Využitie veľkých jazykových modelov na efektívnu analýzu náboženských textov

Bibiána Lajčinová[1] Jozef Žuffa[2] Milan Urbančok[2]

[1]Národné superpočítačové centrum, Dúbravská cesta 9, 845 35 Bratislava 45, Slovenská republika
bibiana.lajcinova@nscc.sk
[2]Teologická fakulta Trnavskej univerzity v Trnave, Kostolná 1, 814 99 Bratislava, Slovenská republika
jozef.zuffa@truni.sk, milan.urbancok@truni.sk

Abstrakt

Štúdia sa zaoberá úlohou získavania informácií (z angl. *information retrieval*) z dokumentov s náboženskými témami v slovenskom jazyku pomocou embeddingových modelov, s cieľom zrýchlenia a zefektívnenia analýzy týchto textov pre odborníkov v danej oblasti. Pomocou verejne dostupných embeddingových modelov *Slovak-BERT* a *BGE M3* a proprietárneho modelu *text-embedding-3-small* od OpenAI sme generovali embeddingové indexy z textových blokov dostupných dát a vyhodnocovali metriku recall naprieč piatimi rôznymi témami pomocou testovacích otázok. Skúmali sme tiež rôzne techniky predspracovania, ako je kontextová augmentácia testovacích otázok a odstraňovanie vyradených slov (z angl. *stopwords*). Výsledky naznačujú, že táto metodológia môže byť užitočná pre zefektívnenie výskumu náboženských textov a môže pomôcť odhaliť skryté interpretácie a významy obsiahnuté v týchto textoch. Naše zistenia tiež zdôrazňujú dôležitosť výberu vhodnej techniky predspracovania pre daný model a dáta.

1 Úvod

Analýza a štúdium textov s náboženskými témami boli historicky doménou filozofov, teológov a ďalších špecialistov v sociálnych vedách. S príchodom umelej inteligencie, konkrétne veľkých jazykových modelov, naberá výskum v tejto oblasti nové dimenzie. Tieto moderné technológie môžu byť využité na odhaľovanie skrytých nuáns v náboženských textoch, čím umožňujú hlbšie pochopenie rôznych symbolizmov a odhalenie významov, ktoré sú pre tieto texty charakteristické a môžu byť nejasné. Takéto zrýchlenie analytického procesu umožňuje výskumníkom sústrediť sa len na špecifické aspekty textu relevantné pre ich výskum. Jednou z úloh, ktorou sa vedci v tejto oblasti zaoberajú, je štúdium diel autorov asociovaných so špecifickými náboženskými skupinami a komunitami. Porovnávaním ich textov s oficiálnymi doktrínami a učeniami ich denominácií môžu výskumníci hlbšie nahliadnuť do presvedčení, viery a uhlov pohľadu komunít, formovaných učeniami vplyvných autorov.

Štúdia sumarizuje prístup využívajúci embeddingové indexy a jazykové modely na efektívnu analýzu textov s náboženskými témami. Primárnym cieľom je vyvinúť nástroj na získavanie informácií, špeciálne navrhnutý na účinné lokalizovanie relevantných častí textu v dokumentoch. Identifikácia nesúladov medzi získanými časťami textu z diel náboženských komunít a oficiálnymi náukami daného náboženstva, z ktorého táto komunita pochádza, nie je cieľom tejto práce a je ponechaná na teológov.

Táto práca vznikla spojeným úsilím Národného superpočítačového centra a Teologickej fakulty Trnavskej univerzity. Na dosiahnutie riešenia, ktoré vyžadovalo numericky náročné spracovanie veľkého objemu dát, boli využité výpočtové zdroje HPC systému Devana.

2 Dáta

Texty analyzované v tejto štúdii pochádzajú z náboženskej komunity známej ako Hnutie Nazaret (bežne nazývanej aj "Beňovci"), ktorá sa začala formovať v sedemdesiatych rokoch minulého storočia. Hnutie, o ktorom niektorí vedci hovoria, že má známky sekty, je stále aktívne aj v dnešnej dobe, avšak v redukovanej a zmenenej forme. Jeho zakladateľ, Ján Augustín Beňo (1921 - 2006), bol tajne vysväteným katolíckym kňazom v totalitnej dobe. Beňo nabádal členov hnutia k aktívnemu žitiu

viery prostredníctvom každodenného čítania biblických textov a uplatňovania ich posolstva v praxi prostredníctvom konkrétnych rozhodnutí a činov. Hnutie sa rozšírilo po celom Slovensku, pričom komunity existovali takmer v každom väčšom meste. Rozšírilo sa aj do susedných krajín, ako Poľsko, Česká republika, Ukrajina a Maďarsko. V roku 2000 bolo v hnutí približne tristo manželských párov, tisíc deti a stotridsať kňazov a študentov pripravujúcich sa na kňazstvo. Hnutie malo tri hlavné ciele: radikálnu prevenciu v oblasti vzdelania, podporu kňazov, ktorí by mohli pôsobiť ako rodičovské postavy na identifikáciu a rozvoj kňazských povolaní u detí, a výrobu a distribúciu samizdatových materiálov potrebných na katechézu a evanjelizáciu.

Pre výskum bolo k dispozícii 27 dokumentov vytvorených touto komunitou. Tieto dokumenty, ktoré významne vplývali na formovanie ideológie komunity Beňovci, boli reprodukované a distribuované počas komunistického režimu vo forme samizdatov - literatúry zakázanej komunistickým režimom. Po politickom prevrate boli viaceré z týchto dokumentov vytlačené a distribuované verejnosti mimo hnutia. Väčšina z dokumentov pozostávala z textov určených pre "ranné úvahy" — krátke meditácie nad biblickými textami. Dokumenty taktiež obsahovali zakladateľove komentáre o učeniach Katolíckej cirkvi a vybraných témach týkajúcich sa výchovy detí, spirituálneho vedenia a katechézy pre deti.

Dokumenty obsahovali niekoľko duplicít, avšak pre úlohu získavania a vyhľadávania informácií to nepredstavuje problém. Všetky dokumenty sú napísané výhradne v slovenskom jazyku.

Jeden z dokumentov bol anotovaný pre účely testovania expertom z partnerskej fakulty, ktorý sa dlhodobo venuje Hnutiu Nazaret. Anotáciami myslíme časti textu (zvyčajne odseky, prípadne vety) označené ako patriace do jednej z piatich tried, pričom tieto triedy reprezentujú päť tém:

1. Direktívna poslušnosť
2. Hierarchická výchova
3. Radikálnosť v prevzatí modelu života
4. Ľudské potreby realizované len v spoločenstve/hnutí a v rodine
5. Divné/čudné/silné

Každá z týchto tém je doplnená o súbor otázok (dopytov/výrazov), ktoré boli navrhnuté na testovanie riešenia získavania informácií. Cieľom týchto testovacích otázok je vyhodnotiť, koľko relevantných častí textu týkajúcich sa danej témy dokáže náš nástroj získať z anotovaného dokumentu.

Text	**Anotácia**
Veď ak milujeme svojho Boha, ako si to myslíme, alebo aj hovoríme, nemôže nám byť ľahostajný nijaký odklon od jeho svätej vôle.	Direktívna poslušnosť

Tabuľka 1: Príklad anotovaného textu patriaceho do triedy Direktívna poslušnosť.

3 Postup riešenia

Existuje viacero metód vhodných na riešenie tejto úlohy, vrátane klasifikácie textu, modelovania témy textu, RAG (z angl. *Retrieval-Augmented Generation*), alebo optimalizácie predtrénovaného jazykového modelu. Avšak, použiadavkou partnerských teológov, zaoberajúci sa analýzou týchto dokumentov, bola identifikácia konkrétných častí textu relevantných k daným témam, a teda získanie ich presného znenia. Práve preto bola vybraná metóda získavania informácií (z angl. *information retrieval*). Tento prístup sa líši od metódy RAG, ktorá bežne obsahuje časť získavania informácií a tiež časť generovania nového textu, v tom, že sa sústredí výhradne na identifikáciu relevantných častí textu v dokumentoch a negeneruje žiadny nový text.

Metóda získavania informácií využíva jazykové modely na transformovanie komplexných dát, ako je text, do numerickej reprezentácie, ktorá zachytáva celý význam a kontext daného vstupu. Táto numerická reprezentácia, nazývaná embedding (vo zvyšku textu budeme kvôli jednoduchosti využívať už len tento termín), môže byť použitá na sémantické vyhľadávanie v dokumentoch analyzovaním pozícií a blízkosti embeddingov v multidimenzionálnom vektorovom priestore. Použitím otázok (dopytov) dokáže systém nájsť v dokumentoch relevantné časti textu meraním podobnosti medzi embeddingami

otázok a embeddingami segmentovaného textu. Tento prístup nevyžaduje žiadnu optimalizáciu existujúceho jazykového modelu, takže modely môžu byť použité bez akýchkoľvek úprav a pracovaný postup zostáva pomerne jednoduchý.

3.1 Výber modelu

Keďže všetky analyzované dokumenty v rámci tejto štúdie sú v slovenskom jazyku, je potrebné, aby zvolený jazykový model "rozumel" slovenčine, čo značne zúžilo možnosti jeho výberu. K dnešnému dňu existuje len jeden verejne dostupný model, ktorý rozumie výhradne slovenskému jazyku, a niekoľko multilingválnych modelov, ktoré rozumejú slovenčine do určitej miery. Štyri predtrénované modely boli vybrané z malého množstva dostupných možností, prvým z nich je model *Slovak-BERT* [1]. *Slovak-BERT* je verejne dostupný model založený na architektúre transformerov. Ďalším vybraným modelom je *text-embedding-3-small* model. Ide o výkonný proprietárny embedding model dostupným len cez API spoločnosti OpenAI. Tretím modelom je verejne dostupný embedding model *BGE M3* [2], ktorý je výkonným multilingválnym modelom podporujúcim viac než 100 jazykov. Posledným modelom je taktiež multilingválny model z dielne Microsoftu nazývaný *E5* [3], ktorý je rovnako verejne dostupný.

Tieto štyri modely boli použité na získanie vektorových reprezentácií textu. Ich výkon bude detailne diskutovaný v nasledujúcich častiach reportu.

3.2 Predspracovanie dát

Prvým krokom predspracovania dát je segmentovanie textu (z angl. *chunking*). Hlavným dôvodom pre tento krok bolo splniť požiadavku teológov na vyhľadávanie (získavanie) krátkych častí textu. Okrem toho bolo potrebné dokumenty rozdeliť na menšie časti, aj kvôli obmedzenej dĺžke vstupu niektorých jazykových modelov. Na túto úlohu bola použitá knižnica *Langchain* [4]. Poskytuje hierarchické segmentovanie textu, ktoré produkuje prekrývajúce sa bloky textu definovanej dĺžky (s definovaným prekrytím) tak, aby v nich bol zachovaný kontext. Takto boli vytvorené bloky s dĺžkami 300, 400, 500 a 700 znakov. Následne spracovanie pozostávalo z odstránenia diakritiky, úprava textu na veľké/malé písmená, podľa podmienok modelov a odstránenie vylúčených slov (z angl. *stopwords*). Odstraňovanie týchto slov je bežnou praxou v úlohách spracovania prirodzeného jazyka, keďže vylúčené slová nenesú žiadnu významovú informáciu. Niektoré modely môžu profitovať z odstránenia vylúčených slov na zlepšenie relevantnosti získaných blokov textu, ale iné môžu ťažiť z ponechania týchto slov, aby bol zachovaný celý kontext nevyhnutný na pochopenie textu.

Index	Blok textu
8	Podľa tohoročnej sa rozvádza už každé tretie. Tento bolestný spoločenský jav vysvetľujú niektorí skutočnosťou, že dnešní manželia sú náročnejší a od svojho manželstva viac očakávajú než tí, čo žili pred nami. Že by to bola pravda? Od manželstva môže človek čakať, len toľko, koľko doň vloží.
9	Že by to bola pravda? Od manželstva môže človek čakať, len toľko, koľko doň vloží. Ak minulosť týmto bláznovstvom rozvodovosti netrpela, tak zaiste preto, že v manželstve nevidela iba tú sentimentálnu príjemnú lásku, ale aj tú obetavú, živú z viery v Boha a z poslušnosti voči Cirkvi. Zaujímajú nás začiatky, priebeh a dôsledky rozvodov? Nie je ťažko spoznať ich.

Tabuľka 2: Príklad dvoch blokov textu s prekrytím.

3.3 Vektorové embeddingy

Vektorové embeddingy boli vytvorené z blokov textu s použitím vybraných predtrénovaných jazykových modelov.

V prípade modelu *Slovak-BERT*, sme pre generovanie embeddingov použili model bez pridaných predikčných vrstiev, a následne sme ukladali iba prvý embedding, ktorý obsahuje celý význam vstupného textu. Ďalšie používané modely priamo produkujú embeddingy vo vhodnej forme, preto nebolo potrebné žiadne dodatočné spracovanie výstupov.

V nasledujúcej časti s výsledkami analyzujeme výkon všetkých vybraných embedding modelov a porovnávame ich schopnosti zachytiť kontext.

4 Výsledky

Pred uskutočnením kvantitatívnych testov prešli všetky embeddingové indexy predbežným hodnotením, aby sa zistila úroveň porozumenia slovenského jazyka a špecifickej náboženskej terminológie evaluovaných modelov. Predbežné hodnotenie zahŕňalo subjektívne posúdenie relevantnosti získaných častí textu.

Tieto testy odhalili, že embeddingy získané pomocou modelu *E5* nie sú dostatočne efektívne pre naše dáta. Keď sme pomocou testovacej otázky hľadali informácie v dokumentoch, väčšina získaných blokov textu obsahovala kľúčové slová použité v otázke, ale neobsahovala kontext otázky. Možným vysvetlením by mohlo byť, že tento model uprednostňuje zhody na úrovni slov pred zhodami kontextu v slovenskom jazyku. Ďalším dôvodom môže byť aj to, že tento model bol natrénovaný na dátach, ktoré neobsahovali veľké množstvo textu v slovenčine, resp. výber textov nebol dostatočne rozmanitý, čo môže viesť k nižšiemu výkonu modelu *E5* v slovenčine, aj keď v iných jazykoch dosahuje výborné výsledky. Podotýkame, že tieto pozorovania nie sú definitívne závery, ale skôr hypotézy založené na súčasných, obmedzených výsledkoch. Rozhodli sme sa ďalej nevyhodnocovať výkon embeddingových indexov získaných z *E5* modelu, keďže je to irelevantné vzhľadom na neschopnosť modelu zachytiť nuansy náboženského textu.
Na druhej strane, schopnosti modelu *Slovak-BERT*, ktorý je založený na architektúre RoBERTa charakteristickej jej relatívne jednoduchou topológiou, prekonali očakávania. Navyše, výkon *text-embedding-3-small* a *BGE M3* embeddingov splnil očakávania, keďže prvý, subjektívne vyhodnotený, test ukázal veľmi dobré porozumenie kontextu a nuáns v textoch s náboženskými témami a taktiež výborné porozumenie slovenského jazyka.

Preto boli kvantitatívne testy vykonané len pre vektorové databázy využívajúce *Slovak-BERT*, OpenAI *text-embedding-3-small* a *BGE M3* embeddingy.

Vzhľadom na povahu riešeného problému a charakter testovacích anotácií existuje potenciálna obava týkajúca sa ich kvality. Niektoré časti textu mohli byť nesprávne klasifikované, pretože môžu patriť do viacerých tried. Táto skutočnosť, spolu s možnosťou ľudskej chyby, mohla ovplyvniť konzistentnosť a presnosť anotácií.

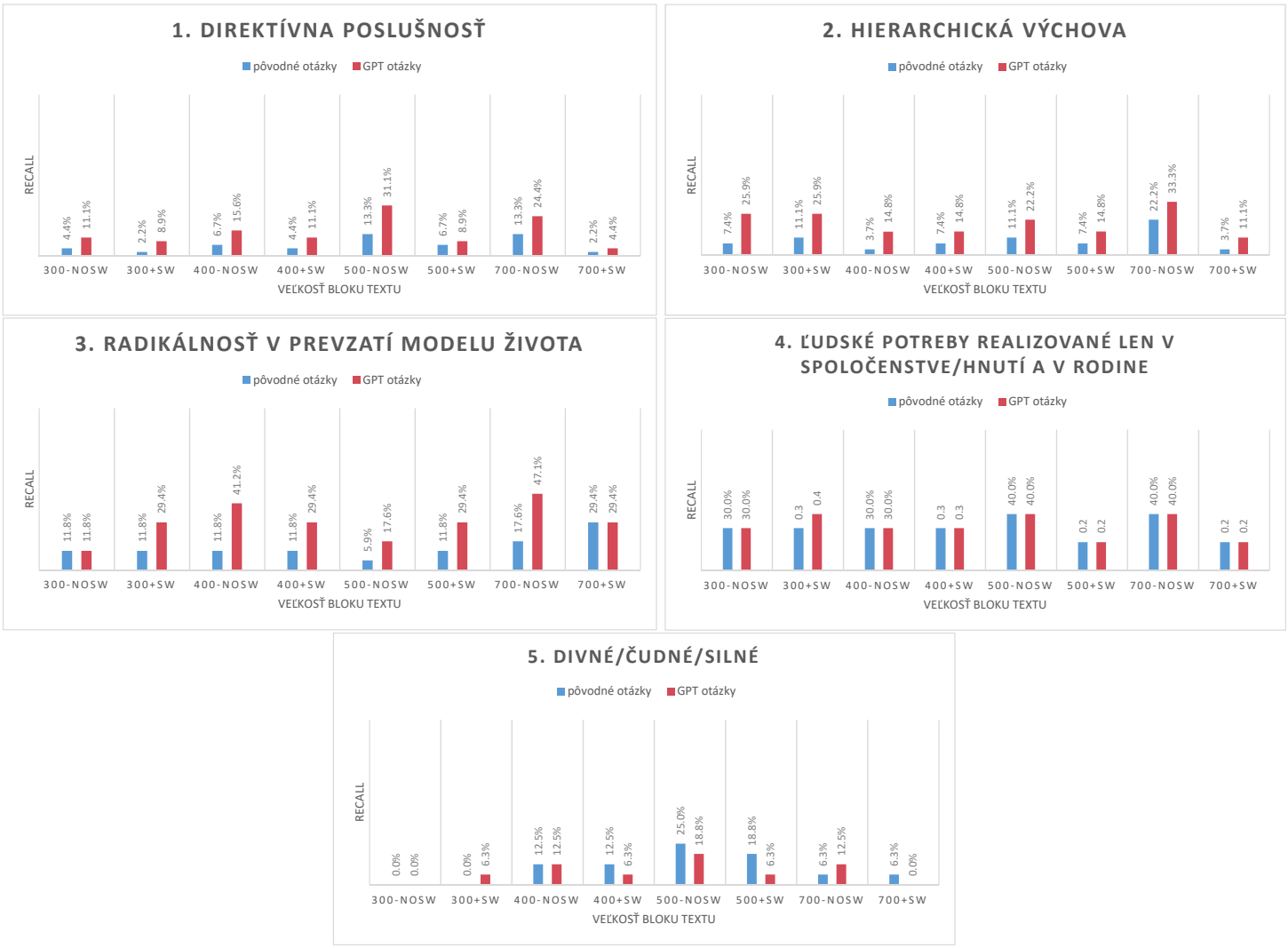
Berúc do úvahy túto skutočnosť, sme sa rozhodli zamerať výhradne na vyhodnotenie metriky zvanej *recall*. Hodnotu tejto metriky meriame ako pomer počtu získaných blokov zhodných s anotáciami, k celkovému počtu anotovaných blokov textu (bez ohľadu na podiel falošne pozitívnych blokov). Recall vyhodnocujeme pre každú tému a pre všetky vektorové databázy s rôznymi dĺžkami blokov textu.

Komplexnosť a interpretačná povaha náboženských štúdií sa pravdepodobne prejavuje nielen v kvalite testovacích anotácií, ale aj v samotných testovacích otázkach. Ako príklad môžme uviesť testovaciu otázku "Božia vôľa" pre tému *Direktívna poslušnosť*. Hoci pozorný čitateľ rozumie, ako táto otázka súvisí s danou témou, nemusí to byť očividné pre jazykový model. Preto, okrem vyhodnotenia pomocou dodaných testovacích otázok budeme vyhodnocovať výkon embeddingov aj s použitím ďalších otázok, ktoré boli získané metódou kontextovej augmentácie.
Kontextová augmentácia je technika v prompt inžinieringu používaná na zlepšenie kvality textových dát a je dokumentovaná vo viacerých vedeckých článkoch [5], [6]. Táto technika spočíva v tom, že sa zvolený jazykový model použije na vytvorenie novej otázky (príp. nového textu) na základe pôvodnej otázky (textu) a doplneného kontexu s cieľom formulovania lepšej otázky. Jazykový model použitý na generovanie nových otázok pomocou tejto techniky bol *GPT 3.5* a tieto otázky budeme ďalej v texte označovať ako "GPT otázky".

4.1 Slovak-BERT embeddingové indexy

Vyhodnotenie metriky recall pre embeddingové indexy využívajúce *Slovak-BERT* embeddingy pre štyri rôzne veľkosti blokov textu s použitím a bez použitia metódy odstraňovania vylúčených slov je zobrazené na Obrázku 1. Toto vyhodnotenie zahŕňa každú z piatich tém špecifikovaných v Časti 2 a pokrýva pôvodné aj GPT otázky.

Je očividné, že GPT otázky produkujú vo všeobecnosti lepšie výsledky než pôvodné otázky, okrem prípadu posledných dvoch tém, pri ktorých obe sady otázok produkujú podobné výsledky. Je tiež zrejmé, že *Slovak-BERT* embeddingy vo väčšine prípadov profitujú z odstránenia vylúčených slov. Najvyššia hodnota recall bola dosiahnutá pre tému *Radikálnosť v prevzatí modelu života*, s veľkosťou blokov textu 700 znakov, s odstránenými vylúčenými slovami, dosahujúc viac než 47%. Na druhej strane, najhoršie výsledky boli získané pre tému *Divné/čudné/silné*, kde ani jedna sada otázok nedokázala úspešne získať relevantné častí textu z dokumentov. Dokonca, v niektorých prípadoch neboli získane absolútne žiadne relevantné bloky textov.



Obr. 1: Hodnoty recall pre všetky témy vyhodnotené pomocou pôvodných aj GPT otázok (pre všetky skúmané veľkosti blokov textu) pre embeddingy generované modelom *Slovak-BERT*. Indexy embeddingov označené ako +SW obsahujú vylúčené slová, zatiaľ čo -NoSW znamená, že vylúčené slová boli odstránené.

4.2 OpenAI text-embedding-3-small embeddingové indexy

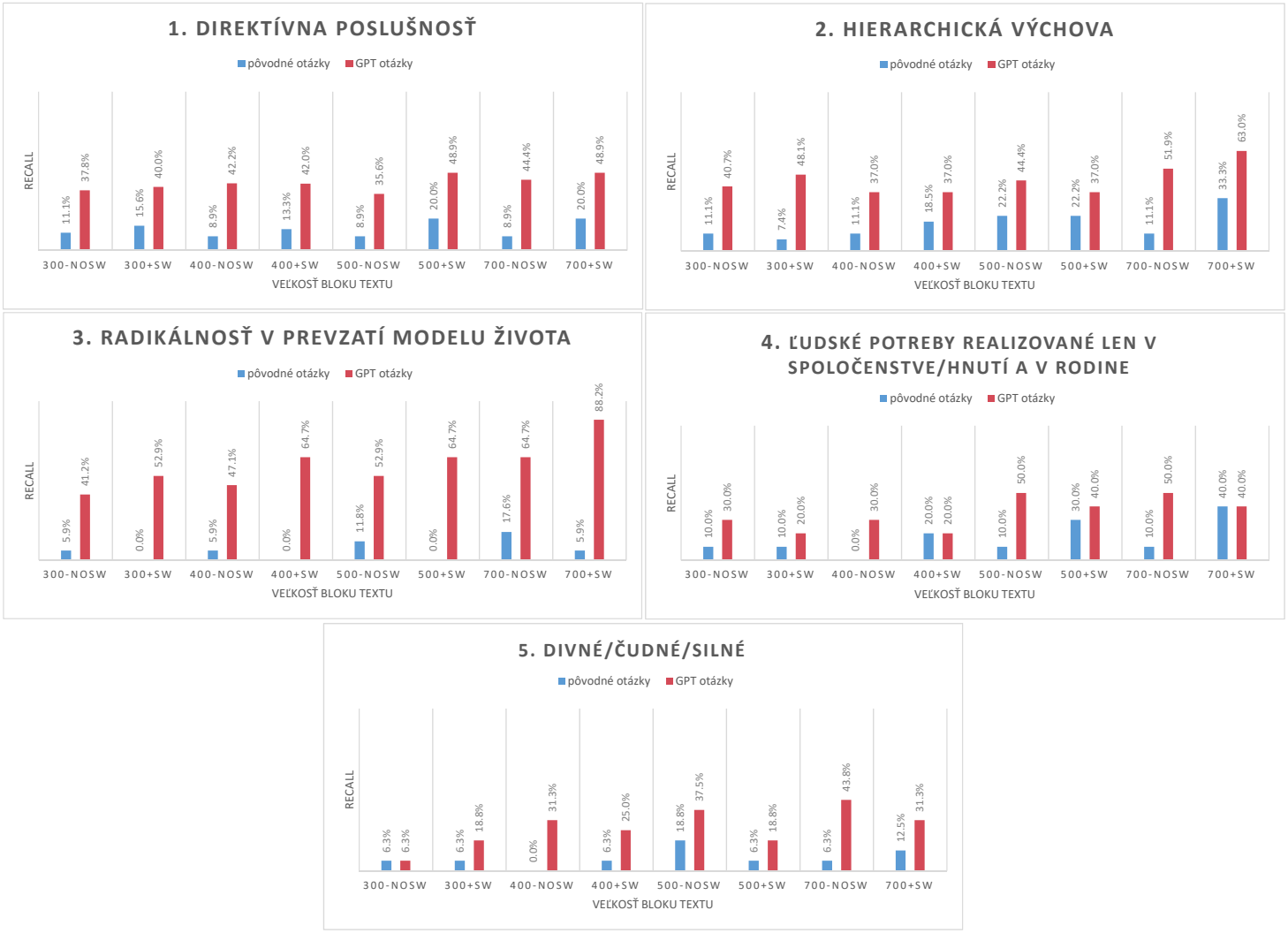
Analogicky k vyhodnoteniu *Slovak-BERT* embeddingových indexov, grafy s výsledkami pre embeddingy získané modelom *text-embedding-3-small* sú zobrazené na Obrázku 2. Hodnoty metriky recall sú všeobecne vyššie než tie získané so *Slovak-BERT* embeddingami. Podobne ako v predchádzajúcom prípade, GPT otázky produkujú lepšie výsledky. Pozorovateľný je taktiež istý trend medzi hodnotou metriky recall a veľkosťou textových blokov — dlhšie bloky textu zvyčajne vykazujú vyššie hodnoty recall.

Zaujímavé zistenie sa týka témy *Radikálnosť v prevzatí modelu života*. S použitím pôvodných otázok sme nezískali takmer žiadne relevantné výsledky. Naopak, pri použití otázok generovaných pomocou GPT modelu, boli hodnoty recall metriky výrazne vyššie a dosahovali takmer 90% pre bloky textu s veľkosťou 700 znakov.

Čo sa týka odstraňovania vyradených slov, vplyv tejto techniky na embeddingy sa líši. Pre témy 4 a 5 sa ukazuje, že odstránenie vyradených slov je prospešné. Avšak, pre ostatné témy tento krok výhody neprináša.

Témy 4 a 5 vykazovali najslabšie výsledky medzi všetkými témami. Môže to byť spôsobené povahou otázok pre tieto dve témy, keďže sú to citáty a celé vety, na rozdiel od otázok pre ostatné témy, ktoré sú frázy, kľúčové slová alebo výrazy. Zdá sa, že model *text-embedding-3-small* funguje lepšie s frázovitým typom otázok. Ale na druhej strane, keďže otázky pre témy 4 a 5 sú celé vety, zdá sa embeddingy profitujú z odstránenia vyradených slov, keďže v tomto prípade to môže pomôcť pri zachytení kontextu v dlhých otázkach.

Téma 4 je veľmi špecifická a preto možno vyžaduje detailnejšie testovacie otázky, keďže poskytnuté otázky pravdepodobne neobsahujú všetky nuansy danej témy. Naopak, téma 5 je veľmi všeobecná, vďaka čomu je celkom pochopiteľné, prečo je zachytávanie kontextu tejto témy pomocou embeddingov náročné. Všeobecný charakter tejto témy by mohol profitovať z iného analytického prístupu. Napríklad metóda analýzy sentimentu by mohla zachytiť zvláštnu, čudnú a silnú náladu vo vzťahu k študovaným náboženským témam.



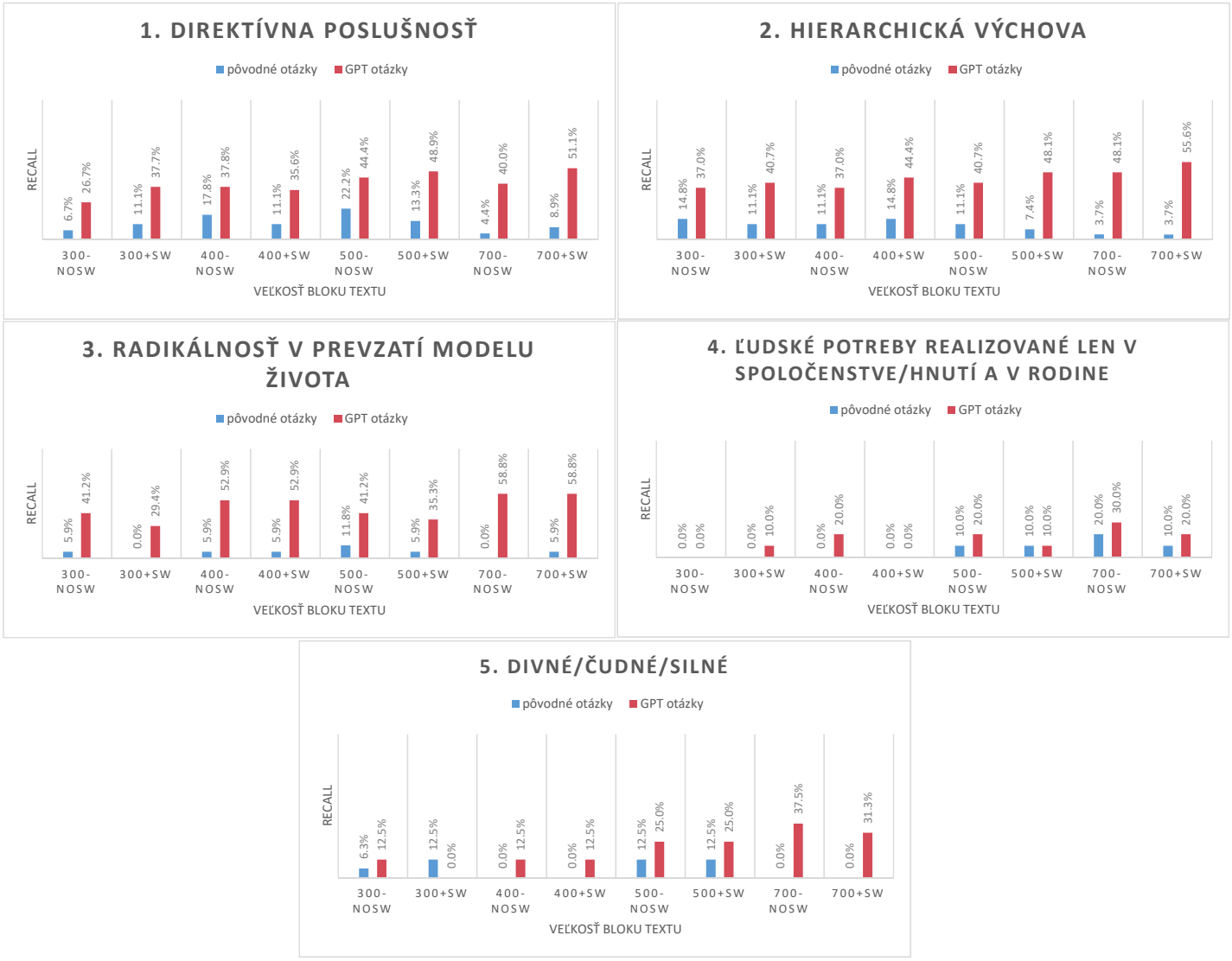
Obr. 2: Hodnoty recall vyhodnotené pre všetky témy pomocou pôvodných aj GPT otázok, pre všetky embeddingové indexy generované modelom *text-embedding-3-small*. Embeddingové indexy označené +SW obsahujú vylúčené slová, zatiaľ čo indexy označené -NoSW majú vylúčené slová odstránené.

4.3 BGE M3 embeddingové indexy

Grafy s vyhodnotenou metrikou recall pre embeddingové indexy využívajúce *BGE M3* embeddingy sú zobrazené na Obrázku 3. Tieto hodnoty ukazujú výkon spadajúci medzi *Slovak-BERT* a OpenAI *text-embedding-3-small* embeddingy. V niektorých prípadoch sa nepodarilo dosiahnuť také vysoké hodnoty metriky recall ako pri OpenAI embeddingoch, avšak *BGE M3* embeddingy stále vykazujú konkurencieschopný výkon, hlavne ak prihliadneme na skutočnosť, že sa jedná o verejne dostupný model, na rozdiel od OpenAI embeddingového modelu, ku ktorému sa dá pristupovať len cez API, čo môže byť niekedy problémom kvôli zdieľaniu súkromných alebo citlivých dát a taktiež kvôli finančným nákladom.

S týmito embeddingami môžeme pozorovať rovnaký fenomén ako s *text-embedding-3-small* embeddingami: krátke, frázovité otázky sú preferované pred dlhšími otázkami podávanými formou viet a citátov. Preto sú hodnoty recall pre prvé tri témy vyššie, ako sme diskutovali už v predchádzajúcej časti.

Odstránenie vylúčených slov sa zdá byť užitočné, hlavne pre posledné dve témy.



Obr. 3: Hodnoty metriky recall pre všetky témy získané s použitím pôvodných aj GPT otázok pre embeddingy vytovrené modelom *BGE M3*. Značky +SW označujú indexy obsahujúce vylúčené slová, zatiaľčo -NoSW indikuje, že vylúčené slová boli v daných indexoch odstránené.

5 Záver

Štúdia prezentuje prístup pre analýzu textov s náboženskými témami pomocou numerických reprezentácií textu zvaných embeddingy, generovanými troma vybranými predtrénovanými jazykovými modelmi: *Slovak-BERT*, OpenAI *text-embedding-3-small* a *BGE M3* model. Výberu modelov predchádzalo posúdenie ich schopnosti "rozumieť slovenčine" a náboženskej terminológii. Pre zvolené tri modely sme konštatovali dostatočnú schopnosť, čo ich predurčilo ako vhodných kandidátov na zvládnutie úlohy získavania informácií z danej sady dokumentov.

Výzvy týkajúce sa kvality testovacích otázok boli adresované pomocou techniky kontextovej augmentácie. Tento prístup pomohol pri formulovaní vhodnejších otázok, čo viedlo k získavaniu relevantnejších častí textu, ktoré zachytávali všetky nuansy tém, ktoré teológovia v texte hľadajú.

Výsledky demonštrujú, že efektívnosť embeddingov generovaných týmito modelmi, hlavne modelom *text-embedding-3-small* od OpenAI, je dostatočná na hlboké porozumenie kontextu, aj v slovenskom jazyku. Hodnoty metriky recall pre embeddingy tohto modelu sa líšia v závislosti od témy a použitých testovacích otázok, pričom najlepšia hodnota bola dosiahnutá pre tému *Radikálnosť v prevzatí modelu života* dosahujúc takmer 90%, s použitím GPT otázok a dĺžky textových blokov 700 znakov. Vo všeobecnosti, *text-embedding-3-small* model mal najlepšie výsledky s najväčšou analyzovanou dĺžkou blokov textu, vykazujúc mierny trend zvyšujúcej sa hodnoty recall so zväčšujúcou sa dĺžkou blokov textu. Téma *Divné/čudné/silné* mala najnižšiu hodnotu recall, čo môže byť dôsledkom neurčitosti v špecifikácii tejto témy.

Pre *Slovak-BERT* embeddingové indexy sú hodnoty recall o niečo nižšie, ale stále pomerne pôsobivé vzhľadom na jednoduchosť tohto jazykového modelu. Lepšie výsledky boli získané v použitím GPT otázok, s najlepšou hodnotou 47.1% pre tému *Radikálnosť v prevzatí modelu života* s dĺžkou blokov 700 znakov, a s embeddingami vytvorenými z textu s odstránenými vylúčenými slovami. Celkovo, tento model najviac ťažil z odstraňovania vylúčených slov.

Čo sa týka *BGE M3* embeddingov, výsledky boli taktiež veľmi dobré, dosahujúc vysokú hodnotu recall metriky, aj keď nie až takú vysokú ako v prípade OpenAI embeddingov. Ale vzhľadom na to, že BGE M3 je verejne dostupný model, sú tieto výsledky pozoruhodné.

Tieto zistenia zdôrazňujú potenciál využitia veľkých jazykových modelov pre špecializované oblasti ako analýza textu s náboženskými témami. Výskum by sa ďalej mohol zaoberať zhlukovaním embeddingov za účelom odhalenia asociácií a inšpirácií autorov týchto diel. Pre teológov, budúca práca spočíva v analýze získaných častí textu s cieľom identifikovať odchýlky od oficiálneho učenia Katolíckej cirkvi, čím sa objasnia interpretácie a pohľady hnutia.

6 Poďakovanie

Výskum bol realizovaný s podporou Národného kompetenčného centra pre HPC, projektu EuroCC 2 a Národného Superpočítačového Centra na základe dohody o grante 101101903-EuroCC 2-DIGITAL-EUROHPC-JU-2022-NCC-01.

Výskum bol realizovaný s využitím výpočtovej infraštruktúry obstaranej v projekte Národné kompetenčné centrum pre vysokovýkonné počítanie (kód projektu: 311070AKF2) financovaného z Európskeho fondu regionálneho rozvoja, Štrukturálnych fondov EU Informatizácia spoločnosti, operačného programu Integrovaná infraštruktúra 2014-2020.

Literatúra

[1] Matúš Pikuliak, Štefan Grivalský, Martin Konôpka, Miroslav Blšták, Martin Tamajka, Viktor Bachratý, Marián Šimko, Pavol Balážik, Michal Trnka, and Filip Uhlárik. Slovakbert: Slovak masked language model, 2021.

[2] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024.

[3] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual e5 text embeddings: A technical report, 2024.

[4] Harrison Chase. Langchain. https://github.com/langchain-ai/langchain, 2022. Accessed: May 2024.

[5] Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. Query rewriting for retrieval-augmented large language models, 2023.

[6] Rolf Jagerman, Honglei Zhuang, Zhen Qin, Xuanhui Wang, and Michael Bendersky. Query expansion by prompting large language models, 2023.