Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

Sémantická podobnosť textov v slovenskom jazyku

Bakalárska práca

2024
Matej Krajčovič

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

Sémantická podobnosť textov v slovenskom jazyku

Bakalárska práca

Študijný program:	Aplikovaná informatika
Študijný odbor:	Informatika
Školiace pracovisko:	Katedra Aplikovanej informatiky
Školiteľ:	Lukáš Radoský

Bratislava, 2024
Matej Krajčovič

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta:
Študijný program:

Študijný odbor:
Typ záverečnej práce:
Jazyk záverečnej práce:
Sekundárny jazyk:

Názov:

Anotácia:

Vedúci:
Katedra:
Vedúci katedry:

Dátum zadania:

Dátum schválenia:

garant študijného programu

.....

študent

.....

vedúci práce

Poďakovanie: Chcel by som poďakovať svojmu školiteľovi, Lukášovi Radoskému, za jeho pomoc a podporu počas zhotovovania mojej práce.

Abstrakt

Táto práca sa zaoberá analýzou a porovnaním rôznych metód určovania sémantickej podobnosti textov v slovenskom jazyku. V práci sú predstavené a testované tri modely: Text-embedding-ada-002, SlovakBert a Paraphrase-multilingual-mpnet-base-v2, ktoré boli použité na tri vybrané dátové sady. Cieľom práce bolo zistiť, ktorý z modelov poskytuje najpresnejšie výsledky pri analýze sémantickej podobnosti slovenských textov. Výsledky ukázali, že model Text-embedding-ada-002 preukázal konzistentnú výkonnosť naprieč všetkými datasetmi, zatiaľ čo SlovakBert vyžaduje ďalšie špecifické trénovanie na zlepšenie jeho výkonu. Model Paraphrase-multilingual-mpnet-base-v2 zase vykázal výborné výsledky na datasete SemEval-2015-example_sk, ale jeho úplné schopnosti neboli možné testovať kvôli obmedzeniam platených služieb. Práca poskytuje základ pre ďalší vývoj a zlepšenie modelov pre sémantickú analýzu textov v slovenskom jazyku.

Kľúčové slová: sémantická podobnosť textov, strojové učenie, rozsiahle jazykové modely, umelá inteligencia, slovenský jazyk

Abstract

This thesis deals with the analysis and comparison of different methods of determining the semantic similarity of texts in the Slovak language. Three models are present and test in this thesis: Text-embedding-ada-002, SlovakBert and Paraphrase-multilingual-mpnet-base-v2, which were used on three selected data sets. The aim was to evaluate which of the models provides the most accurate results when analyzing the semantic similarity of Slovak texts. The results showed that the Text-embedding-ada-002 model demonstrated consistent performance across all datasets, while SlovakBert requires additional specific training to improve its performance. The Paraphrase-multilingual-mpnet-base-v2 model, on the other hand, showed excellent results on the SemEval-2015-example_sk dataset, but its full capabilities could not be tested due to the limitations of paid services. The work provides a basis for further development and improvement of models for the semantic analysis of texts in the Slovak language.

Keywords: semantic similarity of texts, machine learning, extensive language models, artificial intelligence, Slovak language

Obsah

Zoznam tabuliek

Úvod

Sémantická podobnosť textov sa zaoberá porovnávaním dvoch textov. Toto porovnanie sa vyjadruje pridelením číselnej hodnoty v určitom intervale. Najčastejšie intervaly pre toto porovnanie sú 0-1 alebo 0-5. Čím je pridelená hodnota nižšia, ukazuje na menšiu podobnosť. Naopak vyššia hodnota hovorí o väčšej podobnosti. Texty sú sémanticky podobné, ak majú podobný význam alebo vyjadrujú podobné myšlienky, no môžu obsahovať rôzne slová a frázy. Sémantická podobnosť textov je dôležitou súčasťou výskumu textov, ako sú napríklad preklady, generovanie textov, kontrola plagiátorstva[4].

Na zistenie podobnosti sa dajú použiť rôzne prístupy, no pre slovenský jazyk ich nie je toľko ako pre iné, viac svetovo rozšírené jazyky. Preto sme obmedzení na prístupy ako sú napríklad text-embedding-ada-002, SlovakBert a paraphrase-multilingual-mpnet-base-v2. V tejto práci sa zameriame na prácu s týmito modelmi, pretože sú pripravené na okamžité používanie. My budeme evaluovať ich efektivitu nad vybranými dátovými sadami 1.3.1.

Výber týchto modelov nám tiež poskytuje príležitosť preskúmať, ako rôzne prístupy k modelovaniu jazyka môžu ovplyvniť výsledky v sémantickej podobnosti. Zatiaľ čo modely ako SlovakBert sú špecificky trénované pre slovenský jazyk, iné, ako je paraphrase-multilingual-mpnet-base-v2, sú navrhnuté pre viacjazyčné použitie. Toto porovnanie nám umožňuje nielen zistiť, ktorý model je najpresnejší, ale aj pochopiť ich silné a slabé stránky pri práci v slovenskom jazyku.

Kapitola 1

Sémantická podobnosť textov

Táto kapitola sa zaoberá konceptom sémantickej podobnosti textov. Preskúmava, ako moderné algoritmy a modely umožňujú analyzovať a porovnávať texty na základe ich významu, nie len na základe použitých slov. Sústreďuje sa na rôzne metódy určovania sémantickej podobnosti, vrátane vektorových reprezentácií textu a transformačných modelov.

1.1 Súčasný stav

Sémantická podobnosť textov je úloha, ktorá hodnotí mieru, do akej dva textové segmenty vyjadrujú podobný alebo rovnaký význam. Oblasti využitia sú systémy automatického odpovedania, klasifikácia textov, kontrola plagiátorstva [18].

V zahraničí sa sémantickej podobnosti textov venoval rozsiahly výskum, pričom veľkou zásluhou pre tento výskum mal projekt SemEval[1] [1]. V tomto projekte sa vyskytli úlohy zamerané na sémantickú podobnosť textov od roku 2012 až po rok 2017. Pre zahraničné jazyky existuje mnoho modelov zaoberajúcich sa sémantickou podobnosťou textov. Jedny zo známejších sú BERT modely, pre ktoré existujú verzie vo viacerých jazykoch [3, 13, 10]. Existuje niekoľko zahraničných dátových sád. Dve z nich, s ktorými aj budeme pracovať, sú STS Benchmark a SICK [15, 9].

Na slovensku sa výskum sémantickej podobnosti textov postupne vyvíja. Jeden z najvýznamnejších pokrokov pre tento výskum bol vývoj SlovakBERT [7], ktorý je slovenskou verziou globálne populárneho modelu BERT. Dátové sady v slovenskom jazyku doposiaľ neexistujú, no je možné využívať strojovo preložené dátové sady z iných jazykov.

[1] `https://semeval.github.io/`

1.2 Cieľ práce

Cieľom tejto práce je skúmať a analyzovať metódy určovania sémantickej podobnosti textov. Zároveň sa zameriava na porovnanie existujúcich prístupov a zisťovanie najpresnejšej metódy pre slovenský jazyk. Ďalej sa práca venuje hodnoteniu použiteľnosti týchto metód a analýze ich výsledkov.

1.3 Praktický výskum

1.3.1 Dátové sady

Dátové sady sú zbierky informácií, ktoré obsahujú údaje o konkrétnych objektoch alebo javoch v reálnom svete. Tieto dátové sady slúžia na uchovávanie a organizovanie informácií tak, aby boli ľahko prístupné a použiteľné pre analýzu alebo ďalšie spracovanie [11].

Sady, ktoré sme použili, boli stsbenchmark_sk, sick_sk a SemEval-2015-example_sk. Použili sme ich na vyhodnotenie presnosti modelu pri predikcii podobnosti medzi vetami. Porovnávali sme výstupy modelu s anotáciami v týchto dátových sadách. Anotácie sú miery podobnosti ohodnotené ľudskými anotátormi. Tieto anotácie sú destainné čísla v intervaloch od 0 do 5. Neexistujú však dátové sady pre slovenský jazyk, preto sa pre slovenské modely využívajú preložené sady zo zahraničia.

STS Benchmark je dátová sada obsahujúca 8628 dvojíc viet s anotáciami [15]. STS Benchmark je starostlivo vybraná sada anglických dátových sád, ktoré boli použité v súťažiach SemEval medzi rokmi 2012 a 2017 [2]. SemEval je séria medzinárodných workshopov zameraných na výskum spracovania prirodzeného jazyka, ktorej cieľom je posunúť súčasný stav vývoja v oblasti sémantickej analýzy[2]. My sme preto využili strojovo preloženú verziu tejto dátovej sady do slovenského jazyka.

Ďaľšou dátovou sadou bol SICK. Je to dátová sada, ktorá bola vyvinutá s cieľom vyplniť medzeru v existujúcich dátových sadách, týkajúcich sa spracovania prirodzeného jazyka. Obsahuje veľké množstvo dvojíc viet bohatých na lexikálne, syntaktické a sémantické javy, ktoré sa očakávajú od modelov založených na distribučnej sémantike. SICK je špeciálne navrhnutá tak, aby nevyžadovala zaoberanie sa inými aspektmi existujúcich dátových sád obsahujúcich vety, ktoré nie sú v rámci rozsahu kompozičnej distribučnej sémantiky [9].Táto sada obsahuje 9840 dvojíc viet s anotáciami. Rovnako ako pri STS Benchmark, aj túto sadu sme strojovo preložili do slovenského jazyka.

[2] https://semeval.github.io/

Poslednou použitou dátovou sadou je SemEval-2015-example. Je to sada, ktorú sme vytvorili pre túto prácu prekladom dátovej sady, použitej ako príklad v SemEval-2015 [1] tabuľka č. 1.

1.3.2 Kosínusová podobnosť

Na výpočet podobnosti medzi textami je potrebné použiť určitý algoritmus, pričom existuje viacero možností. Najznámejšie sú Jaccardová podobnosť, Kosínusová podobnosť, Euklidovská podobnosť a Manhattanská podobnosť [6].

V našej práci sme si vybrali kosínusovú podobnosť, pretože všetky tri použité metódy pracujú s vektorovými reprezentáciami textov. Kosínusová podobnosť dokáže určiť podobnosť medzi dvoma textami, pomocou výpočtu kosínusovej hodnoty medzi ich vektorovými reprezentáciami. Čím je výsledná hodnota vyššia, tým je podobnosť dvoch textov väčšia. Kosínusová podobnosť sa počíta vzorcom [14] č. 1.1.

$$Sim(\vec{q}, \vec{d}) = \frac{\vec{q} \cdot \vec{d}}{|\vec{q}||\vec{d}|} = \frac{\sum_{k=1}^{t} w_{qk} \times w_{dk}}{\sqrt{\sum_{k=1}^{t}(w_{qk})^2} \cdot \sqrt{\sum_{k=1}^{t}(w_{dk})^2}} \tag{1.1}$$

Tento vzorec predstavuje výpočet kosínusovej podobnosti medzi dvoma vektormi $\vec{q}$ a $\vec{d}$, ktoré reprezentujú dva dokumenty. Tieto vektory vieme zapísať ako

$$\vec{q} = (w_{q0}, w_{q1}, ..., w_{qk}) \tag{1.2}$$

pre $\vec{q}$ a pre $\vec{d}$

$$\vec{d} = (w_{d0}, w_{d1}, ..., w_{dk}), \tag{1.3}$$

kde vektor reprezentuje ľubovolný text v rôznych veľkostiach (slovo, veta, dokument). Vzorec počíta kosínusovú podobnosť ako podiel súčinu vektorov $\vec{q}$ a $\vec{d}$ so súčinom veľkostí vektorov $\vec{q}$ a $\vec{d}$. Čitateľ bude preto obsahovať skalárny súčin dvoch vektorov. Ten zapíšeme ako

$$\sum_{k=1}^{t} w_{qk} \times w_{dk}. \tag{1.4}$$

V menovateli bude súčin veľkostí týchto dvoch vektorov, ktoré získame ako odmocninu zo súčtu štvorcov všetkých prvkov daného vektora.

$$\vec{q} = \sqrt{\sum_{k=1}^{t}(w_{qk})^2} \tag{1.5}$$

$$\vec{d} = \sqrt{\sum_{k=1}^{t}(w_{dk})^2} \tag{1.6}$$

Výsledná hodnota je v intervale od 0 do 1, lebo pre w_{qi} a w_{di} platí, že $(0 \leq i \leq k)$ [14]. Hodnota 1 predstavuje, že vektory majú rovnaký smer, čo znamená, že sú si vektory a aj texty, ktoré predstavujú, rovnaké. Naopak, hodnota 0 znamená, že vektory sú na seba navzájom kolmé. To hovorí o tom, že danné vektory nie sú podobné ani odlišné na základe meraných vlastností.

1.3.3 Používanie metód

Text-embedding-ada-002

Model text-embedding-ada-002 je model druhej generácie od spoločnosti OpenAI, vytvorený pre úlohy súvisiace s sémantikou podobnosťou textov [8]. Pri práci s text-embedding-ada-002 je potrebné komunikovať s API serverom. Na túto komunikáciu je potrebné získať API kľúč pre OpenAI, ktorý je nevyhnutný na prácu s OpenAI službami. Jedna z týchto služieb je aj získavanie vektorových reprezentácií viet. Pomocou API kľúču môžeme posielať požiadavku, v ktorej povieme, ktorých viet vektorovú reprezentáciu potrebujeme. Tento server ju spracuje a využije svoje predtrénované modely a pošle nám späť vektorové reprezentácie. Výslednú podobnosť viet sme nakoniec získali pomocou výpočtu Kosínusovej podobnosti. 1.3.2

SlovakBert

SlovakBert je prvý jazykový model vyvinutý len pre slovenský jazyk. Jednou z jeho funkcií je aj sémantická textová podobnosť [12]. Pre dosiahnutie lepších výsledkov je tento model potrebné dodatočne trénovať. Bol predstavený v roku 2021 [7], čo je pomerne neskoro, v porovnaní s ostatnými BERT modelmi. BERT model pre Holandsko (BERTje) bol predstavený v roku 2019 [3], rovnako ako pre Taliansko (AlBERTo) [13] a Francúzsko (CamemBERT) [10]. Rovnako ako pri text-embedding-ada-002, pomocou metódy SlovakBert sme museli získať vektorové reprezentácie viet. Na to sme museli vety tokenizovať do formátu, ktorému by model rozumel. To dokážeme spraviť pomocou tokenizátoru. Tokenizátor je nástroj, ktorý rozdeľuje text na menšie časti. Tieto časti sa nazývajú tokeny. Toto je potrebné, pretože SlovakBert vyžaduje, aby jednotlivé tokeny boli prevedené do číselného formátu, aby ich vedel spracovať. V takejto forme ich už vedel predtrénovaný model spracovať a získať vektorové reprezentácie dvoch viet, ktorých podobnosť sme získali pomocou Kosínusovej podobnosti 1.3.2. Túto podobnosť sme previedli na interval 0-5.

Paraphrase-multilingual-mpnet-base-v2

Model "paraphrase-multilingual-mpnet-base-v2"je verzia modelu MPNet od spoločnosti Microsoft [5]. MPNet využíva dodatočné informácie o pozícii slov, aby mal lepší

prehľad o celej vete. Toto zlepšenie pomáha modelu lepšie pochopiť text, pretože znižuje rozdiely medzi pozíciami slov v trénovacích a reálnych situáciách. Takto zohľadňuje celú vetu a znižuje tým nezrovnalosti v pochopení polohy slov vo vete [17]. Tento model je prístupný na platforme NLP cloud[3], kde je možné ho využívať zadarmo s nie plnou verziou alebo si zaplatiť plnú verziu. S paraphrase-multilingual-mpnet-base-v2 sme pracovali cez NLP cloud. Pracovali s verziou zadarmo, ktorá mala obmedzenie na počet požiadaviek na NLP cloud API. Preto sme vo verzii zadarmo priamo vkladali dva texty do modelu na ich stránke. Z toho dôvodu sme ručne vytvorili dátovú sadu mojtext.txt, ktorá obsahuje len 5 dvojíc viet na porovnanie.

1.3.4 Ohodnotenie výsledkov

Vyhodnotili sme prístupy Text-embedding-ada-002, SlovakBert a Paraphrase-multilingual-mpnet-base-v2 na dátových sadách 1.3.1, s využitím Pearsonovho korelačného koeficientu.

Pearsonov korelačný koeficient ukazuje mieru monotónnej asociácie medzi dvoma premennými. Tento monotónny vzťah medzi premennými rastie pri jednej premennej, ak rastie aj pri druhej, alebo naopak klesá pri jednej premennej, ak klesá aj pri druhej [16].

Tabuľka 1.1: Nasledujúca tabuľka predstavuje výsledky jednotlivých modelov na rôzne dátové sady obsahujúce texty.

Meno modelu v slovenčine	SICK	STS Benchmark	SemEval-2015
Text-embedding-ada-002	**0.6815**	**0.7224**	0.7998
SlovakBert	0.5803	0.6055	0.7883
Paraphrase-multilingual-mpnet-base-v2	N/A	N/A	**0.9475**

Tabuľka 1.1 poskytuje prehľad výsledkov Pearsonovho korelačného koeficientu pre Text-embedding-ada-002, SlovakBert a Paraphrase-multilingual-mpnet-base-v2 na dátové sady. Tieto údaje nám umožňujú porovnať efektívnosť jednotlivých modelov.

Model Text-embedding-ada-002 dosiahol koeficient 0.6815 na dátovej sade SICK_sk, 0.7224 na STS Benchmark_sk a 0.7998 na SemEval-2015-example_sk. Tento model ukázal konzistentnosť a efektívnosť s postupným zlepšením na náročnejších jazykových úlohách, ako sú tie nachádzajúce sa v datátových sadách SICK_sk a STS Benchmark_sk.

[3] `https://nlpcloud.com/home/playground/semantic-similarity`

SlovakBert, špecificky navrhnutý pre slovenský jazyk, má koeficient 0.5803 na SICK_sk a 0.6055 na STS Benchmark_sk, pričom na dátovej sade SemEval-2015-example_sk dosiahol 0.7883. Tieto výsledky poukazujú na to, že SlovakBert má potenciál pre vysokú presnosť, ale môže byť menej konzistentný.

Model Paraphrase-multilingual-mpnet-base-v2, nebol testovaný na prvých dvoch dátových sadách, ale dosiahol najlepší výsledok 0.9475 na SemEval-2015-example_sk. Tento výsledok naznačuje, že tento model je mimoriadne účinný, no nebolo nám to možné ukázať na väčších dátových sadách.

1.4 Výsledky práce

Na dátovej sade SICK_sk dosiahol najlepšie výsledky model Text-embedding-ada-002 s Pearsonovým koeficientom 0.6815, čo ho činí najúspešnejším modelom pre túto dátovú sadu. Tento model sa ukázal byť najefektívnejší aj na dátovej sade STS Benchmark_sk, kde dosiahol koeficient 0.7224, čo ukazuje jeho konzistentnú výkonnosť aj pri náročnejších jazykových úlohách. Na dátovej sade SemEval-2015-example_sk však najlepší výsledok dosiahol model Paraphrase-multilingual-mpnet-base-v2 s hodnotou 0.9475, čo poukazuje na jeho schopnosť porovnávania krátkych textov.

Model SlovakBert, ktorý bol špeciálne vyvinutý pre slovenský jazyk, nepredviedol na žiadnej dátovej sade najlepšie výsledky, avšak dôvodom môže byť, že sme ho dodatočne netrénovali. Hoci dosiahol na dátovej sade SemEval-2015-example_sk koeficient 0.7883, jeho výsledky by mohli byť výrazne zlepšené trénovaním a prispôsobením na konkrétne typy textov alebo špecifické jazykové úlohy. Tento proces by mohol výrazne zvýšiť jeho účinnosť pri riešení náročnejších jazykových úloh.

To, že sme nemohli testovať model Paraphrase-multilingual-mpnet-base-v2 na iných datasetoch kvôli obmedzeniam na platenú verziu platformy NLP cloud, bráni v úplnom vyhodnotení. Napriek tomu, jeho vynikajúci výkon na jednom datasete naznačuje, že z modelov nami testovaných by mohol byť najlepší model na určovanie sémantickej podobnosti textov.

Pri výbere najvhodnejšieho modelu je dôležité zvážiť nielen mieru úspešnosti, ale aj ďalšie faktory, ako sú rýchlosť a náklady na výpočet. Napriek pomalšiemu vykonávaniu môže byť model Text-embedding-ada-002 stále vhodnou voľbou pre úlohy, kde je dôležitejšia vyššia presnosť pred rýchlosťou. Avšak, ak je potrebná rýchlejšia odozva, Paraphrase-multilingual-mpnet-base-v2 s platenou verziou môže byť lepšou alternatívou. SlovakBert ponúka dobrý kompromis medzi presnosťou a rýchlosťou, najmä po ďalšom špecifickom trénovaní pre slovenský jazyk.

Záver

V práci sme predstavili a analyzovali rôzne prístupy k určovaniu sémantickej podobnosti textov v slovenskom jazyku. Naša práca poskytla porovnanie troch rôznych modelov: Text-embedding-ada-002, SlovakBert a Paraphrase-multilingual-mpnet-base-v2 na troch vybraných dátových sadách. Text-embedding-ada-002 preukázal konzistentnú výkonnosť na všetkých testovaných datasetoch, zatiaľ čo SlovakBert, hoci bol špeciálne navrhnutý pre slovenský jazyk, nedosiahol očakávané výsledky bez ďalšieho špecifického trénovania. Paraphrase-multilingual-mpnet-base-v2 vykázal najlepšie výsledky na datasete SemEval-2015-example_sk, čo naznačuje jeho potenciál v úlohách zameraných na krátke texty, ale obmedzenia spojené s prístupom k plateným službám nám zabránili v plnom testovaní jeho schopností.

Budúci výskum by sa mal zamerať na ďalšie trénovanie modelu SlovakBert, aby bol lepšie prispôsobený špecifickým potrebám slovenského jazyka. Taktiež by bolo užitočné získať prístup k plateným nástrojom, ako je napríklad plná verzia platformy, ktorá poskytuje model Paraphrase-multilingual-mpnet-base-v2, aby bolo možné správne vyhodnotiť jeho schopnosti.

Výsledky tejto práce nám dávajú základ pre zlepšenie nástrojov na meranie textovej podobnosti a prispievajú k lepšiemu pochopeniu funkcionality modelov strojového učenia v oblasti spracovania prirodzeného jazyka pre slovenský jazyk.

Literatúra

[1] Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. SemEval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. In Preslav Nakov, Torsten Zesch, Daniel Cer, and David Jurgens, editors, *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263, Denver, Colorado, June 2015. Association for Computational Linguistics.

[2] Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*, 2017.

[3] Wietse De Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. Bertje: A dutch bert model. *arXiv preprint arXiv:1912.09582*, 2019.

[4] Wael H Gomaa, Aly A Fahmy, et al. A survey of text similarity approaches. *international journal of Computer Applications*, 68(13):13–18, 2013.

[5] Álvaro Huertas-García, Javier Huertas-Tato, Alejandro Martín, and David Camacho. Countering misinformation through semantic-aware multilingual models. In *International conference on intelligent data engineering and automated learning*, pages 312–323. Springer, 2021.

[6] Chi-gon Hwang, Chang-Pyo Yoon, and Dai Yeol Yun. Sentence similarity analysis using ontology based on cosine similarity. In *Proceedings of the Korean Institute of Information and Commucation Sciences Conference*, pages 441–443. The Korea Institute of Information and Commucation Engineering, 2021.

[7] Jozef Kubík, Daniel Kyselica, and Martin Takáč. Efficient fine-tuning of slovakbert with epinet. 2023.

[8] Xiu Li, Aron Henriksson, Martin Duneld, Jalal Nouri, and Yongchao Wu. Evaluating embeddings from pre-trained language models and knowledge graphs for educational content recommendation. *Future Internet*, 16(1):12, 2023.

[9] Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 1–8, 2014.

[10] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah, and Benoît Sagot. Camembert: a tasty french language model. *arXiv preprint arXiv:1911.03894*, 2019.

[11] John P McCrae and Paul Buitelaar. Linking datasets using semantic textual similarity. *Cybernetics and information technologies*, 18(1):109–123, 2018.

[12] Matúš Pikuliak, Štefan Grivalskỳ, Martin Konôpka, Miroslav Blšták, Martin Tamajka, Viktor Bachratỳ, Marián Šimko, Pavol Balážik, Michal Trnka, and Filip Uhlárik. Slovakbert: Slovak masked language model. *arXiv preprint arXiv:2109.15254*, 2021.

[13] Marco Polignano, Pierpaolo Basile, Marco De Gemmis, Giovanni Semeraro, Valerio Basile, et al. Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets. In *CEUR workshop proceedings*, volume 2481, pages 1–6. CEUR, 2019.

[14] Faisal Rahutomo, Teruaki Kitasuka, Masayoshi Aritsugi, et al. Semantic cosine similarity. In *The 7th international student conference on advanced science and technology ICAST*, volume 4, page 1. University of Seoul South Korea, 2012.

[15] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[16] Patrick Schober, Christa Boer, and Lothar A Schwarte. Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia*, 126(5):1763–1768, 2018.

[17] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867, 2020.

[18] Jiaqi Yang, Yongjun Li, Congjie Gao, and Yinyin Zhang. Measuring the short text similarity based on semantic and syntactic information. *Future Generation Computer Systems*, 114:169–180, 2021.