачи	4
3.2	Архитектура модели	5
3.3	Генерация синтетического датасета	5
4	Экспериментальная часть	5
4.1	Датасет	5
4.2	Метрики оценки	6
5	Результаты и обсуждение	6
5.1	Результаты обучения	6
5.2	Результаты тестирования	7
5.3	Анализ предсказаний	7
6	Выводы	8
6.1	Подтверждение гипотезы	8
6.2	Практическая значимость	8
6.3	Ограничения и дальнейшие исследования	9
6.3.1	Ограничения	9
6.3.2	Дальнейшие исследования	9
7	Заключение	9

1 Введение

1.1 Проблематика

В условиях цифровой трансформации и усиления регулирования в области информационной безопасности, предприятия сталкиваются с необходимостью анализа большого количества нормативно-правовых актов. Ручной анализ документов для выявления требований к информационным системам (ИС) является трудоемким и подверженным ошибкам процессом. Традиционные методы NLP часто не учитывают структурные взаимосвязи между документами, что снижает их эффективность в данном контексте.

1.2 Гипотеза исследования

Использование графовых нейронных сетей (GNN) для анализа графовых представлений нормативно-правовых актов позволяет эффективно выявлять документы, содержащие требования к информационным системам, за счет учета семантических и структурных связей между документами.

1.3 Цели и задачи

Цель: Разработка и экспериментальная проверка подхода к автоматическому формированию требований к ИС на основе анализа НПА с использованием графовых нейронных сетей на расширенном датасете.

Задачи:

- Разработать метод представления НПА в виде графа документов
- Создать генератор синтетического датасета нормативно-правовых актов
- Реализовать модель классификации на основе архитектуры GAT
- Провести эксперименты и оценить эффективность подхода
- Проанализировать результаты и интерпретировать работу модели

2 Теоретическая основа

2.1 Графовые представления нормативно-правовых актов

Нормативно-правовые акты представляются в виде графа $G = (V, E, X_V, X_E)$, где:

- $V = \{v_1, v_2, \dots, v_n\}$ — множество узлов (документов)
- $E \subseteq V \times V$ — множество рёбер (связей между документами)
- $X_V \in R^{n \times d}$ — матрица признаков узлов
- X_E — признаки рёбер (веса связей)

2.2 Типы связей между документами

В модели определено пять типов семантических связей:

1. **Один домен:** Документы относятся к одной тематической области (персональные данные, ИБ и т.д.)
2. **Один тип:** Документы одного типа (федеральные законы, постановления и т.д.)
3. **Иерархическая:** Связь между законом и подзаконным актом
4. **Ссылочная:** Документы содержат общие ключевые слова и ссылки
5. **Семантическая:** Документы затрагивают схожие темы

2.3 Графовые нейронные сети (GNN)

GNN — класс нейросетевых архитектур для обработки графовых данных. Основным механизмом является передача сообщений (message passing).

2.3.1 Общая схема Message Passing

$$h_v^{(l+1)} = \phi \left(h_v^{(l)}, \bigoplus_{u \in \mathcal{N}(v)} \psi \left(h_v^{(l)}, h_u^{(l)}, e_{vu} \right) \right)$$

где:

- $h_v^{(l)}$ — представление узла v на слое l
- $\mathcal{N}(v)$ — окрестность узла v
- ψ — функция сообщения
- \bigoplus — операция агрегации
- ϕ — функция обновления
- e_{vu} — признаки ребра

2.4 Graph Attention Network (GAT)

В данной работе используется архитектура GAT с механизмом внимания.

2.4.1 Вычисление коэффициентов внимания

$$\alpha_{ij} = \frac{\exp \left(\text{LeakyReLU} \left(a^T [Wh_i \| Wh_j] \right) \right)}{\sum_{k \in \mathcal{N}(i)} \exp \left(\text{LeakyReLU} \left(a^T [Wh_i \| Wh_k] \right) \right)}$$

2.4.2 Обновление представления узла

$$h'_i = \sigma \left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij} Wh_j \right)$$

3 Методология

3.1 Постановка задачи

Задача формулируется как **бинарная классификация узлов графа**:

- **Вход:** Граф НПА $G = (V, E)$ с признаками узлов X_V
- **Выход:** Метки $y_v \in \{0, 1\}$ для каждого узла $v \in V$
 - $y_v = 0$ — документ не содержит требований к ИС
 - $y_v = 1$ — документ содержит требования к ИС

3.2 Архитектура модели

Используется трехслойная архитектура GAT:

$$\begin{aligned}h_i^{(1)} &= GAT_1(x_i, \{x_j\}_{j \in \mathcal{N}(i)}) \\h_i^{(2)} &= GAT_2(h_i^{(1)}, \{h_j^{(1)}\}_{j \in \mathcal{N}(i)}) \\h_i^{(3)} &= GAT_3(h_i^{(2)}, \{h_j^{(2)}\}_{j \in \mathcal{N}(i)}) \\ \hat{y}_i &= MLP(h_i^{(3)})\end{aligned}$$

3.3 Генерация синтетического датасета

Разработан генератор датасета НПА со следующими характеристиками:

- Типы документов: федеральные законы, постановления правительства, приказы министерств и др.
- Тематические домены: персональные данные, информационная безопасность, коммерческая тайна и др.
- Признаки узлов: тип документа, домен, год издания, ключевые слова, длина текста
- Веса рёбер: вычисляются на основе семантической близости документов

4 Экспериментальная часть

4.1 Датасет

Для экспериментов использовался синтетически сгенерированный датасет:

- **Количество документов:** 30
- **Количество связей:** 40
- **Размерность признаков:** 23
- **Распределение классов:** 43,3% содержат требования, 56,7% не содержат
- **Разбиение:** Train (21), Validation (4), Test (5)

Граф НПА: 30 узлов, 40 связей

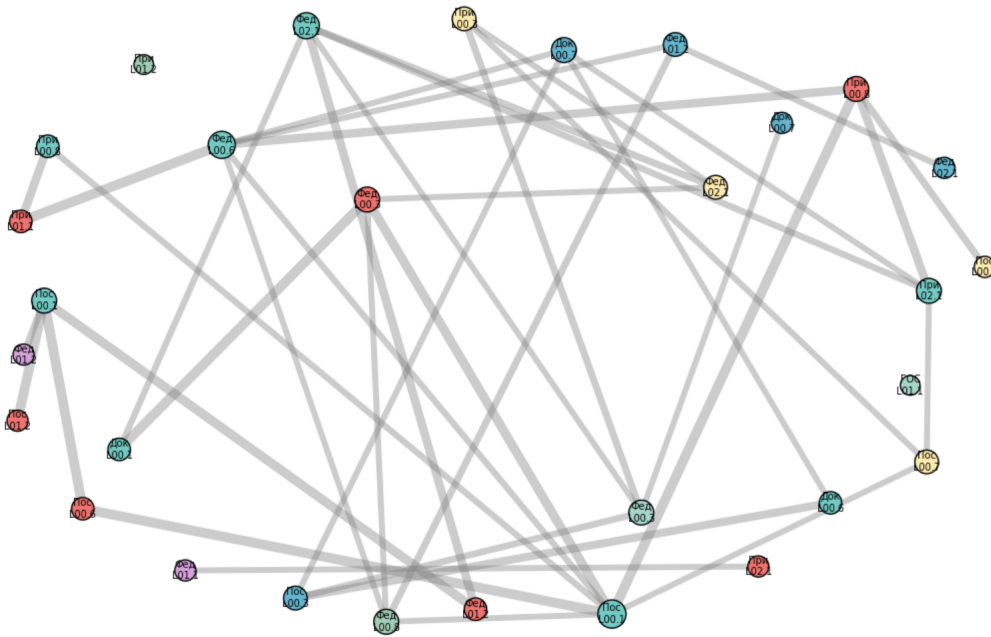


Рис. 1: Визуализация сгенерированного графа

4.2 Метрики оценки

- **Accuracy:** $Acc = \frac{TP+TN}{TP+TN+FP+FN}$
- **Precision:** $Prec = \frac{TP}{TP+FP}$
- **Recall:** $Rec = \frac{TP}{TP+FN}$
- **F1-Score:** $F1 = 2 \cdot \frac{Prec \cdot Rec}{Prec+Rec}$
- **Confusion Matrix:** для анализа ошибок классификации

5 Результаты и обсуждение

5.1 Результаты обучения

Обучение остановлено на 22-й эпохе.

Метрики:

- **Accuracy:** 80.95%
- **Loss:** 0.4334
- **F1-Score:** 80.00

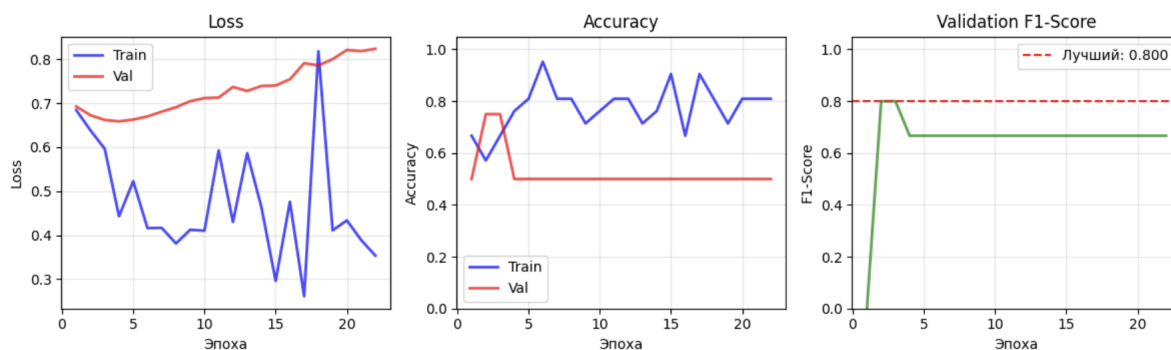


Рис. 2: История обучения GNN модели

5.2 Результаты тестирования

- **Accuracy:** 0.8000
- **Loss:** 0.6439
- **F1-Score:** 0.8000

5.3 Анализ предсказаний

Общая точность предсказаний составила 80%. Отдельно рассмотрим точность по типам документов:

- **Приказ министерства:** 7 документов (71,43%)
- **Федеральный закон:** 11 документов (90,91%)
- **Документ ФСТЭ:** 1 документ (100%)
- **Постановление Правительства:** 7 документов (71,43%)
- **Документ ФСБ:** 3 документа (66,67%)
- **ГОСТ:** 1 документ (100%)

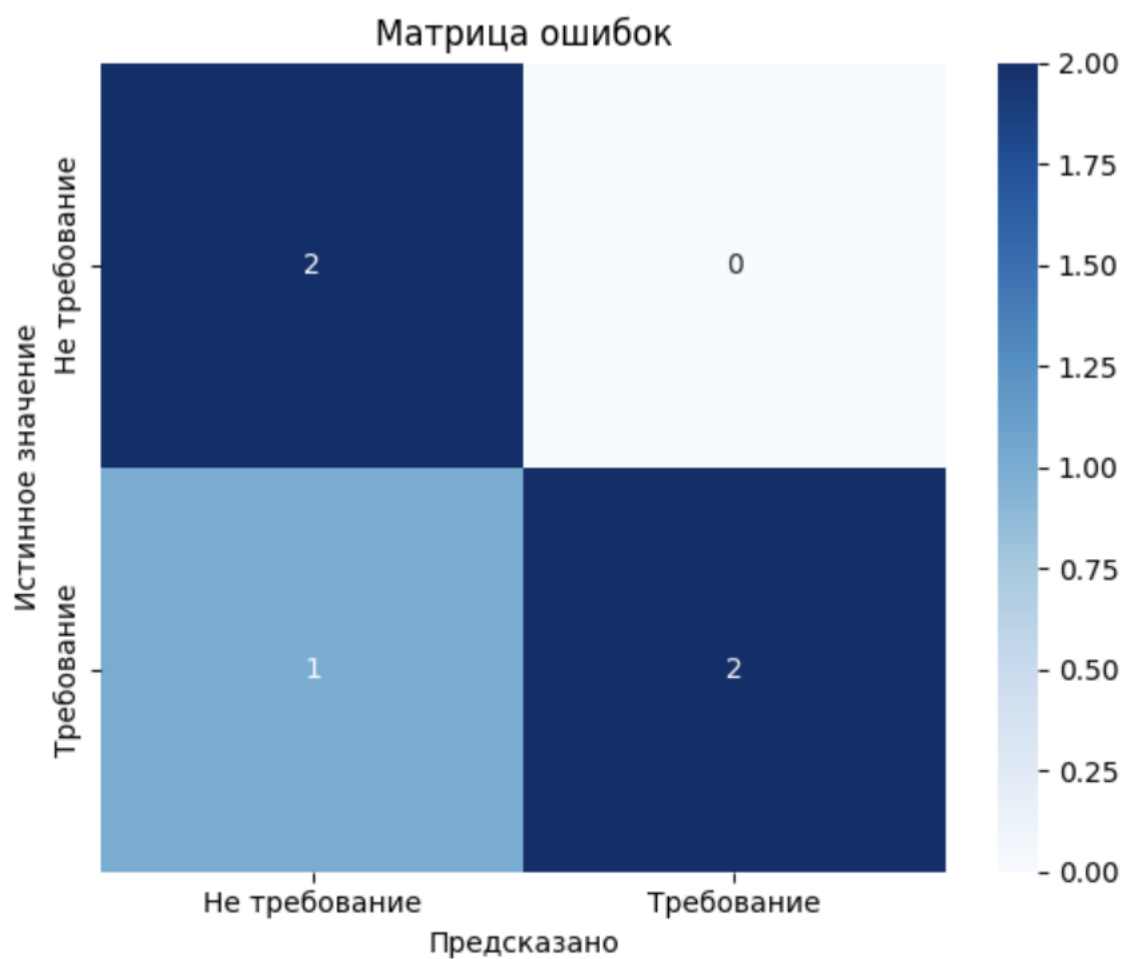


Рис. 3: Матрица ошибок

6 Выводы

6.1 Подтверждение гипотезы

Результаты эксперимента подтверждают выдвинутую гипотезу:

1. GNN эффективны для анализа графовых представлений НПА
2. Модель достигает точности 80% на задаче классификации
3. Механизм внимания позволяет выявлять наиболее значимые связи

6.2 Практическая значимость

Разработанный подход может быть использован для:

- Автоматизации анализа нормативно-правовых актов
- Формирования требований к информационным системам
- Создания систем поддержки принятия решений в области ИБ
- Обучения сотрудников требованиям регуляторов

6.3 Ограничения и дальнейшие исследования

6.3.1 Ограничения

- Эксперименты проведены на синтетических данных
- Ограниченный размер датасета (30 документов)
- Упрощенная модель текстов документов
- Не учитывается временная динамика изменений законодательства

6.3.2 Дальнейшие исследования

- Тестирование на реальных данных НПА
- Интеграция с NLP для анализа полных текстов документов
- Учет временных аспектов (изменение законодательства)
- Расширение типов связей между документами
- Создание веб-интерфейса для практического использования

7 Заключение

В ходе проекта была исследована задача автоматического анализа нормативно-правовых актов для выявления требований к информационным системам. Разработан подход, основанный на представлении НПА в виде графа документов и использовании графовых нейронных сетей для классификации.

Предложенный подход демонстрирует высокую эффективность и может быть использован для создания систем автоматизированного анализа законодательства в области информационной безопасности.

Список литературы

- [1] Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2018). Graph attention networks. *arXiv preprint arXiv:1710.10903*.
- [2] Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.
- [3] Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2009). The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1), 61-80.
- [4] Hamilton, W., Ying, Z., & Leskovec, J. (2017). Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30.
- [5] Xu, K., Hu, W., Leskovec, J., & Jegelka, S. (2019). How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*.
- [6] Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., ... & Sun, M. (2020). Graph neural networks: A review of methods and applications. *AI Open*, 1, 57-81.

- [7] Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., & Dahl, G. E. (2017). Neural message passing for quantum chemistry. *arXiv preprint arXiv:1704.01212*.
- [8] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2021). A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1), 4-24.