**Analýza a návrh riešenia**

**Názov:** *Sémantické a štrukturálne rozdelenie súdnych rozhodnutí*

**Autor:** Bc. Attila Papán

**Vedúci:** RNDr. Peter Gurský, PhD.

**Konzultant:** RNDr. Zoltán Szoplák

**Ciele:**

1. Oboznámiť sa so štruktúrou Slovenských súdnych rozhodnutí a metódami sémantickej segmentácie
2. Navrhnúť a implementovať metódu na automatizáciu členenia textu rozhodnutia na jednotlivé časti podľa obsahu.
3. Vyhodnotiť úspešnosť implementovaného prístupu s anotovanou sadou od právnych expertov.

**1. Analýza problému**

**1.1 Kontext a motivácia**

Súdne rozhodnutia predstavujú kľúčový prvok právnej praxe a právnej vedy. Ich porozumenie, analýza a porovnávanie zohrávajú významnú úlohu nielen v rozhodovacej činnosti súdov, ale aj v príprave právnych podaní, výskume či pri právnom poradenstve. Slovenské súdy každý rok generujú obrovské množstvo rozhodnutí, ktoré sú sprístupňované verejnosti cez rôzne databázy. Problémom však ostáva ich neštruktúrovaná forma – väčšina týchto textov je publikovaná ako súvislý text bez formálneho členenia na jednotlivé logické časti.

Takýto stav výrazne komplikuje možnosť rýchlej orientácie v texte, automatizované vyhľadávanie relevantných informácií a najmä ich ďalšie spracovanie strojovými nástrojmi. Pre právnikov to znamená časovo náročné manuálne čítanie a prechádzanie rozsiahlych textov, pričom často hľadajú iba konkrétnu časť – napríklad právnu argumentáciu alebo samotné rozhodnutie súdu.

Našou ambíciou je tento problém zmierniť pomocou automatizovaného sémantického a štrukturálneho rozdelenia týchto dokumentov na jednotlivé časti. Takto vznikne nový typ pološtruktúrovaných dokumentov, ktoré umožnia efektívnejšie vyhľadávanie, indexovanie a analýzu právnych textov. V dlhodobom horizonte sa otvárajú možnosti pre tvorbu inteligentných právnych asistentov, systematické porovnávanie rozhodnutí, či efektívne monitorovanie súdnej praxe.

**1.2 Charakteristika dát a výzvy**

Základným materiálom našej práce je rozsiahla databáza slovenských súdnych rozhodnutí, ktorá obsahuje viac ako 4 milióny dokumentov. Tieto dokumenty pochádzajú z rôznych druhov konaní (napr. trestné, občianske, správne, rodinné právo) a rôznych súdnych stupňov, čo zvyšuje variabilitu ich obsahu i štylistickej realizácie.

Každé rozhodnutie obsahuje viacero implicitne prítomných častí, ako napríklad:

* hlavičku (údaje o súdnom orgáne, spisovej značke a účastníkoch),
* úvodné konštatovanie predmetu sporu,
* skutkový stav (fakty prípadu, opis priebehu konania),
* právnu analýzu a aplikáciu právnych noriem,
* samotný výrok (rozhodnutie súdu),
* odôvodnenie rozhodnutia.

Tieto časti však nie sú explicitne oddelené formálnymi nadpismi, štruktúrovaním alebo značkami, ktoré by umožňovali jednoduché spracovanie. Navyše, každý sudca má mierne odlišný štýl, čo vedie k značnej jazykovej variabilite. V porovnaní s inými typmi dokumentov (napr. vedeckými článkami či legislatívnymi textami) sú teda súdne rozhodnutia oveľa menej predvídateľné z hľadiska štruktúry.

Ďalšou výzvou je jazyk samotný – slovenský právny jazyk má špecifické charakteristiky, používa odborné výrazy, archaizmy a štylisticky náročné konštrukcie. Navyše, v oblasti spracovania prirodzeného jazyka (NLP) neexistuje taký rozsiahly ekosystém nástrojov pre slovenčinu ako pre angličtinu. Aj preto je výber vhodných modelov a prístupov rozhodujúci pre úspešnosť riešenia.

**1.3 Doterajší výskum a existujúce prístupy**

Problematika segmentácie právnych textov bola riešená najmä v angloamerickom právnom prostredí. Napríklad výskum Jaromíra Šavelku a Kevina D. Ashleyho sa zameriaval na segmentáciu amerických súdnych rozhodnutí do funkčných častí (napr. výrok, fakty, právna analýza), pričom využívali rôzne metódy strojového učenia a klasifikácie na základe anotovaných datasetov. V Európskom kontexte však nie je dostupná takáto štandardizovaná segmentácia, najmä nie pre menšie jazyky ako slovenčina.

Pri tradičných metódach segmentácie sa využívajú pravidlá (rule-based prístupy), ktoré identifikujú časti textu na základe typických fráz (napr. „súd rozhodol“, „na základe vykonaného dokazovania“ atď.). Tieto pravidlá sú však často príliš rigidné a nedokážu zvládnuť jazykovú variabilitu reálneho textu. Preto v posledných rokoch nastáva presun ku kombinovaným prístupom, ktoré spájajú sémantickú analýzu s jazykovými modelmi trénovanými na právnom korpuse.

Pre slovenské právne texty je vhodným kandidátom **SlovakBERT** – transformerový model optimalizovaný pre slovenčinu. Ďalšou alternatívou je **SlovakSentenceBERT**, ktorý generuje vektorové reprezentácie viet a umožňuje porovnanie ich významovej podobnosti. Tieto modely môžu byť využité na klasifikáciu viet alebo úsekov textu do sémantických kategórií. Navyše, moderné jazykové modely ako **Mistral-7B** a **LLaMA** umožňujú aj generatívny prístup k analýze štruktúry dokumentov.

**1.4 Formulácia cieľov segmentácie**

Cieľom našej práce je vyvinúť a experimentálne overiť metódu, ktorá umožní automatické rozdelenie textov súdnych rozhodnutí do logických a sémantických celkov. Tieto segmenty zodpovedajú častiam, ktoré právnici prirodzene vnímajú ako samostatné jednotky počas čítania dokumentu. Na rozdiel od bežnej segmentácie textu podľa odsekov či interpunkcie, naša segmentácia má mať **sémantickú hĺbku** a **štrukturálnu konzistenciu**.

Konkrétne sa zameriavame na detekciu a extrakciu týchto častí:

* **Hlavička rozhodnutia** – obsahuje identifikačné údaje o súde, čísle konania, menách účastníkov a ich právnom zastúpení.
* **Úvod** – stručne predstavuje právny spor alebo predmet rozhodovania.
* **Opis skutkového stavu** – podrobný popis okolností prípadu, dôkazy a priebeh konania.
* **Právna analýza** – výklad a aplikácia relevantných právnych predpisov.
* **Rozhodnutie (výrok)** – jednoznačné zhrnutie rozhodnutia súdu (napr. rozsudok, uznesenie).
* **Odôvodnenie** – vysvetlenie, prečo súd rozhodol určitým spôsobom, vrátane reakcie na argumenty strán.

Tieto časti nie sú striktne oddelené v texte, a preto musí byť náš systém schopný identifikovať ich **na základe obsahu, kontextu a štylistických znakov**, nie len na základe formálnych znakov (napr. nadpisy či odrážky, ktoré často chýbajú).

Hlavné **ciele segmentácie** môžeme zhrnúť nasledovne:

* **Zvýšiť čitateľnosť** rozhodnutí pre ľudí – zobrazenie textu po segmentoch zlepšuje orientáciu.
* **Podporiť efektívne vyhľadávanie** – možnosť hľadať kľúčové pojmy iba v odôvodnení alebo právnej analýze.
* **Zefektívniť indexovanie textov** – pre právnické databázy alebo právne asistenčné systémy.
* **Vytvoriť základ pre ďalšie úlohy** – napríklad sumarizáciu rozsudkov, detekciu precedensov, či odporúčacie systémy pre právnikov.

Z dlhodobého hľadiska je tiež cieľom preskúmať, či možno vytvoriť **vzorové šablóny segmentácie** pre rôzne typy súdnych rozhodnutí (napr. rozsudky vs. uznesenia), čím by sa zvýšila presnosť segmentácie a možnosť prenosu metódy aj na iné typy právnych dokumentov.

**1.5 Technologické predpoklady a výzvy pri spracovaní slovenského právneho textu**

Spracovanie prirodzeného jazyka (NLP) sa v poslednom desaťročí dramaticky zmenilo vďaka nástupu hlbokých neurónových sietí a najmä transformerových jazykových modelov, ktoré umožňujú efektívne zachytávať sémantiku viet a kontextuálne závislosti. Tieto modely, ako BERT, RoBERTa či GPT, sú však primárne trénované na anglických dátach. V prípade slovenčiny – a najmä právnej slovenčiny – čelíme viacerým špecifickým výzvam:

**1.5.1 Nedostatok anotovaných právnych dát**

Na rozdiel od angličtiny, kde existujú rozsiahle anotované právne korpusy (napr. US Case Law, EUR-Lex dataset), v slovenčine nie sú dostupné verejné datasety s označenými časťami rozsudkov. To komplikuje použitie metód podmienených učením (supervised learning), ktoré vyžadujú ručne anotované trénovacie dáta. Preto je potrebné uvažovať nad polopodmieneným alebo bezdozorovým učením (semi-/unsupervised), prípadne kombinovať strojové učenie s heuristikami alebo transfer learningom.

**1.5.2 Špecifickosť právneho jazyka**

Právny jazyk sa vyznačuje vysokou mierou formalizmu, použitím odborných výrazov, častým citovaním zákonov a archaickými vetnými štruktúrami. Navyše, právne texty často používajú formulácie typu „na základe vyššie uvedeného súd rozhodol…“, ktoré sú síce štylisticky podobné, no ich význam sa môže líšiť v závislosti od kontextu. To kladie vysoké nároky na jazykové modely, ktoré musia byť dostatočne citlivé na významové nuansy.

**1.5.3 Nízka dostupnosť NLP nástrojov pre slovenčinu**

Väčšina populárnych NLP knižníc (spaCy, NLTK, AllenNLP) má len obmedzenú podporu pre slovenský jazyk. Základné nástroje ako tokenizácia či lematizácia sú dostupné, no pokročilé úlohy ako segmentácia textu, extrakcia významových jednotiek či klasifikácia viet podľa obsahu sú zriedkavé alebo chýbajú úplne.

Z tohto dôvodu je potrebné siahnuť po jazykových modeloch špecificky trénovaných na slovenčine. Jedným z najperspektívnejších je **SlovakBERT**, ktorý je trénovaný na veľkom množstve slovenského textu vrátane legislatívnych dokumentov a správ. Tento model je vhodný na extrakciu sémantických reprezentácií slov a viet, a možno ho prispôsobiť na klasifikáciu alebo segmentáciu dokumentov pomocou jemného doladenia (fine-tuning).

Ďalšou možnosťou je **SlovakSentenceBERT**, ktorý rozširuje architektúru BERT o možnosť generovať porovnateľné vektorové reprezentácie celých viet alebo odsekov. Tento prístup je užitočný napríklad na výpočet podobnosti medzi vetami alebo na klastrovanie viet s podobnou funkciou (napr. všetky odôvodnenia).

Zvažujeme tiež experimentálne nasadenie veľkých jazykových modelov ako **Mistral-7B** alebo **LLaMa**, ktoré síce nie sú špecificky trénované na slovenčine, ale v kombinácii s promptovaním a adaptáciou môžu poskytnúť prekvapivo dobré výsledky najmä pri generatívnych úlohách alebo pri transfer learningu.

**1.5.4 Vyváženie presnosti a výpočtovej náročnosti**

Pre praktické nasadenie je dôležité nájsť rovnováhu medzi presnosťou segmentácie a výpočtovými nárokmi riešenia. Veľké modely (ako transformerové architektúry) sú výpočtovo náročné a ich tréning alebo aj inferencia na veľkých datasetoch si vyžaduje GPU infraštruktúru. Preto bude potrebné zvážiť kompromisy – napríklad použiť menšie modely na predspracovanie textu a veľké modely len na finálnu analýzu, alebo využívať dávkované spracovanie.

**1.6 Porovnanie možných prístupov k segmentácii textu**

Segmentácia súdnych rozhodnutí je komplexná úloha, pretože texty nie sú formálne štruktúrované, majú vysokú variabilitu, špecifický jazyk a neobsahujú explicitné nadpisy. V tejto časti analyzujeme rôzne riešenia segmentácie dokumentov – od jednoduchých heuristických pravidiel až po pokročilé modely hlbokého učenia – a hodnotíme ich aplikovateľnosť na slovenské právne texty.

**1.6.1 Heuristické (pravidlové) delenie**

**Popis:**
Pravidlové metódy vychádzajú z identifikácie pevných jazykových štruktúr – napríklad frázy ako „súd rozhodol“, „na základe vykonaného dokazovania“, „vzhľadom na uvedené“, ktoré často signalizujú začiatok alebo koniec určitej časti rozhodnutia.

**Výhody:**

* Jednoduchá implementácia.
* Vysoká interpretovateľnosť (ľahké vysvetlenie, prečo bol text rozdelený).
* Minimálna potreba výpočtových zdrojov.

**Nevýhody:**

* Slabá robustnosť voči štylistickej variabilite textu.
* Zlyháva pri „nečistých“ alebo štylisticky odlišných dokumentoch.
* Neposkytuje sémantické informácie (nerozumie významu vety).

**Použiteľnosť v tomto projekte:**
Skôr ako doplnkový mechanizmus – môže pomôcť predspracovať text, no samostatne neposkytne dostatočnú presnosť.

**1.6.2 Rozdelenie na základe podobnosti embeddingov**

**Popis:**
Vety sú prevedené do embeddingov (napr. pomocou Sentence-BERT), ktoré reprezentujú ich význam. Následne sa porovnáva podobnosť medzi po sebe idúcimi vetami – nízka podobnosť indikuje tematickú zmenu (potenciálnu hranicu segmentu).

**Výhody:**

* Zachytáva skutočný význam textu (sémantika).
* Nevyžaduje ručnú anotáciu (unsupervised).
* Možno vizualizovať a manuálne overiť.

**Nevýhody:**

* Citlivé na výber prahu (manuálne ladenie).
* Niekedy deteguje hranice aj tam, kde nie sú (napr. len kvôli zložitej vete).
* Neurčuje typ segmentu – iba hranicu.

**Použiteľnosť v tomto projekte:**
Už otestované – výborné pre detekciu prechodov. Vhodné ako základná metóda segmentácie. Možno rozšíriť o dynamické prahovanie a klasifikáciu.

**1.6.3 Supervidovaná klasifikácia viet**

**Popis:**
Každá veta je klasifikovaná do jednej z tried (napr. „úvod“, „skutkový stav“, „právna analýza“, „výrok“) na základe trénovacieho datasetu, kde boli vety ručne označené.

**Modely:**

* BERT/Sentence-BERT + lineárna vrstva (softmax),
* klasické modely: Random Forest, SVM (so syntaktickými a lexikálnymi črtami),
* transformerové klasifikátory s fine-tuningom.

**Výhody:**

* Presná identifikácia typu každej vety.
* Umožňuje priame označenie segmentov.
* Pripravené na nasadenie, ak existuje anotovaný dataset.

**Nevýhody:**

* Vyžaduje kvalitné anotované dáta (časovo náročné vytvoriť).
* Modely môžu byť citlivé na doménové zmeny.
* Zložitejšie ladenie a overovanie.

**Použiteľnosť v tomto projekte:**
Vysoko odporúčané ako ďalší krok. Potrebuje však najprv malý ručne anotovaný korpus na trénovanie.

**1.6.4 Klastrovanie embeddingov**

**Popis:**
Vety sa rozdelia do skupín (klastrov) na základe podobnosti embeddingov bez anotácie. Potom sa každému klastru môže priradiť typ segmentu, prípadne sa klastre používajú na vizuálne rozdelenie textu.

**Modely:**

* K-means,
* DBSCAN,
* hierarchické zhlukovanie.

**Výhody:**

* Nevyžaduje anotácie (unsupervised).
* Možnosť objaviť nečakané štruktúry v texte.
* Dá sa kombinovať s vizualizáciou (napr. PCA alebo t-SNE).

**Nevýhody:**

* Klastre nemusia mať jednoznačný právny význam.
* Počet klastrov treba zvoliť (alebo detegovať automaticky).
* Možná „rozmazanosť“ hraníc medzi segmentmi.

**Použiteľnosť v tomto projekte:**
Vhodné ako doplnková analýza alebo na predtriedenie viet, ale nie ako finálne riešenie segmentácie.

**1.6.5 Sekvenčné modely (napr. BiLSTM, CRF, transformer encoder)**

**Popis:**
Text sa spracúva ako sekvencia viet alebo tokenov, pričom sa využíva kontext okolných viet na predikciu segmentu. Typické pre úlohy sekvenčného označovania.

**Modely:**

* BiLSTM + CRF,
* transformer encoder (napr. Longformer, LED),
* BERT + CRF alebo softmax nad výstupmi.

**Výhody:**

* Zachytáva globálny kontext dokumentu.
* Dokáže zohľadniť „tok“ textu – nie len jednotlivé vety.
* Vhodné aj pre dlhé dokumenty.

**Nevýhody:**

* Výpočtovo náročné, náročné na trénovanie.
* Vyžaduje veľa dát alebo doladenie z inej domény.
* Komplexná implementácia.

**Použiteľnosť v tomto projekte:**
Dlhodobo perspektívne riešenie. Momentálne náročné bez silného GPU a väčšieho množstva anotácií, ale vhodné pre rozšírenú verziu systému.

**2. Návrh riešenia**

**2.1 Architektúra systému**

Navrhované riešenie pozostáva z niekoľkých modulov, ktoré zabezpečujú spracovanie textu od surového vstupu až po segmentovaný výstup. Systém je navrhnutý ako **modulárny a rozšíriteľný**, aby bolo možné ľahko experimentovať s rôznymi komponentmi (napr. výmena embedding modelu, klasifikátora, vizualizačnej vrstvy).

**2.1.1 Hlavné komponenty**

1. **Predspracovanie dokumentov**
	* Tokenizácia dokumentov do viet pomocou heuristík alebo nltk.
	* Odstránenie šumu, nadbytočných znakov a neúplných viet.
2. **Generovanie embeddingov**
	* Použitie modelu SlovakBERT alebo SlovakSentenceBERT na prevod viet do vektorovej formy.
	* Ukladanie embeddingov v štruktúrovanom formáte (JSON).
3. **Analýza podobnosti**
	* Výpočet kosínovej podobnosti medzi vetami.
	* Identifikácia hraníc segmentov podľa prahovej hodnoty alebo dynamického prahu (napr. na základe distribúcie).
4. **Segmentácia**
	* Rozdelenie dokumentu podľa identifikovaných hraníc.
	* Voliteľná klasifikácia segmentov (úvod, skutkový stav, analýza, výrok…).
5. **Vizualizácia a export**
	* Zobrazenie podobností a segmentačných bodov cez grafy.
	* Export segmentovaného dokumentu s anotáciami.

**2.2 Návrh vlastného postupu segmentácie**

Pri návrhu vlastného riešenia sme sa inšpirovali jednak prácami z oblasti právneho NLP (napr. Savelka & Ashley, 2018, [Lattisi et al., 2022]), ako aj konkrétnym článkom *Transformer-based Deep Semantic Segmentation of Legal Court Decisions* (arXiv:2012.03619), ktorý navrhuje vetovú klasifikáciu ako základnú metódu segmentácie právnych dokumentov. Naše riešenie vychádza z kombinácie týchto poznatkov s praktickými experimentmi v prostredí slovenského právneho jazyka.

Navrhovaný postup pozostáva z nasledovných krokov:

**2.2.1 Predspracovanie textu a tokenizácia**

Najskôr sa každý dokument rozdelí na jednotlivé vety pomocou vhodného tokenizačného nástroja. Keďže vety v právnych textoch môžu byť zložité a dlhé, uprednostnili sme modely schopné zachytiť ich významovú jednotu. Pre rozdelenie viet používame predspracovanie pomocou knižnice nltk, v kombinácii s heuristikami pre slovenské právne texty.

**2.2.2 Vygenerovanie embeddingov viet**

Každá veta sa následne prevedie na numerickú reprezentáciu pomocou modelu kinit/slovakbert-sts-stsb, ktorý je založený na architektúre Sentence-BERT. Tento model je trénovaný špeciálne na výpočty sémantickej podobnosti medzi vetami v slovenčine, čo je kľúčové pri rozhodovaní o tom, ktoré vety patria do rovnakého segmentu.

**2.2.3 Detekcia hraníc pomocou kosínovej podobnosti**

Vety v rámci rozhodnutia sú zoradené sekvenčne a medzi každými dvoma susednými vetami sa vypočíta **kosínová podobnosť** ich embeddingov. Miesta s nízkou podobnosťou môžu signalizovať prechod do novej tematickej časti. Experimentálne sme overovali rôzne **prahové hodnoty (napr. 0.2)** a posudzovali, kde dochádza k významovému „zlomu“.

Alternatívne testujeme **rolling-average metódu**, kde sa porovnáva priemer z viacerých viet (napr. 5) s nasledujúcou vetou, čím sa dosiahne vyššia odolnosť voči šumu a náhodným jazykovým výkyvom.

**2.2.4 Identifikácia segmentov**

Na základe zistených „zlomových“ bodov rozdelíme dokument na segmenty. Tieto segmenty následne môžeme:

* označiť ručne (v prípade menšieho počtu dokumentov),
* klasifikovať automaticky pomocou modelu (napr. vetová klasifikácia podľa článku z arXivu),
* porovnať ich s anotovanými dátami, ak budú k dispozícii (napr. testovací set od právnikov).

**2.2.5 Výstup a vizualizácia**

Výstupy sú uložené vo formáte JSON, pričom každý segment obsahuje:

* text,
* pozíciu v dokumente,
* embedding,
* prípadne priradený typ segmentu (ak je známy).

Pre testovanie a vizualizáciu sa používajú grafy kosínovej podobnosti (Matplotlib), ktoré umožňujú intuitívne overenie segmentačných hraníc.

**2.2.6 Porovnanie a hodnotenie**

Plánujeme porovnať rôzne verzie výstupu:

* pravidlová segmentácia podľa kľúčových fráz,
* segmentácia na základe kosínovej podobnosti,
* modelová klasifikácia segmentov (SentenceBERT + klasifikátor).

Výsledky budú porovnané podľa presnosti (ak budú anotácie dostupné), počtu správne detegovaných hraníc, a celkovej konzistencie.

**2.3 Priebežné výsledky a experimenty**

Na overenie navrhovaného postupu sme vykonali niekoľko praktických experimentov, ktorých cieľom bolo otestovať, či sémantické rozdiely medzi vetami možno zachytiť pomocou embeddingov a či sa tieto rozdiely dajú využiť na segmentáciu právneho textu.

**2.3.1 Spracovanie dát a generovanie embeddingov**

Na spracovanie dokumentov sme použili reálnu množinu súdnych rozhodnutí vo formáte JSON, v ktorých je každý dokument reprezentovaný textom. Text bol rozdelený na vety a každá veta bola prevedená na embedding pomocou modelu kinit/slovakbert-sts-stsb zo SentenceTransformers. Tento model bol zvolený kvôli jeho optimalizácii na výpočty podobnosti medzi vetami v slovenskom jazyku.

Vygenerované embeddingy boli následne uložené do JSON súboru, čím sme si pripravili dátovú bázu pre offline analýzu a porovnávanie segmentov.

**2.3.2 Výpočet kosínovej podobnosti**

Medzi každými dvoma po sebe idúcimi vetami v dokumente sme vypočítali **kosínovú podobnosť**, pričom výsledná hodnota (v intervale 0 až 1) reprezentovala mieru významovej podobnosti. Nízka hodnota naznačovala potenciálny tematický zlom – napríklad prechod medzi opisom skutkového stavu a právnou analýzou.

Na ilustráciu výstupov bol zvolený prvý dokument s neprázdnymi embeddingmi, kde sa medzi 50–60 vetami pozorovali výrazné poklesy podobnosti v kľúčových miestach.

**2.3.3 Identifikácia segmentačných bodov**

Na základe vypočítaných podobností sme zaviedli **prahovú hodnotu** (napr. 0.2), pod ktorou sme daný pár viet označili ako možný „zlom“. Tieto body sme vizualizovali pomocou grafov, ktoré ukazovali zníženie podobnosti medzi vetami. Takéto zníženia boli interpretované ako možné začiatky nových segmentov.

Alternatívne sme testovali **rolling-average metódu**, v ktorej sa priemerovalo 5 po sebe idúcich embeddingov a následne sa porovnávali s ďalšou vetou. Tento prístup sa ukázal ako stabilnejší voči výkyvom spôsobeným jazykovými variáciami v jednej vete.

**2.3.4 Vizualizácia výsledkov**

Výsledky sme vizualizovali pomocou knižnice Matplotlib. V grafoch boli znázornené:

* všetky kosínové podobnosti medzi vetami,
* prahová čiara pre segmentáciu,
* vyznačené body pod prahom (pravdepodobné zlomy).

Tieto vizualizácie poskytujú vizuálne overenie toho, že zmeny v texte sú detegovateľné aj čisto na základe embeddingov – bez potreby explicitných značiek v texte.

*Obr. 1: Výpočet kosínovej podobnosti na náhodnom dokumente*



*Obr. 2: Výpočet rolling-average metódy na náhodnom dokumente*

**2.3.5 Zhodnotenie**

Predbežné experimenty naznačujú, že zvolený prístup má potenciál na praktické využitie:

* pri väčšine rozhodnutí dochádza k tematickým zlomom, ktoré sú zachytiteľné v embedding priestoroch,
* kosínová podobnosť dobre odráža prechody medzi faktickou časťou a právnym zdôvodnením,
* výber vhodného prahu je kritický a závisí od typu dokumentu.

Ďalším krokom bude automatizované testovanie na väčšej vzorke dokumentov a porovnanie výstupov s anotáciami od právnikov, ak budú dostupné.

**2.4 Návrhy na ďalší postup a vylepšenia**

**2.4.1 Pridanie klasifikácie segmentov**

V aktuálnej fáze detekujeme hranice segmentov bez ich konkrétneho pomenovania. Ďalším krokom je implementácia **klasifikátora viet alebo segmentov**, ktorý každému úseku priradí typ (napr. právna analýza, výrok). To možno dosiahnuť dvoma spôsobmi:

* **Vetová klasifikácia**: každú vetu označiť a následne z týchto označení odvodiť typ segmentu.
* **Segmentová klasifikácia**: po rozdelení na segmenty ich klasifikovať ako celok (napr. pomocou mean-pooled embeddingov).

Tréning klasifikátora možno vykonať pomocou ručne anotovanej vzorky (aj niekoľko desiatok dokumentov postačí pre MVP).

**2.4.2 Experiment s inými embedding modelmi**

Hoci model SlovakBERT prináša kvalitné výsledky, je možné otestovať aj:

* **multilingválne modely** ako xlm-r, LaBSE, laser (kompatibilita so slovenským právnym jazykom),
* **veľké generatívne modely** (napr. LLaMA, Mistral-7B) pomocou promptovania: „Rozdeľ nasledujúci dokument na časti…“,
* **právne špecializované modely**, ak sa objavia v budúcnosti pre slovenčinu.

**2.4.3 Dynamická detekcia prahových hodnôt**

Pevná prahová hodnota (napr. 0.2) môže byť pre niektoré dokumenty nevhodná. Navrhujeme preto:

* použiť **statistickú metódu** – napr. detekovať odľahlé hodnoty v distribúcii podobností,
* aplikovať **adaptive thresholding** – učenie optimálneho prahu pre daný typ dokumentu,
* kombinovať **viacero podobnostných metód** (okrem kosínovej podobnosti napr. Jaccard alebo TF-IDF).

**2.4.4 Automatizované hodnotenie výsledkov**

V spolupráci s právnymi expertmi možno vytvoriť **referenčnú anotovanú množinu**, na ktorej sa budú testovať metriky ako:

* **presnosť hraníc segmentov** (ako blízko je náš zlom k manuálne označenému),
* **presnosť klasifikácie segmentov** (precision/recall/F1 pre každú triedu),
* **porovnanie s inými metódami** (napr. heuristické delenie podľa fráz).

**2.5 Zhrnutie návrhu**

Návrh riešenia kombinuje sémantickú silu transformerových modelov s pragmatickým prístupom segmentácie podľa podobnosti. Hoci výsledky zatiaľ nie sú dokonalé, ukazujú veľký potenciál pre praktické nasadenie. Kľúčom k ďalšiemu pokroku bude:

* prepojenie s anotovanými dátami,
* rozšírenie o klasifikáciu,
* a dlhodobo aj spolupráca s právnymi profesionálmi pri validácii a aplikácii.

**2.6 Možnosti budúceho rozšírenia a praktického využitia systému**

Systém na sémantickú a štrukturálnu segmentáciu súdnych rozhodnutí má široký potenciál nielen ako akademický experiment, ale aj ako súčasť reálneho právno-informačného ekosystému. V tejto časti popisujeme hlavné oblasti, v ktorých možno systém rozšíriť a prakticky využiť.

**2.6.1 Integrácia do právnických vyhľadávačov**

Najpriamejšie uplatnenie segmentácie je vo vyhľadávaní právnych dokumentov. Aktuálne právnici často hľadajú kľúčové výrazy v celom texte rozhodnutia, čo vedie k veľkému množstvu irelevantných výsledkov. Ak by boli rozhodnutia segmentované, vyhľadávanie by sa mohlo obmedziť len na konkrétne časti (napr. právna analýza, odôvodnenie), čím by sa dramaticky zvýšila relevantnosť výsledkov.

**2.6.2 Príprava šablón a generovanie metadát**

Na základe segmentácie možno identifikovať štrukturálne vzory rôznych typov rozhodnutí – napr. rozdiely medzi rozsudkami a uzneseniami, medzi civilnými a trestnými vecami. Tieto vzory môžu slúžiť ako základ pre:

* **automatickú klasifikáciu typu rozhodnutia**,
* **tvorbu šablón**, ktoré pomáhajú pri generovaní alebo overovaní nových rozhodnutí,
* **anotáciu dokumentov** pre ďalší výskum v právnej informatike.

**2.6.3 Podpora právnického vzdelávania a výskumu**

Segmentované dokumenty sú cenným zdrojom aj pre právnické vzdelávanie. Študenti môžu trénovať svoju schopnosť porozumenia rôznym častiam rozhodnutia, či analyzovať, ako súd odôvodňuje svoje rozhodnutia v rôznych prípadoch. Pre výskumníkov otvárajú segmenty možnosť:

* porovnávať právnu argumentáciu medzi podobnými prípadmi,
* sledovať vývoj rozhodovacej praxe v čase,
* identifikovať precedensy a analýzy založené na judikatúre.

**2.6.4 Automatizované sumarizovanie a generovanie výťahov**

Segmentácia predstavuje základný krok k budovaniu systémov na automatické sumarizovanie právnych dokumentov. V kombinácii s modelmi typu GPT alebo LLaMA možno vytvoriť systémy, ktoré budú schopné:

* vytvoriť stručné zhrnutie právnej analýzy a rozhodnutia,
* zodpovedať otázky položené k danému prípadu,
* zvýrazniť relevantné pasáže pre konkrétny právny problém.

**2.6.5 Export pre ďalšie NLP úlohy**

Segmentované rozhodnutia možno využiť aj ako vstup pre ďalšie NLP úlohy, napríklad:

* **Named Entity Recognition** pre právne subjekty (súdy, osoby, inštitúcie),
* **extrakciu právnych noriem** a ich použitia v rozhodnutí,
* **tematickú klasifikáciu prípadov** podľa oblasti práva.

**2.6.6 Dlhodobé vízie: inteligentný právny asistent**

V dlhodobom horizonte môže byť segmentačný systém súčasťou tzv. **inteligentného právneho asistenta**, ktorý bude schopný:

* prečítať nové rozhodnutie a rozčleniť ho,
* extrahovať kľúčové právne argumenty,
* porovnať ich s existujúcimi precedensmi,
* poskytnúť používateľovi zrozumiteľné odporúčanie alebo analýzu.

**3. Vyhodnotenie a diskusia**

**3.1 Silné stránky navrhnutého riešenia**

Navrhovaný prístup využívajúci transformerové jazykové modely pre segmentáciu súdnych rozhodnutí sa v počiatočných experimentoch ukázal ako sľubný. Medzi jeho hlavné výhody patria:

* **Jazyková flexibilita**: Použitý model SlovakBERT preukázal schopnosť dobre reprezentovať významové rozdiely medzi vetami aj v právnej doméne, čo je kľúčové pre segmentáciu.
* **Modularita systému**: Architektúra riešenia je navrhnutá tak, aby umožňovala jednoduchú výmenu komponentov (embedding modely, segmentačná logika, klasifikátor).
* **Zrozumiteľnosť výstupu**: Vizualizácie podobností medzi vetami umožňujú intuitívne porozumieť, kde sa text tematicky mení.
* **Možnosť rozšírenia**: Riešenie nie je viazané na konkrétny typ dokumentu a môže byť adaptované na iné jazykové úlohy, napríklad sumarizáciu alebo tematické zoskupovanie.

**3.2 Obmedzenia a výzvy**

Napriek sľubným výsledkom je potrebné otvorene pomenovať aj slabšie miesta súčasného riešenia:

* **Chýbajúce anotované dáta**: Bez ručne označených segmentov je objektívne vyhodnotenie presnosti segmentácie veľmi obmedzené. Subjektívna kvalita výstupu je síce povzbudzujúca, ale nestačí pre vedeckú validáciu.
* **Citlivosť na prah**: Nastavenie prahovej hodnoty pre zlom je zatiaľ manuálne a závislé od typu dokumentu. Automatické prahovanie alebo učenie adaptívnej hranice by bolo presnejšie a robustnejšie.
* **Riziko pretrénovania**: V prípade neskoršej klasifikácie viet bude potrebné dbať na generalizovateľnosť modelu. Malý počet anotácií môže viesť k nadmernému prispôsobeniu konkrétnym vzorom v trénovacej množine.
* **Výpočtová náročnosť**: Použitie transformerových modelov vyžaduje výrazné výpočtové prostriedky (najmä pri spracovaní veľkých datasetov), čo obmedzuje real-time nasadenie bez optimalizácie.

**3.3 Diskusia o alternatívnych prístupoch**

Počas navrhovania riešenia sme uvažovali aj o iných možnostiach:

* **Heuristický prístup** založený na kľúčových frázach (napr. „súd rozhodol“, „na základe vykonaného dokazovania“) je rýchly, ale menej presný a zle zvláda jazykovú variabilitu.
* **Klasické ML modely** ako SVM či Random Forest na ručne vytvorených črtách boli zavrhnuté pre nižšiu schopnosť pracovať s významom viet.
* **Generatívne modely** (napr. GPT, Mistral) predstavujú zaujímavú možnosť do budúcnosti, no ich nasadenie v slovenskom právnom jazyku je zatiaľ obmedzené.

Výber sémantických embeddingov ako základného nástroja segmentácie sa teda ukazuje ako **kompromis medzi interpretovateľnosťou, výkonom a rozšíriteľnosťou**.

**3.4 Reflexia dosiahnutých cieľov**

Súčasný stav projektu napĺňa viaceré z cieľov zadania:

Do finálnej fázy projektu ostáva najmä zamerať sa na:

* vytvorenie testovacej množiny s anotovanými rozhodnutiami,
* zapojenie právnikov pre spätnú väzbu a ručnú validáciu,
* optimalizáciu klasifikácie segmentov a ich pomenovanie.

**3.5 Možnosti segmentácie dokumentov pomocou embeddingov**

Využitie jazykových embeddingov pri spracovaní textu je dnes jednou z najefektívnejších techník na zachytenie významu vetných alebo odsekových jednotiek. Embeddingy umožňujú previesť text na numerické reprezentácie, ktoré zachytávajú sémantickú podobnosť medzi rôznymi vetami. V tejto časti sa podrobne venujeme tomu, ako možno embeddingy využiť na rozdelenie súdnych rozhodnutí do zmysluplných segmentov.

**3.5.1 Čo sú embeddingy?**

**Embeddingy** sú vektorové reprezentácie textových jednotiek (slov, viet, odsekov, dokumentov), ktoré zachytávajú význam na základe kontextu, v ktorom sa slová vyskytujú. Sú generované pomocou jazykových modelov, ako napr.:

* **Word embeddings**: napr. Word2Vec, GloVe – fungujú na úrovni slov.
* **Contextual embeddings**: BERT, RoBERTa, SlovakBERT – berú do úvahy kontext vety.
* **Sentence embeddings**: Sentence-BERT, SlovakSentenceBERT – reprezentujú celé vety alebo odseky ako jediné vektory.

V našom prípade využívame model kinit/slovakbert-sts-stsb, ktorý je trénovaný špeciálne na výpočet významovej podobnosti medzi vetami v slovenčine.

**3.5.2 Základná technika: podobnosť medzi po sebe idúcimi embeddingmi**

Najpriamejší spôsob využitia embeddingov je výpočet **kosínovej podobnosti** medzi embeddingmi po sebe nasledujúcich viet:

* Vysoká podobnosť (blízka 1) → vety sú tematicky súvisiace.
* Nízka podobnosť (napr. < 0.2) → možný prechod do novej časti textu.

Týmto spôsobom vieme identifikovať „zlomy“ v sémantickej kontinualite a použiť ich ako segmentačné body.

Možné varianty:

* **Pevný prah**: manuálne zvolená hodnota podobnosti (napr. 0.2).
* **Percentilový prah**: prah založený na spodných 5 % podobností v dokumente.
* **Zmena rozdielu (derivácia)**: nie priamo hodnota podobnosti, ale jej zmena (napr. prudký pokles signalizuje hranicu).

**3.5.3 Priemerovanie embeddingov (sliding window)**

Ak je dokument dlhý, porovnávanie len medzi dvoma po sebe idúcimi vetami môže byť príliš citlivé. Namiesto toho môžeme:

* **Vypočítať priemer z viacerých embeddingov** (napr. z posledných 5 viet),
* Porovnať tento priemer s embeddingom nasledujúcej vety,
* Nízka podobnosť → pravdepodobný zlom.

Tento postup je robustnejší a potláča náhodné fluktuácie medzi vetami.

**3.5.4 Klastrovanie embeddingov**

Embeddingy všetkých viet môžeme skupiť do **klastrov** podľa ich významovej podobnosti pomocou algoritmov ako k-means alebo DBSCAN. Klastre následne reprezentujú rôzne časti dokumentu:

* Skupina viet v rovnakom klastri → rovnaká funkcia (napr. opis skutkového stavu).
* Prechod medzi klastrami → segmentačný bod.

Alternatívne: klastre môžeme následne **označiť** manuálne alebo modelom (napr. „úvod“, „záver“...).

Výzva: počet klastrov sa musí zvoliť rozumne (alebo odhadnúť automaticky).

**3.5.5 Vizuálna analýza embeddingov**

Embeddingy je možné **znížiť na 2D alebo 3D priestor** (napr. PCA, t-SNE, UMAP) a vizuálne zobraziť ich rozloženie v priestore.

* Vety patriace do rovnakého segmentu by mali tvoriť zhluky.
* Zhluky môžu indikovať štrukturálne časti dokumentu.

To slúži ako doplnková diagnostika – vhodná najmä na manuálne overenie alebo prezentáciu výsledkov.

**3.5.6 Segmentácia založená na zhlukovaní zmien (changepoint detection)**

Moderné algoritmy detekcie zmien v časových radoch môžu byť aplikované aj na sekvenciu podobností medzi embeddingmi. Medzi najčastejšie patria:

* **Ruptures** – detekcia hraníc podľa zmien v distribúcii dát,
* **Bayesian changepoint detection** – pravdepodobnostné modelovanie zmien.

Výhodou je, že systém **sám rozhodne**, kde sú najpravdepodobnejšie zmeny tém v texte – bez nutnosti nastavovania prahu.

**3.5.7 Kombinované metódy**

Najlepšie výsledky možno dosiahnuť kombináciou viacerých prístupov:

* Prvotná detekcia zlomov pomocou podobností.
* Následné overenie pomocou klasifikátora segmentov.
* Vizualizácia + heuristiky na finálne doladenie.

Takýto **viacstupňový prístup** je robustnejší, prenosnejší medzi dokumentmi a ľahšie adaptovateľný na zmeny v štýle súdneho písania.

**4. Záver**

Spracovanie súdnych rozhodnutí predstavuje náročnú, ale súčasne mimoriadne dôležitú oblasť v aplikácii spracovania prirodzeného jazyka (NLP) na špecifickú právnu doménu. Táto práca sa zameriavala na segmentáciu týchto dokumentov – t. j. rozdelenie dlhého, formálne neštruktúrovaného textu na sémanticky ucelené časti, ako je úvod, skutkový stav, právna analýza či výrok.

Cieľom práce bolo navrhnúť a experimentálne overiť prístup, ktorý využíva moderné jazykové modely – konkrétne embeddingy generované modelom SlovakBERT – na identifikáciu tematických zlomov v texte. Na rozdiel od klasických heuristických metód je náš prístup založený na sémantickej podobnosti medzi vetami, čo umožňuje identifikovať štruktúru dokumentu aj bez explicitných značiek alebo nadpisov.

V práci sme:

* analyzovali štruktúru slovenských súdnych rozhodnutí,
* implementovali prototyp systému, ktorý spracováva dokumenty, generuje vetové embeddingy, počíta kosínové podobnosti a vizualizuje zmeny v texte,
* identifikovali segmentačné hranice pomocou rôznych metód: pevného prahu, priemerovania embeddingov (sliding window), ako aj možnosti využitia klastrovania a changepoint detection algoritmov,
* pripravili návrh ďalšieho postupu vrátane klasifikácie segmentov, tvorby anotovaného datasetu

**Hlavným prínosom práce je návrh prakticky využiteľnej metódy segmentácie právneho textu, ktorá je jazykovo a štruktúrne prispôsobená slovenským podmienkam.** Taktiež sme demonštrovali, že aj bez rozsiahlych anotácií možno dosiahnuť zmysluplné a interpretovateľné výsledky.

**Možnosti pokračovania**

Túto prácu možno rozvíjať viacerými smermi:

* **Anotácia dát**: Vytvoriť ručne anotovanú množinu dokumentov, ktorá umožní trénovanie klasifikátorov segmentov a objektívne vyhodnotenie presnosti.
* **Vetová klasifikácia**: Na základe embeddingov trénovať klasifikačný model, ktorý rozpozná typ každej vety alebo segmentu.
* **Experimentovanie s inými modelmi**: Vyskúšať väčšie jazykové modely (napr. Mistral, LLaMA) a ďalšie embeddingové architektúry.
* **Zavedenie dynamických prahov a pokročilých detektorov zmien**: Vylepšiť presnosť identifikácie hraníc medzi časťami dokumentu.
* **GUI nástroj pre spätnú väzbu právnikov**: Umožniť manuálne označovanie segmentov, ktoré môže slúžiť ako základ pre ďalší výskum a zber tréningových dát.

**Záverečná myšlienka**

Segmentácia právnych textov nie je len technická úloha – je to krok smerom k sprístupneniu právnych informácií širokej odbornej verejnosti, zvýšeniu transparentnosti súdnych rozhodnutí a lepšiemu využitiu právnej praxe. Táto práca ukazuje, že s využitím moderných nástrojov umelej inteligencie možno robiť praktické kroky aj v doméne, kde sú štruktúra a jazyk mimoriadne náročné.