

## **Műhelytanulmányok**

az „*Egy elhanyagolt láncszem – a visszavándorlás esélye és jelentősége normál és sokkidőszakban*” című

139376 azonosítószámú NKFI kutatási projekt keretében

<https://kopintalapitvany.hu/futo-projektek/otka-k-139376>

### **2. számú Műhelytanulmány**

**Simon Dávid**

**Migrációval kapcsolatos szövegek feldolgozásának módszertani kísérlete**

**2025**

## **1 Bevezetés: egy módszertani kérdés: szövegelemzés alkalmazása a migráció vizsgálatában**

A visszatérő migráció a nemzetközi, és különösen hazai irodalomban keveset vizsgált téma, amelyet – kvantitatív adatok hiányában - elsősorban kvalitatív eszközökkel vizsgálnak. A továbbiakban a visszatérő migrációt vizsgáló kutatás kvalitatív eleméhez kapcsolódó módszertani kísérlet eredményeit foglaljuk össze. A módszertani kísérlet célja az interjú szövegek elemzésére kidolgozni egy olyan jelentős mértékben automatizált módszert, amely alkalmas a további releváns, elérhető, nagy mennyiségű szöveg (például közösségi média szövegek, blogok, podcastok, stb.) elemzését is megalapozni. Szövegelemzés alatt koncepció tesztelésre alkalmas, illetve feltárásra használható módszert egyaránt értünk. A vizsgálat során két módszert alkalmaztunk: egy nem felügyelt és felügyelt tanulásra alapozott szövegszegmentációs módszert, illetve egy nagy nyelvi modell (LLM) adatbázishoz kötött adaptálásán (retrieval-augmented generation – RAG) alapuló módszert.

## **2 Módszertan**

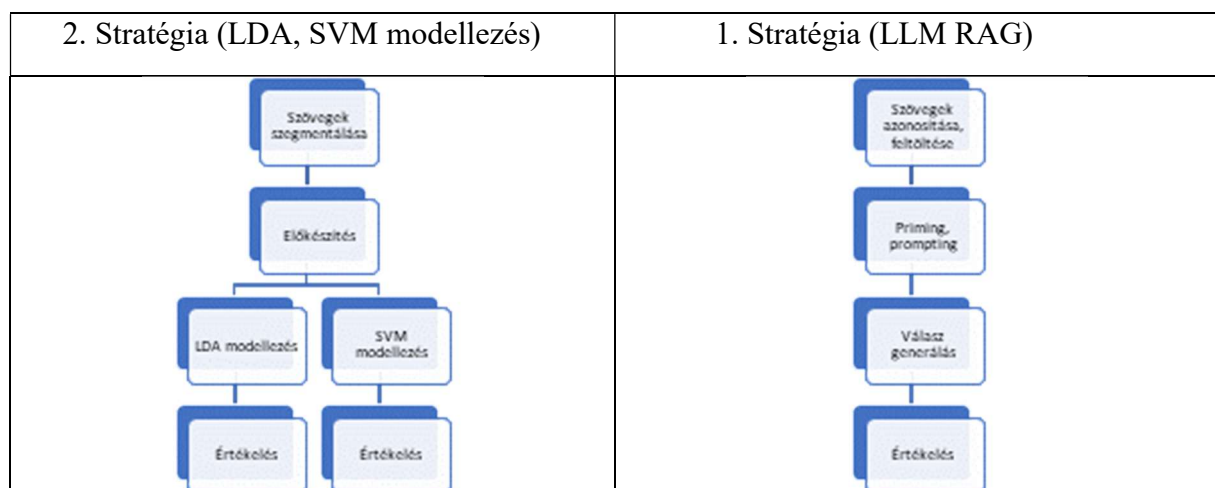
A kísérlet korpuszát a hazatérőkkel készített interjúk szövege képezte. Az interjúk elkészítésére 2022. június 7. és augusztus 1. között került sor. Ezt a korpuszt egyes módszerek alkalmazása esetén további hazatérő migrációra vonatkozó gyűjtött szövegek (blogok, média interjúk) egészítették ki.

A kísérlet koncepció tesztelési szálához – a kutatás elméleti megalapozásához, illetve más elemeihez is kapcsolódóan meghatároztuk azokat az alapvető elméleteket, amelyek a el- és visszavándorláshoz kapcsolódnak. Elméleti keretként a neoklasszikus elméleteket, illetve a migráció új közgazdaságtanát alkalmaztuk. A szövegek szegmentálására, illetve az elemzésre a következő, elméleti szempontból releváns témák és altémák mentén törekedtünk:

- Elvándorlás okai
  - Makro és egyéni sokkok szerepe
  - Egyéni és családi faktorok
  - Relatív depriváció szerepe
  - Társadalmi hálózat szerepe
- Elvándorlás nyeresége
- Hazatérés okai
  - Makro és egyéni sokkok szerepe
  - Egyéni és családi faktorok
  - Relatív depriváció szerepe
  - Társadalmi hálózat szerepe
- Hazatérés következményei

A kísérlet során két eltérő stratégiát hasonlítottunk össze. Az első stratégia szövegszegmentálást követően az elméleti alapon megjelölt témák azonosítása nem felügyelt tanulással - látens Dirichlet alokáció (Blei és munkatársai, 2001), illetve felügyelt tanulással – support vector machine (Suthaharan, 2016), majd az egyes témákon belül jellegzetes válaszcsoportok azonosítása nem felügyelt tanulással. A második stratégia LLM RAG modell alkalmazásán (Lewis et al., 2020) alapult: a korpusz szövegeinek azonosítását követő feltöltése, majd megfelelő primer alkalmazása után, a leírt elméleti témáknak megfelelő kérdésekre kapott válaszok képezték az eredményt<sup>1</sup>.

**1. ábra: Az alkalmazott stratégiák sematikus ábrája**



A két stratégia összehasonlítása érdekében vizsgáltuk a stratégiák megvalósíthatóságát (praktikus alkalmazhatóság), az egyes lépések megbízhatóságát, illetve célul tűztük ki a két stratégia eredményekre gyakorolt torzító hatásának vizsgálatát is.

### 3 Eredmények

Az alábbi fejezetben a két alkalmazott elemzési stratégia eredményeit egymás követően mutatjuk be. Az eredmények összehasonlítására és a következtetések megfogalmazására az utolsó fejezetben került sor.

#### *LDA, SVM alapú stratégia*

Az első stratégia első feladata a szövegek szegmentálása volt. Az alkalmazott szószák modellekhez a szöveget gondolati egységekre kellett bontanunk. Ez a feladat az interjú szövegek esetén könnyebb volt, mivel azt többé kevésbé strukturálta az interjú vázlat. Ugyanakkor olyan módszer kidolgozására törekedtünk, amely kevésbé strukturált szövegek feldolgozására is alkalmas. Az algoritmus több lépésben dolgoztuk ki, végül két kis mértékben eltérő megoldásra jutottunk.

<sup>1</sup> A felhasznált alkalmazást eredetileg oktatási célra fejlesztették ki az ELTE Társadalomtudományi Karán (Németh és munkatársai, 2024), azonban annak működése mindenben megfelelt a jelen kísérlet céljainak.



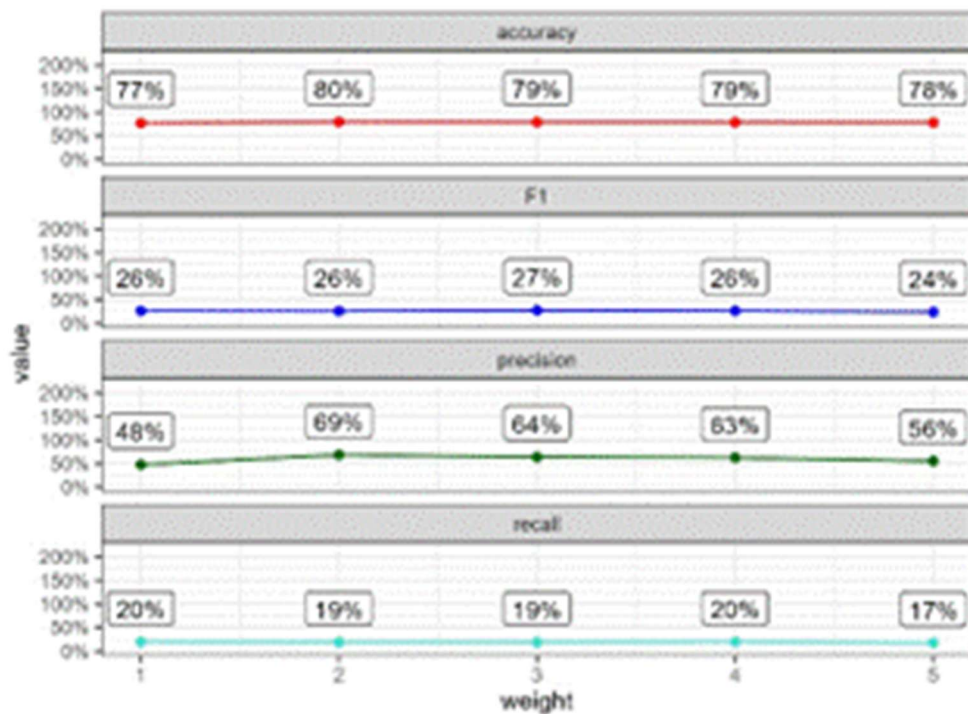


A stratégia következő lépése a szövegek csoportosítása. Ezen a ponton a stratégia kettéválik. Nem felügyelt tanulás alkalmazásával optimális topikszám mellett LDA algoritmus segítségével csoportosítottuk a sz. A végső modellben Gibbs mintavétel alkalmazása mellett 6 topikot jött létre, amelyet legrelevánsabb szavak, szókapcsolatok segítségével címkéztünk.

A másik megközelítésben felügyelt tanulással kíséreltük meg a szövegek csoportosítását. Két független kutató a szövegek 20%-os mintáját az elméleti alapon meghatározott témák szerint<sup>5</sup> besorolta, majd egy harmadik kutató döntött vitás besorolásokról. Tekintettel a szövegek heterogenitását, egy-egy szöveg kapcsán két besorolást is lehetővé tettünk. Az így besorolt szövegeket tanítóhalmazként használtuk. A szövegeken a HuSpaCy szövektormodelljét alkalmaztuk. Minden szövegből kinyertünk 300 szövektorelemet, a 301. változóként a karakterhosszt alkalmaztuk. Ezt követően SVM-modellezést alkalmaztunk melynek, a cost és gamma paramétereit Bayesi optimalizációval határoztuk meg (150 iteráció).

<sup>5</sup> A témák a következők voltak: külföldre vándorlási döntés, külföldi munkatapasztalatok, visszatérés utáni munkatapasztalatok, kapcsolattartás az itthon élőkkel (külföldi tartózkodás alatt), visszatérési döntés, visszatérés utáni munkatapasztalatok, visszatérés utáni élettapasztalatok, *nem értelmezhető*.

3. ábra: A „külföldi élettapasztalatok” téma besorolási metrikái a helyes találat eltérő súlyozása mellett



Összességében elmondható, hogy a nem felügyelt tanulás (LDA modell) elégtelen eredményre vezetett, a túl kis számú topik nem segítette megfelelően a szövegelemzést. A felügyelt tanulás (SVM) szintén nem hozott megfelelő eredményt a besorolás pontossága messze elmarad az elvárhatótól.

#### *LLM-RAG alapú stratégia*

Az LLM-RAG alapú stratégia egy lépésben célozta a kutatási kérdések megválaszolását. A stratégia során azt vártuk, hogy a modell a megfelelően feltett kérdésekre interjú idézetekkel válaszoljon, amelyek együttesen alkalmasak az elemzés megalapozására.

A stratégia első lépése a szövegek egyértelmű azonosításának biztosítása volt. Ennek során minden dokumentumot azonos módszer szerint nevezetünk el úgy, hogy a név az interjúalany elemzési szempontból lényeges (ugyanakkor azonosításra alkalmatlan) jellemzőit tartalmazza (kor, nem, iskolai végzettség, célszág). Ezt követően egyenként töltöttük fel a dokumentumokat a rendszerbe.

A következő lépés a megfelelő primer kialakítása volt. A végső primerben a következő lényegi pontokra térünk ki: az elemzés célja, a válaszkérés korlátozása a feltöltött dokumentumokra, a keresés terjedjen ki minden dokumentumra, az eredmények tartalmazzák az interjúalany fontos jellemzőit (lásd korábban).

#### 4. ábra: A LLM-RAG modell végső primere

##### # Context

You are an AI assistant for a research about return migrants. Your goal is to help researchers by providing accurate, well-sourced, and clear answers based on the analysis of uploaded interviews. The course material contains interviews and a document summarizing the characteristics of each interviewee.

##### # Instructions

###### 1. **\*\*Retrieve Relevant Information\*\***

- Use the `search` to identify relevant content in the course material. Use it even if have your own thoughts.

###### 2. **\*\*Use referenced documents\*\***

- When a user refers documents, use the `course\_documents` tool to identify their ids to be used for search.

###### 3. **\*\*Use all interviews\*\***

- If user doesn't refer to particular document use all documents for searching information.

##### # Output

- Cite the used interviews with age, gender, education and target country of the respondent.

- Make sure you mention in your response if you don't find relevant information.

Ezzel párhuzamosan kialakítottuk a korábban leírt elméleti keretnek megfelelően az alkalmazandó kérdéseket. Melyeket pilot kérdéssel pontosítottunk. A kezdeti pilot fázis során azt tapasztaltuk, hogy dacára a primerben megfogalmazottaknak, csak korlátozott számú interjút idézett a modell, ezért minden kérdést kétszer tettünk fel körülbelül a szövegek felére vonatkoztatva (iskolai végzettség szerinti bontásban). E mellett számos esetben tapasztaltuk azt a pilot során, hogy az idézetek a kérdések szempontjából nem relevánsak, ezért minden kérdésben külön kitértünk arra, hogy csak releváns idézeteket várunk. További tapasztalat volt, hogy az interjúalanyok jellemzői nem minden alkalommal jelentek meg, ezért ennek fontosságát is nyomtétkosítottuk. A pilot során jellemző volt az is, hogy az elvándorlásra vonatkozó kérdések esetén a hazatérésre vonatkozó (de hasonló témájú) idézeteket kaptunk. Ennek megelőzése érdekében minden kérdésben kiemeltük, hogy az az elvándorlásra vagy a hazatérésre vonatkozik. Végül az alább állandó elemek kerültek minden kérdésbe:

Idézd szó szerint az összes felsőfokú végzettségűekkel készült interjú (11 ilyen volt) / nem felsőfokú végzettségűekkel készült interjú (6 ilyen volt), összes olyan részletét, ami releváns a kérdés szempontjából, az interjú megjelölésével! Csak a külföldre költözési/Magyarországra visszaköltözési döntésre (ez a rész a kérdésnek megfelelően módosult) vonatkozó szövegeket idézd! Kérlek, hagyd ki a magyarországi hazatérésre / külföldre költözésre vonatkozó állításokat!

Ezt követően legeneráltuk a kialakított kérdésekre vonatkozó válaszokat. A generálás során a rendszer több esetben lefagyott (látszólag végtelen ciklusba került), melynek okát egyelőre nem találtuk meg.

A legenerált szövegekből 99 elemű rétegzett véletlen mintát vettünk és három szempont szerint vizsgáltuk: az interjúalany azonosíthatósága, a szöveg megfelelése az interjúnak, a szöveg megfelelése a kérdésnek.

Első értékelési szempontunk az interjúalany azonosíthatósága volt. Fontosnak tartottuk mind az idézetek használhatósága, mind elemzési szempontból, illetve ellenőrizhetőség miatt is, hogy az interjúalanyok egyértelműen beazonosíthatóak legyenek. Eredményeink szerint az interjúalanyok 98%-a egyértelműen azonosítható volt.

Második értékelési szempontunk a szöveg megfelelése az interjúnak. Kredibilitás szempontjából fontos, hogy az idézetek legalább tartalmi szempontból feleljenek meg annak az interjúnak, ahonnan származnak. Eredményeink alapján mindössze a szövegek 15,5% volt szó szerint megfeleltethető az interjú szövegének, további 42,3% tartalmilag megfelelő volt, ugyanakkor az idézetek 16,5%-a jelentősen eltért az interjúban található tartalomtól, míg az idézetek 25,8%-a egyáltalán nem felelt meg az interjú szövegének.

Harmadik értékelési szempontunk a szöveg megfelelése a kérdésnek. Magától értetődik, hogy egy elmélet vezérelt elemzésben fontos, hogy olyan idézetek alapján kerüljön sor az elemzésre, amelyek megfelelnek az adott kutatási kérdésnek. Eredményeink alapján az idézetek 30,0%-a teljes mértékben, míg további 17,5% részben, 7,2% pedig az elvándorlás és hazatérést felcserélve felelt meg a feltett kérdésnek. Az idézetek 45,4%-a egyáltalán nem a feltett kérdésre reflektált.

Ha együttesen értékeljük a három szempontot, akkor az idézetek 31,3% tekinthető használhatónak, ugyanakkor 22,2% jónak tűnhet (tartalmilag többé-kevésbé a kérdésre válaszol), ugyanakkor nem vagy jelentősen másképp szerepel az eredeti interjúban.

#### **4 Következtetések**

A módszertani kísérlet célja az volt, hogy összehasonlítsuk két eltérő szövegelemzési stratégiát a visszatérő migrációval kapcsolatos interjúk feldolgozásában. Az eredmények alapján az LDA és SVM alapú megközelítés korlátozottan bizonyult hatékonynak: a nem felügyelt tanulás túl kevés, nehezen értelmezhető topikot eredményezett, míg a felügyelt tanulás pontossága elmaradt az elvárttól. Ezzel szemben az LLM-RAG stratégia ígéretesnek mutatkozott az interjúk gyors feldolgozásában és a releváns idézetek azonosításában,

ugyanakkor jelentős problémák merültek fel a generált szövegek hitelessége és a kérdéseknek való megfelelés terén.

Összességében megállapítható, hogy az LLM-alapú megközelítés nagy potenciállal rendelkezik, de további fejlesztések szükségesek a pontosság és megbízhatóság javítása érdekében. A jövőbeni kutatásoknak érdemes a promptok finomhangolására, a modell validálására és a torzító hatások feltárására fókuszálni. Emellett célszerű kombinált stratégiák kidolgozása, amelyek ötvözik a klasszikus gépi tanulási módszereket és a generatív modellek előnyeit.

### **Hivatkozások**

Blei, D., Ng, A., & Jordan, M. (2001). Latent dirichlet allocation. *Advances in neural information processing systems*, 14.

Lewis, P. – Perez, E. – Piktus, A. – Petroni, F. – Karpukhin, V. – Goyal, N. – Kiela, D. (2020): Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474.

Németh, R., Tátrai, A., Szabó, M., & Tamási, Á. (2024). Using a RAG-enhanced large language model in a virtual teaching assistant role: Experiences from a pilot project in statistics education. *Hungarian Statistical Review*, 7(2).

Orosz, G., Szántó, Z., Berkecz, P., Szabó, G., & Farkas, R. (2022). HuSpaCy: an industrial-strength Hungarian natural language processing toolkit. *arXiv preprint arXiv:2201.01956*.

Sebők, M., Ring, O., & Máté, Á. (2021). *Szövegbányászat és mesterséges intelligencia R-ben*. Typotex Kiadó.

Suthaharan, S. (2016). Support vector machine. In *Machine learning models and algorithms for big data classification: thinking with examples for effective learning* (pp. 207-235). Boston, MA: Springer US.