Klasifikácia intentov pre bankové chatboty pomocou veľkých jazykových modelov

Bibiána Lajčinová [1], Patrik Valábek [1], [3], Michal Spišiak [2]

[1] Národné superpočítačové centrum, Bratislava, Slovenská republika
[2] nettle, s.r.o., Bratislava, Slovenská republika
[3] Ústav informatizácie, automatizácie a matematiky, Slovenská technická univerzita v Bratislave, Slovenská republika

Abstrakt

Tento článok hodnotí použitie veľkých jazykových modelov na klasifikáciu intentov v chatbote s preddefinovanými odpoveďami, určenom pre webové stránky bankového sektora. Zameriavame sa na efektivitu modelu SlovakBERT a porovnávame ho s použitím multilingválnych generatívnych modelov, ako sú *Llama 8b instruct* a *Gemma 7b instruct*, v ich predtrénovaných aj fine-tunovaných verziách. Výsledky naznačujú, že SlovakBERT dosahuje lepšie výsledky než ostatné modely, a to v presnosti klasifikácie ako aj v miere falošne pozitívnych predikcií.

1 Úvod

Príchod digitálnych technológií výrazne ovplyvnil aj sektor zákazníckych služieb, pričom výrazný posun je pozorovateľný hlavne v integrácii chatbotov do zákazníckej podpory. Tento trend možno pozorovať najmä na webových stránkach firiem, kde chatboty slúžia na zodpovedanie zákazníckych otázok týkajúcich sa daného biznisu. Títo virtuálni asistenti sú kľúčoví pri poskytovaní základných informácií zákazníkom, čím znižujú množstvo pracovných úloh, ktoré by inak museli riešiť pracovníci zákazníckej podpory.

V oblasti vývoja chatbotov bolo možné v posledných rokoch pozorovať obrovský nárast využitia generatívnej umelej inteligencie na tvorbu personalizovaných odpovedí. Napriek tomuto technologickému pokroku niektoré firmy stále uprednostňujú štruktúrovaný prístup k interakciám chatbota. V tomto prípade sú odpovede vopred definované, nie generované počas interakcie. Týmto je zaručená presnosť informácií v odpovediach bota a zároveň je zabezpečené konzistentné dodržiavanie komunikačného štýlu danej firmy. Vývoj chatbotov zvyčajne zahŕňa definovanie špecifických kategórií nazývaných intenty. Každý intent predstavuje konkrétny dopyt zákazníka, čo umožňuje chatbotu poskytnúť adekvátnu odpoveď. Najväčšou výzvou tohto systému je preto presná identifikácia zákazníkovho zámeru (intentu) na základe jeho textového vstupu.

2 Popis problému

Tento článok je výsledkom spoločného úsilia Národného kompetenčného centra pre vysokovýkonné počítanie a spoločnosti nettle, s.r.o., ktorá je slovenským start-upom zameraným na spracovanie

prirodzeného jazyka, chatboty a voiceboty. V rámci tejto spolupráce sa sústredíme na návrh jazykového klasifikátora chatbota pre online prostredie banky. Na spracovanie rozsiahlych výpočtov potrebných na vývoj tohto riešenia boli použité zdroje HPC systému Devana.

V chatbotoch spomenutých v úvode je preferovaná vopred definovaná odpoveď namiesto generovanej. Kľúčovým krokom v počiatočnej fáze vývoja takéhoto chatbota je preto identifikácia súboru intentov v danej doméne. Tento krok je zásadný pre správne fungovanie chatbota a pre poskytovanie presných odpovedí na každý konkrétny intent. Takéto chatboty bývajú vysoko sofistikované a často zahŕňajú široké spektrum intentov, niekedy až niekoľko stoviek. Vývojári vytvárajú rôzne ukážkové frázy pre každý intent, ktoré by mohli používatelia použiť pri otázkach súvisiacich s konkrétnym zámerom. Tieto frázy zohrávajú zásadnú úlohu pri definovaní jednotlivých intentov a slúžia ako trénovacie dáta pre klasifikačný algoritmus.

Náš základný model na klasifikáciu intentov, ktorý nevyužíva hlboké učenie, dosahuje presnosť 67% na reálnych testovacích dátach, podrobnejšie popísaných v ďalšej časti tohto článku. Cieľom práce je vyvinúť model založený na hlbokom učení, ktorý prekoná výkon tohto základného modelu.

Prezentujeme dva rôzne prístupy k riešeniu tejto úlohy. Prvý z nich skúma aplikáciu modelu BERT (Bidirectional Encoder Representations from Transformers), ako základ pre klasifikáciu intentov. Druhý prístup sa zameriava na využitie generatívnych veľkých jazykových modelov (LLM z angl. large language models) pomocou prompt inžinieringu na identifikáciu vhodného intentu, pričom skúmame využitie týchto modelov s fine-tuningom aj bez neho.

2.1 Dáta

Naša trénovacia dátová sada pozostáva z párov (text, intent), kde každý text predstavuje príklad dopytu adresovaného chatbotovi, ktorý vyvolá príslušný intent. Táto dátová množina je precízne skomponovaná tak, aby pokrývala celé spektrum preddefinovaných intentov, zaručujúc dostatočný objem textových príkladov pre každú kategóriu.

V našej štúdii pracujeme s rozsiahlym súborom intentov, pričom každý je doplnený o príslušné príklady dopytov. Používame dve trénovacie množiny: "simple" množinu, ktorá obsahuje 10 až 20 príkladov pre každý intent, a "generated" množinu, ktorá zahŕňa 20 až 500 príkladov na intent. Množina "generated" poskytuje väčší objem dát, avšak s opakujúcimi sa frázami v rámci jednotlivých intentov.

Tieto zoskupenia dát sú pripravené na spracovanie supervizovanými klasifikačnými modelmi. Tento proces zahŕňa konverziu množiny intentov do číselného poradia a priradenie každého textového príkladu k príslušnému číslu intentu, po čom nasleduje samotné trénovanie modelu.

Okrem trénovacej sady využívame aj testovaciu dátovú sadu, ktorá obsahuje približne 300 párov (text, intent), získaných z reálnej prevádzky chatbota, čo nám poskytuje autentický obraz interakcií používateľov. Všetky texty v tejto dátovej sade sú manuálne anotované ľudskými anotátormi. Táto sada slúži na hodnotenie výkonu našich klasifikačných modelov porovnávaním predikovaných intentov so skutočnými.

Všetky spomínané dátové množiny sú vlastníctvom spoločnosti nettle, s.r.o., a preto nebudú detailnejšie diskutované.

2.2 Vyhodnotenie

V tomto článku sú modely hodnotené predovšetkým na základe ich presnosti meranej na reálnej testovacej dátovej sade obsahujúcej 300 pozorovaní. Každé z týchto pozorovaní patrí do jedného z preddefinovaných intentov, na ktorých boli modely trénované. Presnosť počítame ako pomer správne klasifikovaných vzoriek k celkovému počtu vzoriek. Pre modely, ktorých výstupom je aj

pravdepodobnosť predikcie, ako napr. BERT, je vzorka považovaná za správne klasifikovanú iba vtedy, ak jej pravdepodobnosť predikcie do správnej triedy (intentu) prekročí stanovený prah.

Druhotnou metriku používanou na vyhodnotenie modelov je miera falošne pozitívnych predikcií (FPR z angl. false positive rate), kde je preferovaná čo najnižšia hodnota. Na vyhodnotenie tejto metriky používame syntetické dáta, ktoré nepatria do žiadneho intentu. Očakáva sa, že modely budú v tomto prípade produkovať nízke hodnoty pravdepodobnosti predikcie (pre model BERT), alebo klasifikovať tieto vzorky do triedy "invalid" (pre generatívne jazykové modely).

V celom článku sa pod pojmami presnosť a FPR rozumejú metriky vypočítané týmto spôsobom.

2.3 Prístup 1: Klasifikácia intentov pomocou modelov BERT

2.3.1 SlovakBERT

Keďže dáta sú v slovenskom jazyku, bolo nevyhnutné vybrať model, ktorý "rozumie" slovenčine. Preto sme sa rozhodli pre model s názvom SlovakBERT [5], ktorý je prvým verejne dostupným veľkým modelom pre slovenčinu.

Na dosiahnutie najlepšieho výkonu sme vykonali viacero experimentov s optimalizáciou tohto modelu. Tieto pokusy zahŕňali ladenie hyperparametrov, rôzne techniky predspracovania textu a, hlavne, výber trénovacích dát.

Vzhľadom na existenciu dvoch trénovacích dátových množín s relevantnými intentami ("simple" a "generated"), ako prvé boli vykonané experimenty s rôznymi pomermi vzoriek z týchto dvoch množín. Výsledky ukázali, že optimálny výkon modelu sa dosahuje pri trénovaní pomocou "generated" dátovej sady.

Po výbere dátovej množiny boli vykonané ďalšie experimenty, zamerané na výber správneho predspracovania dát. Testovali sme nasledujúce možnosti:

- prevod celého textu na malé písmená,
- odstránenie diakritiky z textu, a
- odstránenie interpunkcie z textu.

Ďalej boli testované aj kombinácie týchto troch možností. Keďže použitý model SlovakBERT je citlivý na veľké a malé písmená a tiež na používanie diakritiky, všetky tieto transformácie textu ovplyvňujú celkový výkon modelu trénovaného na týchto dátach.

Zistenia z experimentov odhalili, že najlepšie výsledky sú dosiahnuté, keď je text prevedený na malé písmená a je odstránená diakritika aj interpunkcia.

Ďalším skúmaným aspektom počas experimentálnej fázy bol výber vrstiev, ktoré budú fine-tunované. Testovali sme fine-tunovanie štvrtiny, polovice, troch štvrtín a celého modelu, pričom sme skúmali aj variácie ako napríklad fine-tunovanie celého modelu niekoľko epoch a následné fine-tunovanie zvoleného počtu vrstiev až do konvergencie. Výsledky ukázali, že priemerné zlepšenie získané týmito úpravami je štatisticky nevýznamné. Keďže cieľom je vytvoriť čo najjednoduchší algoritmus, tieto zmeny neboli vo výslednom modeli realizované.

Každý experiment bol vykonaný trikrát až päťkrát na zabezpečenie spoľahlivosti výsledkov.

Najlepší model dosiahol priemernú presnosť 77.2% so smerodajnou odchýlkou 0.012.

2.3.2 Banking-Tailored BERT

Keďže naše dáta obsahujú terminológiu špecifickú pre bankový sektor, rozhodli sme sa využiť model BERT, ktorý bol fine-tunovaný špeciálne na dátach pre sektor bankovníctva a financií. Avšak, keďže tento model rozumie výlučne angličtine, bolo nutné trénovanie dáta preložiť.

Na preklad sme použili DeepL API[1]. Najprv sme preložili trénovaciu, validačnú a testovaciu množinu. Vzhľadom na povahu angličtiny, nebol text ďalej upravovaný (predspracovaný), ako tomu bolo v prípade použitia modelu SlovakBERT v sekcii 2.3.1. Následne sme optimalizovali model BERT pre bankovníctvo na preložených dátach.

Fine-tunovaný model dosiahol sľubné počiatočné výsledky, s presnosťou mierne prevyšujúcou 70%. Bohužiaľ, ďalšie trénovanie a ladenie hyperparametrov nepriniesli zlepšenie. Testovali sme aj ďalšie modely trénované na anglickom jazyku, no všetky priniesli podobné výsledky. Použitie anglického modelu sa ukázalo ako nedostatočné na dosiahnutie lepších výsledkov, pravdepodobne kvôli chybám v preklade. Preklad obsahoval nepresnosti spôsobené "šumom" v dátach, hlavne v testovacej dátovej sade.

2.4 Prístup 2: Klasifikáca intentov pomocou veľkých jazykových modelov

Ako bolo uvedené v sekcii 2, okrem fine-tunovania modelu SlovakBERT a ďalších modelov založených na architektúre BERT, sme skúmali aj využitie generatívnych veľkých jazykových modelov pre klasifikáciu intentov. Zamerali sme sa na inštrukčné modely, kvôli ich schopnosti pracovať s inštrukčnými promptami a tiež schopnosti odpovedať na otázky.

Keďže neexistujú verejne dostupné inštrukčné modely trénované výhradne na slovenčinu, vybrali sme niekoľko multilingválnych modelov: *Gemma 7b instruct* [6] a *Llama3 8b instruct* [1]. Na porovnanie ukážeme aj výsledky proprierárneho modelu OpenAI *gpt-3.5-turbo*, používaného za rovnakých podmienok ako vyššie uvedené verejne dostupné modely.

Podobne ako v článku [4], našou stratégiou je využitie promptov s možnosťami intentov a ich popismi na vykonanie predikcie intentu v režime zero-shot. Očakávame, že výstupom bude možnosť so správnym intentom. Keďže kompletná sada intentov s ich popismi by produkovala veľmi dlhé prompty, používame náš základný model na výber troch najlepších intentov. Dáta pre tieto modely boli pripravené nasledovne:

Každý prompt obsahuje vetu (otázku od používateľa) v slovenčine, štyri možnosti intentov s popismi a inštrukciu na výber najvhodnejšej možnosti. Prvé tri možnosti intentov sú vybrané základným modelom, ktorý má hodnotu Top-3 recall metriky 87%. Posledná možnosť je vždy "invalid" a mala by byť vybraná, keď žiadna z prvých troch možností nezodpovedá otázke používateľa, alebo ide o otázku mimo rozsahu intentov. V tomto nastavení je najvyššia možná dosiahnuteľná presnosť 87%.

2.4.1 Implementácia predtrénovaného LLM

Na úvod sme implementovali neoptimalizovaný predtrénovaný veľký jazykový model, čo znamená, že daný inštrukčný model bol použitý bez fine-tunovania na našich dátach.

Na zlepšenie výsledkov sme využili prompt inžiniering. Tento proces jemne preformuluje prompt, v našom prípade sme upravovali pokyny pre model, aby odpovedal napr. iba názvom intentu alebo číslom/písmenom, ktoré označuje správnu možnosť. Rovnako sme skúmali rôzne možnosti umiestnenia promptu (rola používateľa/rola systému) a experimentovali sme s rozdelením promptu, kde pokyn pre model bol umiestnený v úlohe systému a otázka spolu s možnosťami v úlohe používateľa.

Napriek týmto snahám tento prístup nepriniesol lepšie výsledky ako fine-tuning modelu SlovakBERT. Avšak, pomohol nám identifikovať najefektívnejšie formáty promptov pre fine-tuning týchto inštrukčných modelov. Tieto kroky boli zásadné pri analyzovaní správania modelov a ich vzorcov odpovedí, čo sme následne využili pri tvorbe stratégií na fine-tunovanie týchto modelov.

[1] www.deepl.com

2.4.2 Optimalizácia LLM

Prompty, na ktoré predtrénované modely reagovali najlepšie, boli využité pri fine-tuningu modelov. Keďže predtrénované veľké jazykové modely nevyžadujú rozsiahle trénovacie dátové množiny, použili sme našu "simple" dátovú množinu, podrobne opísanú v sekcii 2.1. Model bol následne fine-tunovaný tak, aby na zadané prompty odpovedal príslušnými názvami intentov.

Kvôli veľkosti vybraných modelov sme použili metódu nazývanú parameter efficient training (PEFT) [2], čo je stratégia zameraná na efektívne využívanie pamäte a znižovanie času výpočtu. PEFT trénuje len malú podmnožinu parametrov, s hodnotami zvyšných vôbec nehýbe, čím sa znižuje počet trénovateľných parametrov. Konkrétne sme použili prístup Low-Rank Adaptation (LoRA) [3].

Na dosiahnutie najlepšieho výkonu boli ladené aj hyperparametre, vrátane rýchlosti učenia, veľkosti dávky, parametra *lora alpha* v konfigurácii LoRA, počtu krokov akumulácie gradientu a formulácie "chat template".

Optimalizácia jazykových modelov si vyžaduje značné výpočtové zdroje, čo znamená potrebu využitia HPC (High Performance Computing) zdrojov na dosiahnutie požadovaného výkonu a efektivity. HPC systém Devana, ktorý je vybavený 4 GPU akcelerátormi NVidia A100 s 40 GB pamäte na každom uzle, poskytuje potrebnú výpočtovú kapacitu. V našom prípade sa oba fine-tunované modely zmestia do pamäte jedného GPU akcelerátora (v plnej veľkosti) s maximálnou veľkosťou dávky 2.

Aj keď využitie všetkých 4 GPU akcelerátorov na jednom uzle by skrátilo čas trénovania a umožnilo väčšiu veľkosť dávky, pre účely benchmarkingu a na zabezpečenie konzistentnosti a porovnateľnosti výsledkov sme vykonali všetky experimenty iba s jedným GPU akcelerátorom.

Toto úsilie viedlo k určitým zlepšeniam vo výkone modelov. Pre model *Gemma 7b instruct* sa podarilo znížiť počet falošne pozitívnych predikcií. Na druhej strane, pri fine-tuningu modelu *Llama3 8b instruct* došlo k zlepšeniu oboch metrík (presnosť a počet falošne pozitívnych predikcií). Avšak, ani jeden z týchto modelov po optimalizácii neprekročil schopnosti fine-tunovaného modelu SlovakBERT.

Čo sa týka modelu *Gemma 7b instruct*, niektoré množiny hyperparametrov priniesli vyššiu presnosť, ale aj vysokú hodnotu FPR, zatiaľčo ďalšie viedli k nižšej presnosti a nízkej hodnoty FPR. Hľadanie množiny hyperparametrov, ktorá by zaistila vyvážené hodnoty presnosti a FPR bolo náročné. Najlepšia konfigurácia dosiahla presnosť mierne prevyšujúcu 70% s hodnotou FPR 4.6%. Porovnaním týchto hodnôt s výkonom tohto modelu bez optimalizácie zisťujeme, že fine-tunovanie iba zľahka zvýšilo presnosť, ale dramaticky redukovalo počet falošne pozitívnych predikcií, takmer o 70%.

Pre model *Llama3 8b instruct*, najlepšia konfigurácia dosiahla presnosť 75.1% s hodnotou FPR 7.0%. V porovnaní s výkonom modelu bez optimalizácie prinieslo fine-tunovanie vyššiu presnosť a zároveň prispelo k významnému zníženiu hodnoty FPR, ktorá sa znížila na polovicu.

2.4.3 Porovnanie s proprietárnym modelom

Na porovnanie nášho prístupu s proprietárnym veľkým jazykovým modelom sme vykonali experimenty s modelom *gpt-3.5-turbo* od OpenAI[2]. Použili sme identické prompty na zabezpečenie spravodlivého porovnania a testovali sme ako predtrénovanú, tak aj fine-tunovanú verziu tohto modelu. Bez fine-tuningu dosiahol *gpt-3.5-turbo* presnosť 76%, hoci vykazoval značnú mieru falošne pozitívnych predikcií. Po fine-tuningu sa presnosť zvýšila na takmer 80% a miera falošne pozitívnych predikcií sa výrazne znížila.

[2] www.openai.com

3 Výsledky

V našej počiatočnej stratégii, ktorá zahŕňala fine-tuning modelu SlovakBERT, sme dosiahli priemernú presnosť 77.2% so štandardnou odchýlkou 0.012, čo predstavuje nárast o 10% v porovnaní s presnosťou základného modelu.

Fine-tuning modelu BERT, špeciálne trénovaný pre bankovníctvo, dosiahol presnosť tesne pod 70%. Tento výsledok prekonáva presnosť základného modelu, avšak nedosahuje výkon fine-tunovaného modelu SlovakBERT.

Následne sme experimentovali s generatívnymi jazykovými modelmi (predtrénovanými, ale nie fine-tunovanými na našich dátach). Hoci tieto modely preukázali sľubné schopnosti, ich výkon bol nižší v porovnaní s fine-tunovaným modelom SlovakBERT. Preto sme pristúpili k fine-tuningu týchto modelov, konkrétne *Gemma 7b instruct* a *Llama3 8b instruct*.

Fine-tunovaná verzia modelu *Gemma 7b instruct* vykazovala finálnu presnosť porovnateľnú s modelom BERT optimalizovaným pre bankovníctvo, a fine-tunovaný model *Llama3 8b instruct* dosiahol výkon o niečo horší než fine-tunovaný SlovakBERT. Napriek rozsiahlemu úsiliu nájsť konfiguráciu hyperparametrov, ktorá by prekonala schopnosti modelu SlovakBERT, neboli tieto pokusy úspešné. Takže model SlovakBERT je najlepším z porovnávaným modelov.

Všetky výsledky sú zobrazené v Tabuľke 1, vrátane nášho základného modelu a tiež výsledkov proprietárneho modelu od OpenAI pre porovnanie.

Názov modelu	**In-scope Presnosť**	**Out-of-scope FPR**
Základný model	67.6	22.5
SlovakBERT fine-tunovaný	77.2	6.3
BERT pre bankovníctvo	68.5	4.0
Gemma 7b instruct predtrénovaný	69.5	73.6
Gemma 7b instruct fine-tunovaný	70.6	4.6
Llama3 8b instruct predtrénovaný	65.5	14.1
Llama3 8b instruct fine-tunovaný	75.1	7.0
gpt-3.5-turbo predtrénovaný	76.6	32.4
gpt-3.5-turbo fine-tunovaný	79.5	4.3

Tabuľka 1: Porovnanie hodnôt metrík presnosť a FPR, definovaných v časti 2.2, pre všetky analyzované modely. Hodnoty sú uvedené v percentách.

4 Záver

Cieľom tohto článku bolo nájsť prístup na riešenie úlohy klasifikácie intentov, ktorý využíva predtrénovaný jazykový model (fine-tunovaný ako aj pôvodný bez fine-tuningu) ako základ pre chatbot pre sektor bankovníctva. Dáta pre našu prácu pozostávali z párov textu a intentu, kde text predstavuje dopyt používateľa (zákazníka) a intent predstavuje príslušný zámer.

Experimentovali sme s viacerými modelmi, vrátane modelu SlovakBERT, BERT pre bankovníctvo a generatívnych modelov *Gemma 7b instruct* a *Llama3 8b instruct*. Po pokusoch s dátovými množinami, konfiguráciami hyperparametrov pre fine-tuning a prompt inžinieringu, sa ukázalo, že optimalizácia modelu SlovakBERT je najlepším prístupom, s finálnou presnosťou o niečo vyššou než 77%, čo predstauje nárast o 10% v porovnaní so základným modelom.

Táto štúdia zdôrazňuje efektivitu optimalizácie predtrénovaných jazykových modelov pre vývoj robustného chatbota s presnou klasifikáciou zámerov užívateľov. Tieto poznatky budú v

budúcnosti využité na ďalšie zlepšenie výkonu a efektivity v reálnych bankových aplikáciách.

5 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.

Literatúra

[1] AI@Meta. Llama 3 model card. 2024. URL: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

[2] Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. Parameter-efficient fine-tuning for large models: A comprehensive survey, 2024. arXiv:2403.14608.

[3] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685, 2021. URL: https://arxiv.org/abs/2106.09685, arXiv:2106.09685.

[4] Soham Parikh, Quaizar Vohra, Prashil Tumbade, and Mitul Tiwari. Exploring zero and few-shot techniques for intent classification, 2023. URL: https://arxiv.org/abs/2305.07157, arXiv:2305.07157.

[5] Matúš Pikuliak, Štefan Grivalský, Martin Konôpka, Miroslav Blšťák, Martin Tamajka, Viktor Bachratý, Marián Šimko, Pavol Balážik, Michal Trnka, and Filip Uhlárik. Slovakbert: Slovak masked language model. *CoRR*, abs/2109.15254, 2021. URL: https://arxiv.org/abs/2109.15254, arXiv:2109.15254.

[6] Gemma Team, Thomas Mesnard, and Cassidy Hardin et al. Gemma: Open models based on gemini research and technology, 2024. arXiv:2403.08295.