Eötvös Loránd Tudományegyetem

Társadalomtudományi Kar

MESTERKÉPZÉS

Klaszterelemzés online depresszió fórumok bejegyzésein

A scatter/gather módszer alkalmazása szöveges adatokon

Konzulens:

Dr. Németh Renáta

Készítette:

Csala-Ferencz Bernadett

EIT05G

survey statisztika és adatanalitika szak

2021. április

Tartalomjegyzék

Táblázatjegyzék

Ábrajegyzék

1. A dolgozat főbb megállapításai

A depressziós megbetegedések egyre elterjedtebbek korunkban, az internetes fórumok pedig jó lehetőséget nyújtanak a mentális betegség természetének alaposabb megismerésére, és súlyosabb állapotban lévő, veszélyeztetett személyek kiszűrésére. Ez utóbbihoz névmáshasználati különbségek használhatóak fel eredményesen. A kutatásban angol nyelvű, depresszió témájú fórumokról gyűjtött 66295 bejegyzés klaszterezésével vizsgáltam, hogy milyen csoportokba szerveződhetnek a vizsgált bejegyzések. A depresszió fórumok megismerésén túl módszertani céljai is voltak a kutatásnak: egyrészt megvizsgálni, hogy a szövegek milyen előfeldolgozásával végezhető hatékonyabban klaszterezés a szövegeken, valamint, hogy a kiválasztott scatter/gather klaszterezési algoritmus mennyiben tudja segíteni jól interpretálható klaszterek megtalálását. Az elemzés során 15 viszonylag jól értelmezhető klaszter került létrehozásra, és elmondható, hogy az alkalmazott klaszterezési módszer többnyire hasznos eszközként szolgált jól interpretálható klaszterek elkülönítésére. A névmáshasználat alapján bár detektálásra került egy veszélyeztetettnek tűnő klaszter, azonban érdemes lehet még további markerek bevonásával is vizsgálni a súlyos depressziós bejegyzések detektálhatóságát.

2. A dolgozat felépítése

Először ismertetem a depresszió vizsgálatának az online fórumokon megmutatkozó lehetőségét, majd azt, hogy miért alkalmas ehhez a klaszterezés, mint feltáró elemzési módszer. Ezután tárgyalom a természetes nyelvfeldolgozás lehetséges lépéseit, valamint ismertetek különböző főbb klaszterezési eljárásokat. A kutatási előzményeknél részletezem a korábbi depresszió témájú fórumok vizsgálatának eredményeit, valamint azt, hogy szöveges adatokban hogyan detektáltak súlyosabb depressziós állapotra, vagy öngyilkossági veszélyre utaló jeleket a névmások használatán keresztül. Emellett még ismertetek egy összetettebb klaszterezési eljárást (scatter/gather módszer), mely hatékonyan alkalmazható szöveges adatokon is, és tárgyalom az ezzel kapcsolatban felmerülő ellentétes állaspontokat is.

A kutatási kérdések bemutatása utána a következő fejezetben először az adatokat ismertetem, valamint leírom a szöveges adatok előfeldolgozásánál alkalmazott lépéseket. Itt részletesen kitérek arra a döntésre, hogy milyen előfeldolgozottságú adatokon érdemes

végeznem az elemzéseket. Majd a scatter/gather módszer lépéseit végigkövetve ismertetem a klaszterezés folyamatát. Ebben az elemzés pontos leírása, és kapott eredmények leírása mellett, kitérek az elemzés során felmerülő dilemmákra is. Az adatokon futtatott hierarchikus klaszterezés tapasztalatairól is beszámolók, valamint ismertetem, hogy a végső klaszterekben milyen arányban szerepelnek a különböző névmások.

A záró fejezetben az eredmények értelmezése történik. Kitérek arra, hogy milyen nehézségek merültek fel az adatelemzés során, és ezek mennyiben korlátozhatják az eredmények érvényességét. Megvitatom a kutatási eredményekből leszűrhető tapasztalatokat, és ezek összevetésre kerülnek korábbi kutatási eredményekkel is. Itt érintem a szövegelemzési, klaszterezési, és súlyos depresszió detektálása kapcsán létrejövő eredményeket is. Végül felvetek további fejlesztési, kutatási lehetőségeket a téma kapcsán.

3. Háttér

3.1. A téma jelentősége

A mentális betegségek egyre elterjedtebbek a fejlett társadalmakban, jelentős terhet róva a bennük szenvedő egyénekre. Tartós szenvedést tekintve Európában az unipoláris depresszió az egyike a leginkább megterhelő betegségeknek (Spinney, 2009). Hidaka (2012) azt vizsgálta, hogy a depresszió hogyan válhatott a modernitás betegségévé. A kutató több megközelítésben is tárgyalja a lehetséges magyarázatokat: egyrészt a fizikális jóllét csökkenését sejti fő kiváltó oknak, ideértve a testi inaktivitást és az egészségtelen étkezést, másrészt a társadalmi környezet átalakulását, a fokozódó versengést, az egyenlőtlenségeket és az elmagányosodást.

Bár vannak feltevések és magyarázatok arra, hogy miért terjedhet el egyre inkább a depressziós megbetegedések jelensége, mégis nehezen férhetünk hozzá olyan kutatási igényességű adathoz, melyből többet is megtudhatunk a depresszióban szenvedő személyek élményvilágáról, gondolatai dinamikájáról. Különböző szociológiai kérdések vizsgálatára hagyományosan kérdőíves módszerek használata a legelterjedtebb módszer. A kérdőíves kutatási módszer széleskörben alkalmazott, annak ellenére, hogy számos nehézsége ismert a tudományos közegben (ld. Groves et al., 2011). Egy a teljes betegpopulációra reprezentatív minta összeállítása költséges, és nehézséges is a speciális alapsokasság miatt, és mivel a

mentális betegségek eleve szenzitív témák. Emellett kiemelendő az is, hogy a kérdőíves adatfelvétel kontrollált körülmények között történik, és ilyen mesterséges körülmények között félő, hogy a személyek válaszai is valamelyest mesterségesek, befolyásolva vannak az által, hogy a személy direkt kutatási céllal válaszol a kérdésekre. Ezek kiküszöbölésére érdemes lenne olyan adatforrást használni, mely természetesebb válaszokat eredményez, és könnyebben hozzáférhető.

Ahelyett, hogy kutatóként mi magunk próbálnánk adatokat generálni, felhasználhatunk korábban, „természetes" módon létrejött adatforrásokat is. Gondolhatunk itt például az internetes fórumokra, melyeket sokan használnak ahhoz, hogy többet tudhassunk meg a mentális betegségek természetéről. Az interneten elterjedtek az olyan fórumbeszélgetések, melyek direkt valamilyen betegség témája köré szerveződnek, így a depresszió kapcsán is sok ilyen fórumbeszélgetést találhatunk meg. Ezekben a fórumokban jellemző, hogy valamilyen mértékben érintett személyek jelennek meg, akik megoszthatják egymással tapasztalataikat és tanácsaikat, érintkezhetnek hozzájuk hasonló problémákkal küzdőkkel, és társas támogatást nyújthatnak egymásnak (Kummervold et al., 2002). Mivel pedig nagyon sok depresszióban érintett személy társalog ilyen fórumokon, ami a digitális világ elterjedésével, a digitális platformokon létrejövő közösségek elterjedésével állhat összefüggésben, az ezeken a felületeken folytatott beszélgetések alkalmas eszköznek bizonyulhatnak számunkra kutatóként ahhoz, hogy bár nem reprezentatív mintán, de célzottan a fórumokra nézve megvizsgálhassuk, hogy mi foglalkoztatja korunkban a depressziós személyeket.

Ez a kérdés annál is inkább fontos lehet, mivel egy súlyos depressziós állapot fokozott rizikót hordoz magában az öngyilkosságra nézve. Az öngyilkossági gondolatokkal, szándékkal küzdő személyek gyakran jelzik ilyen irányú terveiket, vagy utalnak szándékaikra, melyre alkalmas platformot kínálnak a depresszió témájú fórumok is (Horne és Wiggins, 2009). Így a depresszió témájú fórumok vizsgálatával arról is ismeretet szerezhetünk, hogy hogyan detektálhassunk öngyilkossági veszélyt magukban hordozó bejegyzéseket, melyre már történtek is korábban próbálkozások (pl. Cohan, Young és Goharian, 2016).

Az internet széles körű elterjedésével rengeteg adat került digitális közegbe. Ez az új helyzet új adatforrást is biztosít számunkra. Nagy mennyiségű adatforrás és adat vált számunkra könnyen elérhetővé. Elég csak a közösségi médiafelületekre gondolnunk, ahol intenzív

diskurzusok zajlanak különböző témákban. Egyes vélekedések szerint ezek az adatok nem használhatók elemzésekre, hiszen a digitális közeg zárt világában jönnek létre, amely nem vonatkoztatható az offline élettérre. Azonban tapasztalhatjuk és tudhatjuk is, hogy a digitális világ már nagymértékben beleivódott a mindennapjainkba. A közösségi felületeken létrejövő interakciók és cselekedetek általában azon kívül álló eseményekből erednek, vagy azokhoz kapcsolódnak, valamint ugyanúgy belehelyezkednek a társadalmi, politikai és gazdasági kontextusba, mint a digitális platformokon kívüli cselekedetek (Sloan, Morgan, Burnap és Williams, 2015). Természetesen ezek az adatok nem tekinthetőek reprezentatívnak semmilyen közösségi felületen kívüli populációra, azonban annak ismeretében, hogy a közösségi média felületeken is tükröződik a digitális világon kívüli társadalmi kontextus, érdemes ezen a felületen is vizsgálni különböző társadalmi jelenségek reprezentálódását. A közösségi médiafelületeken keletkező adatok elemzésével kiküszöbölhetők a kérdőíves vizsgálatok során keletkező torzítások, azonban felmerülnek másfajta torzítási lehetőségek helyettük (Hargittai, 2020). Ugyanakkor a közösségi média adatok gazdag forrást jelentenek, amennyiben a vizsgált kérdésnek megfelelően vannak feldolgozva.

Közösségi médiához sorolható minden olyan internetes felület, ahol a felhasználók interakciókba kerülhetnek egymással, közösségeket hozhatnak létre, és maguk is alakíthatják a felület tartalmát. Sokféle felület beleesik így a közösségi média fogalmába, például a közösségi háló építő oldalak (Facebook), microblogok (Twitter), médiamegosztó oldalak (Youtube) és a különböző tematikájú blogok és fórumok is (McCay-Peet és Quan-Haase, 2017). Az ilyen digitális felületeken, közösségi médiában megszülető adatok jellegzetessége, hogy elsősorban nem kutatási célokra vannak szánva. Bár az így létrejövő adatok sokkal zajosabbak és elemzésük kihívásokkal teli (ld. pl. Williams, Burnap és Sloan, 2017), viszont olyan előnyös tulajdonsággal is bírnak, hogy a személyek nem érzik azt, hogy kutatási adatgyűjtés céljából lennének megfigyelve.

A digitális adatokkal kapcsolatban felmerülő egyik probléma az etikai dilemmákra vonatkozik. Mit szólnak ahhoz a bejegyzésírók, hogy kutatási célokra használják fel az adataikat? Elterjedt ajánlás szerint a szabadon olvasható, regisztráció nélküli oldalak tartalmai felhasználhatóak kutatási célokra. Azonban fontos szempont ilyenkor az is, hogyha idézünk egy-egy konkrét bejegyzést a kutatás során, akkor annak az írója az interneten visszakeresve se legyen

pontosan azonosítható, biztosítva legyen mindenképp az anonimizáltság. Ennek ellenőrzése és biztosítása a kutató feladata (ESOMAR, 2011).

Az interneten rejlő adatok jelentős része szöveges formában van. Ilyenek a különböző cikkek, blogok, de a közösségi médiafelületeken található bejegyzések és a fórumbeszélgetések is. Ezek az adatok is ugyanúgy érdemesek az elemzésekre, mint a számszerű adatok, hiszen rengeteg információ rejlik bennük az emberek attitűdjéről, helyzetéről, véleményeiről, így jó lehetőséget kínálnak a szociológiai kérdések vizsgálatára. Ugyanakkor a hatalmas adatmennyiség miatt a hagyományosabb szövegelemzési eszközök, melyek a szövegek tényleges elolvasásán alapulnak, itt nem vehetőek számításba. Bár gondolhatnánk azt, hogy a szöveg kvalitatív adatforrás, melyet kvalitatív eszközökkel lehet csak megfelelően elemezni, ez a megközelítés nem helytálló. Kifinomult elemzési eszközökkel a szöveges adat számszerűsíthető, és így kvantitatív módszerekkel elemezhető. Ehhez figyelembe kell venni a szövegek sajátosságait, valamint a kutatási kérdésünk figyelembevételével kell számszerűsítenünk a szövegeket, de a megfelelő technikákat alkalmazva mindenképp végezhetőek kvantitatív elemzések a szövegeken. Az eredmények értelmezéséhez ugyanakkor lehetséges, hogy kvalitatív eszközöket is hasznosítanunk kell. Lehetséges, hogy ez túl sok energiabefektetésnek és túlságosan bizonytalan eredménynek tűnik, de nem az! A szövegelemzési technikák fejlesztésével, finomításával és a gyakorlat elterjedésével olyan adatforrás nyílik meg az emberiség számára, amely könnyen és gyorsan elérhető, és „ingyen” áll rendelkezésünkre.

A közösségi médiaoldalakon megszülető szövegek strukturálatlan adatokat eredményeznek, melyek meglehetősen zajosak. Ahhoz, hogy az elemzési algoritmus helyesen tudja elemezni a szöveget, értenie kell, hogy mely szavak tartoznak össze kifejezésként, milyen rövidítési formái léteznek egy-egy kifejezésnek, és mely szavak azok, amelyek bár különböző formában szerepelnek toldalékolás miatt, mégis egy tőről erednek. Emellett az is fontos, hogy mely szavak bírnak valódi önálló tartalommal, és melyek azok, amelyeknek csak toldalékolási, szintaktikai szerepük van. A szövegek feldolgozásához, és statisztikai elemzésre alkalmassá tételéhez a természetes nyelvfeldolgozás (natural language prosessing – NLP) eszközeit vehetjük segítségül, melyekről a következő fejezetben írok részletesebben.

A probléma körüljárásában a gépi tanulási technikák lehetnek a leghatékonyabb eszközök. A gépi tanulás (machine learning) egy olyan terület az adattudományban, amely az adatokból való modellépítéssel foglalkozik. Olyan modelleket igyekszik létrehozni, amellyel jobban megérthetőek az adatok. Ezzel a technikával nagy adatbázisban is kereshetünk csoportosulásokat, struktúrát az adatainkban (VanderPlas, 2017, 332. o.). Az elemek csoportokba rendezéséhez két különböző gépi tanulási megközelítésből indulhatunk ki. Az egyik esetben rendelkezünk előzetes elképzelésekkel, információkkal arról, hogy milyen csoportok lelhetők fel az elemek halmazában, és ezekbe a csoportokba szeretnék besorolni a lehető legnagyobb hatékonysággal (azaz legkisebb hibaaránnyal) az elemeinket. Ezeket nevezzük konfirmatív módszereknek. Más esetekben viszont nincsenek előzetes elképzeléseink arról, hogy milyen csoportokba rendeződhetnek az elemeink, vagy nem szeretnénk élni ilyen feltevésekkel. Ilyenkor feltáró módszereket alkalmazhatunk a csoportok felderítéséhez. Ilyen módszer a klaszterezési eljárás is, amelyben az elemek jellemzői alapján a számítógép rendezi a lehető leghomogénebb csoportokba az elemeket. Az eljárás célja az, hogy az ugyanabba a csoportba sorolt elemek a lehető leginkább hasonlítsanak egymáshoz, és a lehető leginkább különbözzenek a más csoportba sorolt elemektől. Valójában nem egyféle bevett módszerről, hanem algoritmusok szerteágazó csoportjairól beszélünk, melyek különböző gyakorlati megközelítéssel igyekeznek elkülöníteni csoportokat.

Azért érdemes a depresszió témájú fórumokat feltáró technikákkal vizsgálni, mivel bár lehetnek előzetes elképzeléseink róla, hogy milyen témák jelennek meg egy ilyen fórumon, mégsem ismerünk minden felmerülő témát. A feltáró módszerekben induktív logika mentén haladunk (Ignatow és Mihalcea, 2018, 74. o.). Először feltevések nélkül kezdjük elemezni a korpuszt, majd a kapott eredmények alapján fogalmazunk meg elméleti konklúziókat a vizsgált témában. Azért hasznos számunkra ez a megközelítés, mert ha konfirmatív módszerekkel elemeznénk a korpuszt, akkor csak az általunk választott szempontok mentén tudnánk kategóriákba sorolni, fennáll azonban a lehetősége, hogy egyes csoportokat így nem vennénk figyelembe. Ennek kezelésében lehet segítségünkre a klaszterezés.

3.2. Elterjedt fogalmak meghatározása

3.2.1. Natural language processing eszközei

Ahhoz, hogy a szöveget, mint strukturálatlan adatot kvantitatív módszerekkel tudjuk elemezni, előzetes lépésként alkalmassá kell tennünk arra, hogy az algoritmus futhasson rajta. Ahhoz, hogy a szöveg a különböző klaszterezési eljárások által feldolgozható formába kerüljön, a természetes nyelvfeldolgozás (natural language processing, a továbbiakban: NLP) eszközeit használjuk fel. A következőkben ezeken a lehetséges lépéseken haladunk végig.

A szöveget a legegyszerűbb megközelítésként kezelhetjük úgy, mint szavak gyűjteményét, melyben az nem is számít, hogy a szavak milyen sorrendben szerepelnek egymás után a szövegben, csak az a fontos információ, hogy a szöveg mely szavakat tartalmazza, és azokat hányszor. Ezt nevezhetjük szózsák megközelítésnek (ld. Aggarwal és Zhai, 2012). Itt minden szöveget egy-egy vektor reprezentál, melyben a számok a szöveg szavainak gyakoriságát mutatják. Az, hogy milyen sorrendben követik egymást a szavak, teljesen tetszőleges lehet. Ezekből a vektorokból már egy közös mátrixba rendezhetőek egy korpusz dokumentumai. A mátrix soraiban és oszlopaiban a korpusz dokumentumai és a teljes korpusz szótára (illetve kifejezései) találhatóak meg, a cellákban pedig az, hogy adott dokumentumban hányszor szerepel az adott szó. Ebből adódóan egy meglehetősen nagy, de ritka mátrixot kapunk, melyet kifejezés-dokumentum mátrixnak (term-document matrix) nevezünk. Ez azt jelenti, hogy sok nulla szerepel a mátrixban, mivel a legtöbb szó a korpuszszótárból csak a dokumentumok egy részében szerepel. Ennek a mátrixsajátosságnak a kezelésére külön figyelmet kell fordítani. Bár ez egy meglehetősen leegyszerűsítő megközelítés, mégis meglepően hatékonyan alkalmazható korpuszok elemzésére például olyan problémakörben, mint hogy mérhessük a korpusz dokumentumainak hasonlóságát egymáshoz (Turney és Pantel, 2010).

A legtöbb előfeldolgozási lépés a szózsák megközelítésen alapszik. A feldolgozás első részeként történik a tokenizálás, azaz a szövegnek, mint karakterek sorozatának a szóbeli egységekre bontása. Ehhez mankóul szolgálnak a szóközök, valamint az írásjelek is, melyek a folyamat közben eltávolításra is kerülnek, azonban a feladat nem annyira egyértelmű, hiszen az írásjelek nem feltétlen jelentenek szóhatárokat (pl. kötőjeles szavak), melynek kezelésére külön figyelmet kell fordítani (ld. Ignatow és Mihalcea, 2018, 122. o.). Ahhoz, hogy az elemzés során

ne legyen különbség téve az egyes szavak alakváltozatai, ragozott formái között, kétféle technikát alkalmazhatunk. Az egyik eljárásban (stemming) különböző nyelvspecifikus átalakítási szabályokat felhasználva levágásra kerülnek a toldalékok és előtagok, és így próbálja meg a szótöves alakra hozni a szavakat az algoritmus (Porter,1980). Azonban ezek a szavak gyakran nem szótári szavak, mivel az alkalmazott szabályok nem általánosíthatók kivétel nélkül minden esetre, valamint gyakran előfordulhat, hogy akkor is levág toldaléknak tűnő szóvégeket, amikor valójában nem kéne, és ezzel olyan szavakat hoz közös bázisalakra, amelyek valójában nem ugyanarra vonatkoznak. Ennél finomabb eljárás a lemmatizálás, melyben a toldalékolt, ragozott szavakat ténylegesen a szótári alakra fordítja át az algoritmus. Ez olyan szótárak felhasználásával történhet, amelyek ismerik a kivételt képező szavak különleges formáit is (ld. Bird, Klein és Loper, 2009, 108. o.).

A fenti módszerektől eltérve, ahol csak szavak szintjén elemeztük a dokumentumokat, míg eltekintünk a szavak sorrendjétől, vannak olyan egyszerűbb eljárások, amelyekkel valamelyest megragadhatók a szavak szövegbéli helyzete is. Ilyen módszer a szignifikáns bigramok és trigramok detekciója, melyben nem külön-külön kezeljük a szavakat, hanem olyan szópárosításokat, vagy akár hármas csoportokat keresünk, amelyek gyakran fordulnak elő egymást követve a szövegben (ld. Bird, Klein és Loper, 2009, 141. o.). Gyakran használt eszköz emellett még a névelemfelismerés (named entity recognition), melyben valamilyen konkrét entitásra utaló szavak kerülnek felismerésre (pl. szervezetek vagy személyek nevei). Amellett, hogy az ezeket alkotó szavakat is fontos egy egységként kezelni, még az is cél, hogy az entitás különböző névformáit is felismerje, és azonos alakra hozza az algoritmus. Ehhez használhatóak automatizált megoldások (Collins és Singer, 1999).

A korpuszban mindig szerepelnek olyan szavak, melyek túl gyakoriak, és emiatt nem bírnak akkora diszkrimináló erővel a bejegyzésekre nézve, hiszen természetükből adódóan általában minden bejegyzésben szerepelnek. Ilyenek lehetnek nyelvspecifikusan a névmások, kötőszavak, vagy más funkciószavak, melyek nem segítik elő a szövegek diszkriminálását, de megnövekszik miattuk a zaj a korpuszban. Ezeknek a kezelésére több eszköz is rendelkezésünkre áll. Egyrészt használhatunk úgynevezett stopszó listákat az ilyen szavak eltávolításához, melyek direkt nyelvspecifikusan tartalmazzák az ilyen szavakat. Ugyanakkor, ha klasszikus listák alapján távolítjuk el a stopszavakat, akkor vigyázni kell, hogy ne távolítsunk el olyan szavakat, melyek értékes információul szolgálhatnak a kutatási kérdés vizsgálatához

(pl. Saif, Fernandez, He és Alani, 2014). Hasonlóan robosztus eljárás az is, ha a saját adatbázisunkból kiindulva távolítjuk el azokat a szavakat, amelyek a bejegyzések adott százalékánál többen is szerepelnek. Ehhez hasonlóan eltávolíthatjuk a túlzottan ritka szavakat is, melyek alig szerepelnek egy-két dokumentumban. Ez azért lehet hasznos, mert annak ellenére, hogy amely szavak csak arányaiban nagyon kevés dokumentumban szerepeltek, azok nem hordoznak kellő információt magukban a dokumentumok differenciálásához. Elég sok ilyen szó lehet az adatbázisban, amelyeknek a feldolgozása így feleslegesen lassítja a későbbi elemzési algoritmusokat.

Finomabb módszerként alkalmazhatók olyan eljárások, amelyek nem csak az alapján rendelnek számot a szavakhoz, hogy milyen gyakran szerepelnek az adott dokumentumban, hanem figyelemben veszik a szónak a teljes korpuszban megjelenő gyakoriságát is. Ilyen módszer a tf-idf (term frequency and inverse document frequency: kifejezés-gyakoriság és inverz dokumentum-gyakoriság) súlyozás. Ez abból áll, hogy a kifejezés bejegyzésbeli gyakoriságát megszorozzuk egy olyan tényezővel, amely inverze a kifejezés összes dokumentumban megmutatkozó gyakoriságával, vagyis inverze annak, hogy a kifejezés hány dokumentumban szerepelt összesen. Ez az eljárás visszasúlyozza az olyan szavak gyakoriságát, melyek a korpusz sok dokumentumában szerepelnek, míg nagyobb súlyt hagy azoknak, amelyek kevés dokumentumban szerepelnek. Így azokat a szavakat látja el legnagyobb értékkel, melyek kevesebb dokumentumban fordulnak elő, de az adott dokumentumban gyakrabban megjelentek (ld. Aggarwal and Zhai, 2012).

3.2.2. Klaszterezés

A klaszterezés egy olyan eljárás, amely a vizsgált elemeket a köztük lévő hasonlóság alapján rendezi azonos csoportokba, klaszterekbe. Az algoritmus arra törekszik, hogy az egyes klasztereken belüli elemek közel legyenek egymáshoz, míg a különböző klaszterekhez tartozó elemek távol legyenek egymástól. Különböző szöveges dokumentumok klaszterezésénél általában vektortérmodellből készítjük el a dokumentumok klaszterezését. Az alkalmazott vektortérmodell csupán azt reprezentálja, hogy a bejegyzésekben milyen szavak milyen gyakran, vagy mekkora súllyal szerepelnek, de arról nem tartalmaz információt, hogy a szavak milyen kapcsolatban állnak egymással. Ilyen reprezentáció tulajdonképpen a fentebb említett tf-idf súlyokból létrejövő mátrix is. A klaszterezés nem egy konkrét eljárás, hanem inkább

algoritmusok összessége, melyek különböző altípusokba sorolhatóak. Két ilyen fontos altípus a partícionáló és a hierarchikus módszerek csoportja.

A partícionáló módszerekben közös, hogy nem hierarchikus klaszterstruktúrát építenek fel az adatokból, hanem előre megadott számú klaszterbe rendezik az elemeket, törekedve arra, hogy a klaszterek a lehető legtávolabb legyenek egymástól, és minden elem a hozzá legközelebb eső klaszterbe kerüljön. Az egyik legelterjedtebb ilyen módszer a k-közép (k-means) algoritmus (ld. Aggarwal és Zhai, 2012). Ez az algoritmus adott számú kezdő középpontot jelöl ki a vektortérben, majd minden mintaelemet hozzárendel a hozzá legközelebbi középponthoz. Ezután a hozzárendelt mintaelemekből újraszámolja a középpont helyét, majd újra hozzárendeli az elmozdított középpontokhoz azokat az elemeket, amelyek hozzájuk vannak a legközelebb. Ez az iterációs folyamat addig zajlik, míg már nem történik (vagy csak nagyon apró, elhanyagolható mértékű) változás a középpontokban. Ehhez az algoritmushoz meg kell adnunk előre, hogy mennyi klasztert szeretnénk, azaz hány darab kezdőpontból induljon az eljárás, azonban ezen kívül nem kell semmilyen előzetes feltevéssel élnünk az adatok struktúrájáról, eloszlásáról. Ugyanakkor előnye, hogy meglepően kevés iteráció alatt lefut az algoritmus. Hátrányos azonban, hogy az algoritmus eredménye függ attól, hogy honnan lettek indítva a kezdőpontok. Ennek kiküszöbölése általában már implementálva van a számolófüggvényekbe, de egy sajátos megoldási lehetőséget ismertetni fogok a későbbiekben (4.2. fejezet).

A klaszterezés futtatása előtt fontos megválasztani, hogy milyen módon definiáljuk két dokumentum távolságát egymástól. Bár a kifejezés-dokumentum mátrix hatékony forrás arra, hogy megállapítsuk két dokumentum hasonlóságát egymáshoz, az mégsem evidens, hogy milyen módon mérjük le ezt a hasonlóságot, hiszen több metrika is rendelkezésünkre áll. A nagy és sok dimenziójú adatbázisokban pedig döntő jelentőségű, hogy milyen távolságfogalmat használunk fel. Míg egyik távolságmetrikát felhasználva két elemet közelinek mérhetünk, másik metrika alapján ugyanazon két elem akár távoli is lehet. A távolságmetrika kiválasztásánál fontos figyelembe vennünk, hogy milyen szempontból szeretnénk mérni az elemek hasonlóságát vagy távolságát egymástól.

Klaszterezésnél általánosságban az egyik leggyakrabban használt távolságmetrika az euklidészi távolság. A $p = \{p_1, p_2, ...p_n\}$ és $q = \{q_1, q_2, ...q_n\}$ vektorok euklidészi távolsága n dimenziós térben a következő:

$$y(p,q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_i^{n=1}(q_i - p_i)^2}$$

Ez alapján a dokumentumok mindig ahhoz a klaszterközépponthoz kerülnek hozzárendelésre, amelytől a legkisebb az euklidészi távolságuk, vagy azok a dokumentumok kerülnek összevonásra, melyek vektorainak a legkisebb egymástól az euklidészi távolsága (Huang, 2008). Azonban így túlreprezentálásra kerülhetnek a hosszú dokumentumok, amik több szót tartalmaznak, és így az őket reprezentáló vektorok is hosszabbak lesznek. Szövegelemzési esetekben ezért kifejezés-dokumentum mátrixok esetén a leghasznosabb a koszinusz hasonlóság alkalmazása (Dhillon and Modha, 2001), mely a két vektor közötti szög koszinuszát méri:

$$cossim(x,y) = \frac{x^T \cdot y}{\|x\| \cdot \|y\|}$$

Ez az érték 1 és -1 közé eshetne, de mivel a kifejezés dokumentum mátrixok nem tartalmaznak negatív értékeket, ezért ezekben az esetekben 0 és 1 közé eshet az érték. Ebből fakadóan két vektor koszinusz távolsága megkapható akként, ha 1-ből kivonjuk a koszinusz hasonlóságot. A koszinusz hasonlóság gyakran alkalmazott távolságmetrika szózsák típusú szövegelemzéseknél, hiszen előnyös tulajdonsága, hogy nem befolyásolja az, hogy egy dokumentum milyen hosszú, csak a benne szereplő szavak aránya. Ez úgy értendő, hogyha *A* dokumentum egy mondatból áll, *B* dokumentum pedig ugyanebből a mondatból csak háromszor megismételve, akkor a koszinusz hasonlóság szerint a két dokumentum teljesen hasonló, hiszen pontosan ugyanabba az irányba mutatnak a dokumentumokat reprezentáló vektorok. Ezzel szemben az euklidészi távolság különbséget mérne a dokumentumok között, hiszen az a vektorok végpontjai alapján számolja ki a távolságot. Így a koszinusztávolsággal jobban mérhetők a szövegek közötti tényleges különbségek, melyet azok a mérési tapasztalatok is alátámasztanak, melyekben azt vizsgálták, hogy ismert kategóriájú szövegeket mely távolságmetrikával lehet klaszterezéssel a leghomogénebb csoportokba sorolni (Subhashini and Kumar, 2010).

A partícionáló módszerek mellett a másik nagy csoportja a klaszterezési eljárásoknak a hierarchikus módszereké. Ezekben az eljárásokban az elemek egy faszerű klaszterhierarchiába épülnek fel, mely az egyes különálló elemektől az összes elem egy, közös klaszterbe sorolásáig tart. Ez a hierarchikus klasztersturktúra dendogramon ábrázolható, melynek ágrajzából látható, hogy milyen sorrendben kapcsolódnak össze az egyes elemek, vagy csoportok, és azt is, hogy ezek milyen távol voltak egymástól. A hierarchikus klaszterezés tovább bontható két alesetre: az agglomeratív (összevonó) és felbontó eljárásokra. Agglomeratív klaszterezésnél kezdéskor minden elem egy-egy külön klaszterbe van sorolva, majd ezután lépésenként egybevonásra kerülnek az egymáshoz legközelebb álló elemek, illetve klaszterek. Mindez addig folytatódik, míg egy nagy klaszterbe sorolódik minden elem. A felbontó eljárások hasonló logikával, de épp ellentétes irányba működnek: minden elem kezdetben egy közös klaszterben van, majd a klaszter lépésenként feldarabolódik úgy, hogy az új klaszterek a lehető legtávolabb legyenek egymástól. Ez addig folytatódik, míg minden elem külön klaszterbe kerül. Mindkét eljárás ábrázolható dendogramon (Popat, Deshmukh és Metre, 2017). A hierarchikus eljárás előnye, hogy robosztusabb eredményt adhat, mint a k-közép klaszterezés, mivel összehasonlítja az összes dokumentumpárt. Azonban az eredménye nagyban függhet a helyes távolságmetrika megválasztásától.

Bár az elemek közötti hasonlóság megállapítására már ismertettünk lehetséges megoldásokat, a hierarchikus eljárásnál azt is meg kell határoznunk, hogy hogyan definiáljuk két klaszter távolságát. Talán az egyik legegyszerűbben érthető eljárás a legközelebbi szomszéd módszere (single linkage), melyben azt a két klasztert kapcsoljuk össze, melyeknek a két legközelebbi eleme között a legkisebb a távolság a többi klaszter legközelebbi elemeihez képest. Bár egyszerűen értelmezhető logikájú algoritmus, ritkán alkalmazott, mivel aránytalan méretű csoportokat hozhat létre, és gyakori a láncalkotás (chanining) jelensége, vagyis ugyanahhoz az egyre nagyobb klaszterhez kapcsol újra és újra egy-egy elemet vagy kisebb klasztert, folyton ugyanazt a klasztert táplálva egyre nagyobbra. Ezzel szemben a legtávolabbi szomszéd módszere (complete linkage) azt a két klasztert keresi meg, melyeknek az egymástól legtávolabb eső elemeik a legközelebb vannak a többi klaszter egymástól legtávolabb eső elemeihez képest. Ez a módszer kevésbé érzékeny az adatok közötti zajra és a kiugró értékekre, és így stabilabb klaszterezést képest létrehozni, mint a legközelebbi szomszédok módszere (Hubert, 1974). A két módszer ötvözése az átlagos csoportok közötti páronkénti

távolság módszere, mely azt a két klasztert kapcsolja össze, amelyek elemei közötti távolság átlagosan a legkisebb. Ez egy szintén viszonylag robosztus módszer, és a legtávolabbi szomszéd módszerével szemben jobban képes figyelembe venni a klaszterstruktúrát. Szövegek klaszterezésénél megvizsgálva a leghatékonyabbnak a csoportok közötti átlagos páronkénti távolság mutatkozik, utána pedig a legtávolabbi szomszéd módszere (El-Hamdouchi és Willett, 1989). Ezek a módszerek arra a távolságmetrikára támaszkodnak, amelyet az elemek közötti távolság megállapítására is használunk.

A fentiek mellett vannak olyan módszerek, amelyek bár képesek viszonylag robosztus és egyenletes méretű klaszterek detektálására, csak euklidészi térben és távolságban értelmezett adatokon alkalmazhatóak. Ilyen például Ward módszere mely a csoportokon belüli és csoportok közötti négyzetösszegek arányán keresztül állapítja meg, hogy mely összevonással kapható a leghomogénebb klaszterezés. A centroid módszerrel pedig a klaszterközéppontok (súlypontok) távolságát veszi figyelembe (ld: Everitt, Landau, Leese és Stahl, 2011).

Szükséges lehet ellenőrizni azt, hogy a kapott hierarchikus rendezés, a dendogram mennyire reprezentálja jól az elemek közötti tényleges távolságokat. Erre használható a kofenetikus (cophenetic) korrelációs együttható (Sokal és Rohlf, 1962), mely mérni képes, hogy mennyire illeszkedik jól az adatokra az adott hierarchikus klaszterezés. Ehhez a kofenetikus távolságokat használja fel, melyek azt adják meg, hogy a hierarchikus klaszterezésben két elem a dendogram mely szintjén (mely távolságnál) kerül először egy objektumba. Az ebből kialakult kofenetikus távolságmátrixot hasonlítja a kofenetikus korreláció az elemek eredeti távolságmátrixához. Az eredmény minél közelebb van 1-hez, annál jobban közelíti a hierarchikus klaszterező felosztás a valódi távolságokat.

Nehézséget jelenthet a klaszterezés módszerénél az ideális klaszterszám megtalálása, azaz, hogy milyen számú klaszter megállapítása reprezentálja legjobban az adatokban ténylegesen meghúzódó struktúrát. Ehhez segítségünkre lehet egyrészt az, ha megpróbáljuk interpretálni a klasztereket. Amennyiben jól és egyértelműen interpretálható klasztereket kapunk, akkor biztosabbak lehetünk abban, hogy valóban látensen meghúzódó csoportokat sikerült feltárni. Az ideális klaszterszám megtalálását segítheti, ha megvizsgáljuk egy dendogram alapján, hogy mennyire távoli csoportok kerültek összevonásra. Amennyiben itt találunk egy olyan szintet, ahonnan kezdve láthatóan nagyobb távolságú klaszterek kerültek egybevonásra, mint

korábban, akkor ideális vágás lehet ebben a pontban megállítani a klaszterezést. Rendelkezésünkre állnak emellett ugyanakkor egzaktabb mérőszámok is, melyekkel felderíthető a legoptimálisabb klaszterszám. Ilyen például a Silhouette-együttható (Rousseeuw, 1987), mely azt vizsgálja, hogy melyik modellben jöttek létre jobban meghatározott klaszterek. Az egy elemhez tartozó mutatót a következő képlet alapján kaphatjuk meg:

$$s = \frac{b - a}{max(a, b)}$$

ahol *a* az átlagos távolság az elem és az azonos klaszterbe tartozó elemek között, *b* pedig az átlagos távolság az elem és a legközelebbi klaszter elemei között. A Silhouette együtthatót az elemekhez tartozó értékek átlagaként kaphatjuk meg. Ez az együttható -1 és 1 közötti értéket vehet fel. 1-es értékhez közel azt láthatjuk, hogy kifejezetten sűrű, és egymástól távol eső klasztereket kaptunk, 0 értéknél nagyon átfedő klasztereink vannak, míg -1 felé eső értéknél az elemek helytelenül lettek besorolva, közelebb esnek más klaszterekhez, mint a sajátjukhoz. Egy értelmes klaszterezést várhatóan 0.5-nél nagyobb Silhouette-együttható jelez, míg 0 körüli együttható arra utalhat, hogy nincsen valódi detektálható klaszterstruktúra az adatokban.

Másik elterjedt és jól működő mutató a Calinski-Harabasz index (Caliński & Harabasz, 1974), mely a csoportok közötti (B_k) és a csoporton belüli (W_k) diszperziós mátrix nyomából számolandó ki a következőképpen:

$$s = \frac{tr(B_k)}{tr(W_k)} * \frac{n_E - k}{k - 1}$$

ahol n_E a teljes elemszám, k pedig a klaszterek száma. Minél nagyobb a mutató értéke, annál jobban definiált klasztereink vannak. Ennek az indexnek hátránya azonban, hogy csak euklidészi térben értelmezendő távolságokra alkalmazható.

Fontos megjegyezni még, hogy milyen módon határozzuk meg a klaszterek középpontját. Az euklidészi távolságok alapján létrehozott klasztereket úgy optimalizáljuk, hogy a bejegyzéseket reprezentáló vektorok végpontjai a lehető legközelebb legyenek egymáshoz egy klaszteren belül, míg a többi klaszterhez tartozó vektorok pontjaitól a lehető legtávolabb essenek. Így a klaszterek középpontját is az egy klaszterbe tartozó pontok súlypontjaként

kaphatjuk meg a vektortérben. Koszinuszhasonlóság-alapú klaszterezésnél azonban úgy optimalizáljuk a klasztereinket, hogy a bejegyzéseket reprezentáló vektorok irányát vesszük figyelembe. Így a vektor hossza irreleváns információ számunkra. Így, ha az ilyen módon kialakított klaszterek középpontját határoznánk meg, akkor először normalizálni kell a vektorokat a saját hosszukkal, hogy egység hosszúságúak legyenek. Majd az így kapott vektorokat klaszterenként összegezni kell. Az így kapott vektor a vektorok iránya alapján értelmezett klaszter középpontjában fut. A vektor hossza valójában itt is irreleváns, csak az iránya számít, de ahhoz, hogy elegánsabb megoldást kapjunk, ezt a vektor is normalizálhatjuk a hosszával, hogy egység hosszúságú legyen.

A klaszterezésnél nagy kihívás a kapott eredmények értékelése, a klaszterek értelmezése. Fontos szem előtt tartani, hogy a klaszterezés mindig ad eredményt, azonban nem biztos, hogy ez az eredmény hasznos, „értelmes”. Ezért nagy erőfeszítéseket kell fordítani arra, hogy értékeljük azt, hogy valóban értelmezhető, létező csoportokat kaptunk-e eredményül. A kérdés megválaszolása klaszterezés esetén nem triviális, hiszen nincsen birtokunkban semmilyen a priori ismeret arra vonatkozóan, hogy milyen csoportok léteznek, és melyik elem melyik csoport tagja (Müller és Guido, 2017, 134. o.). Nehezítő körülmény, hogy míg két- vagy háromdimenziós vektortérben elhelyezkedő elemek ábrázolásával és a klasztertagságok jelölésével a vizuális információkból szubjektív ítéletet lehetne meghozni arról, hogy a klaszterek valóban az elemek egy-egy sűrűsödési pontját fedik le, azonban sokdimenziós vektortérben reprezentálódó elemeknél már nem lehetséges az ábrázolás, és így a klaszterezés eredményének szemrevételezése. A klaszterezés eredménye objektíven nem megítélhető, hiszen nem áll rendelkezésünkre információnk arról, hogy mik a valódi csoporttagságok. Sokféle megoldás létezik annak mérésére ugyanakkor, hogy valamelyest a valódi csoporttagságok nélkül is meg tudjuk állapítani, hogy volt sikeres a csoportok létrehozása. Egyrészt alkalmazhatjuk a fentiekben ismertetett Silhouette-együtthatót a klaszterezés jóságának megállapítására, valamint megvizsgálhatjuk azt is, hogy a klasztertagok közötti távolságok alapján mennyire homogén klasztereket kaptunk.

4. Előzmények

4.1. Depressziós fórumok vizsgálata

Fontos megjegyezni elsőként, hogy a depresszió témájú fórumokon nem csak depressziós személyek jelennek meg. Míg a legtöbbjük valóban küzd valamilyen súlyosabb vagy enyhébb mértékű depressziós problémával, semmi sem biztosítja azt, hogy minden bejegyzésíró depresszióban szenved, illetve megjelenhetnek olyan személyek is, akik ismerősükkel kapcsolatban kérnek segítséget, vagy osztanak meg információt. Ezt figyelembe kell venni a fórumokra írt bejegyzéseknél, hiszen nem tekinthetjük úgy, hogy a fórumokon megjelenő személyek a valódi diagnózissal rendelkező depressziósok csoportját reprezentálják.

A bejegyzések írói között különböző fórumbeli viselkedésformák detektálhatók. Egy survey alapú felmérés alapján (Nimrod, 2013) vannak olyan személyek, akik a mindennapi élet aggodalmairól számolnak be, ők jellemzően valóban depressziós egyének. A kevésbé depressziós tünetekkel élők között azonban sokan inkább csak információt keresnek a fórumokon. Vannak emellett olyan fórumtagok is, akik mindegyik fórumtopik iránt érdeklődnek, ők jellemzően nagymértékben használják a fórumokat, valamint olyan tagok is megjelennek, akik nem igazán vannak bevonódva a fórumbeszélgetésekbe.

Nimrod (2012) 9 fő témát talált, melyek köré leginkább szerveződnek a depressziós fórumok közösségei: tünetek, kapcsolatok, megküzdés, élet, formális ellátás, gyógyszerek, okok, öngyilkosság és munka. Sik (2020) online fórumokon végzett etnográfiájában szintén elkülönített különböző tematikájú bejegyzéscsoportokat. Egyrészt megjelenik a depresszió testi-orvosi megközelítéséhez tartozóan az orvoslás ellentmondásossága és a depresszió testi okai. Sok diskurzus szól emellett a pszichológiai vonatkozásáról a depressziónak: automatikus gondolatok, alkalmazkodási képesség romlása és a személyiség eltorzult fejlődése, traumák, személyes hibázások. Emellett azonban megjelenik a depresszió társadalmi diskurzusa is, mind a társadalmi elvárások, mind a megbetegítő társadalmi környezet tekintetében. Németh, Sik és Máté (2020) depresszió témájú fórumok bejegyzéseit épp e 3 keretezési mód (biomedikális, pszichológiai és szociológiai) szerint igyekeztek klasszifikálni a természetes nyelvfeldolgozás eszköztárával. Úgy találták, hogy míg a biomedikális és a pszichológiai keretezés elég jól azonosítható, a szociológiai keretezés kevésbé egyértelműen ragadható meg nyelvi markerek segítségével, melyet az magyarázhat, hogy ez utóbbi még kevésbé elterjedt keretezési forma.

Németh, Sik és Katona (2021) topikmodellezéssel különítettek el különböző témákat depressziós fórumok bejegyzésein. Többféle topikszám mellett arra a következtetésre jutottak, hogy a bejegyzések egyik része jellemzően monológ típusú, míg másik részük inkább interakciós formájú. A monológokon belül vannak olyan bejegyzések, melyek oktulajdonítások fejtegetésére fókuszálnak. Ezek lehetnek egészségügyi, párkapcsolati, családi, munkahelyi vagy akár iskolai problémákon alapuló attribúciók is. A monológok másik nagy típusai inkább valamilyen bemutatások az életükről vagy önmagukról. Ezek lehetnek szenvedésmonológok, küzdelemnaplók, de tartalmaznak a jóllétükről szóló beszámolókat is. A bejegyzések másik nagy halmaza interakciós formájú bejegyzéseket tartalmaz. Egy részük konzultációs típusú akár szerekről, testi terápiákról. Megjelennek laikus szakmai konzultációk, melyek egy részében megjelenik biomedikális vagy pszichológiai perspektíva is. Ezek mellett azonban vannak kvázi terápiás interakciók is, melyekben vannak tapasztalatalapú segítségnyújtások, vallási vigasztalások, és szó esik bennük a mentális problémák médiamegjelenéséről is.

Az öngyilkosság, mint tematika is tetten érhető a depressziós fórumok bejegyzéseiben, hiszen a depresszió az egyik legnagyobb veszélyeztető tényező az öngyilkosságra nézve. Bár elterjedt nézet, hogy „aki beszél róla, az úgysem teszi meg”, ez valójában nem helytálló, mivel a legtöbb öngyilkosságot megkísérlő személy valamilyen módon utal szavaival, vagy tetteivel arra, hogy mire készül. A fórumok alkalmas platformot nyújthatnak arra, hogy a személyek valamilyen módon jelezzék öngyilkossági szándékukat. Horne és Wiggins (2009) vizsgálatukban úgy találták, hogy a fórumokon öngyilkosság témájában beszélő személyek gyakran történetszerűen ábrázolják helyzetüket, és jellemző rájuk annak hangsúlyozása, hogy élet és halál határán állnak pszichésen. Erős törekvés mutatkozik arra, hogy meghatározzák az öngyilkosság veszélyeztető jeleit annak érdekében, hogy könnyebben detektálható legyen az öngyilkossági veszély. Egy megközelítés szerint 10 ilyen jelet lehet elkülöníteni a szakmai tapasztalatok és kutatások alapján (ld. Lester, McSwain és Gunn III, 2011). Ezek az öngyilkosságról való beszéd, a szerhasználat, a céltalanság, a harag, a csapdábaesettség, a reménytelenség, a visszahúzódás, a szorongás, a meggondolatlanság és a hangulatingadozás. Azonban ezek a jelek nem bizonyultak alkalmasnak arra, hogy szöveges feljegyzésekből detektálhatóak legyenek az olyan személyek, akik ténylegesen komoly veszélyben vannak (Lester et al., 2011). McSwain, Lester és Gunn III (2012) elemzéséből, amelyben fórumbejegyzéseket vizsgáltak, úgy mutatkozott, hogy az öngyilkosságról való beszéd, a

céltalanság, a csapdábaesettség, a reménytelenség és a visszahúzódás valóban jelezhetik, ha valakit öngyilkosság veszélyeztet. Ezek azonban kvantitatív módszerekkel nehezen ragadhatóak meg a szöveges adatokból, a fenti kutatásokban is inkább kvalitatívan értékelték a bírálók a szövegek tartalmát.

Többek is vizsgálták, hogy a depresszió hatással van-e a személy nyelvhasználatára, felfedezhetőek-e olyan nyelvi markerek, melyek specifikusan súlyosabb depressziós állapotra jellemzőek. Hiszen amennyiben találhatóak ilyen jól detektálható markerek a nyelvhasználatban, amelyek jelezhetik súlyos depressziós állapot, vagy akár megnövekedett öngyilkossági rizikó létét, akkor kvantitatív elemzéssel is elkülöníthetőek az ilyen szövegek. Kutatási eredmények alátámasztják, hogy az öngyilkosságot megkísérlő személyeknek csökkent a bevonódása a társadalomba, visszahúzódnak a szociális ügyektől, ezzel párhuzamosan pedig inkább befelé, önmagukba fókuszálnak (Breault és Barkey, 1982). Ez súlyos depressziós állapotban is jellemző lehet. Az állapotot detektálni lehet a nyelvben a névmások használatán keresztül, és nem csak a beszélt, hanem az írott nyelvben is megjelenik a tendencia. Stirman és Pennebaker (2001) öngyilkosságot elkövetett költők verseit vetette össze más költők verseivel. Úgy találták, hogy az öngyilkosságot elkövető költők verseikben többször használtak egyes szám, első személyű névmásokat („I”, „me”, „my”), mely mutathatja a fokozott önmagukra fókuszálást. Ezzel szemben kevesebbet használtak többes szám, első személyű névmásokat („we”, „us”, „our”), amely pedig a társas és személyes kapcsolatoktól való visszahúzódást jelezheti. Így a névmások használata hasznos prediktora lehet az öngyilkosságnak, szemben például a pozitív és negatív érzelmi szavak használatával, melytől bár intuitívan hasonló jelzőképességet várhatnánk, a kutatásukban mégsem mutatkozott különbség bennük az öngyilkos és nem öngyilkos költők között. Rude, Gortner és Pennebaker (2004) egyetemistákat vizsgálva úgy találták, hogy a depressziós egyetemisták is több egyes szám, első személyű névmást használnak, azonban ez a hatás az „én” („I”) névmás használatában mutatkozott meg, a ragozott változataiban („me”, „my”, „myself”) nem. Ez azzal magyarázható, hogy a névmás ragozatlan formája inkább a szelfre való utalás, de a ragozott formái egy mondaton belül már gyakran valamilyen társas kapcsolatban jelennek meg. Azt is megfigyelték a szövegekből, hogy a depressziós személyek több negatív érzelmű szót használnak. Ugyanakkor annak nem találták jelét, hogy a depressziós egyetemisták kevesebb társas referenciát alkalmaznának történetmesélésükben, mint a nem depressziósak.

Bernard, Baddeley, Rodriguez és Burke (2016) kutatásában szintén úgy találták, hogy a depressziós személyek több „én” („I”) szót használnak írott szövegeikben, és kevesebb „ő” („he”, she”) szót, azonban nem találtak kapcsolatot a depresszió és a negatív vagy pozitív érzelmi töltetű szavak használatában. Ugyanakkor azt is vizsgálták, hogy a névmások használatára nincs-e ugyanakkora hatással az aktuális érzelmi állapot, mint a depresszió. Úgy találták, hogy amennyiben sikeresen manipulálják negatív irányba a kutatás elején a személy érzelmeit, akkor sem növekszik meg az általuk írott szövegeikben az egyes szám első személyű névmás használata, viszont a negatív érzelmi szavak használata igen. Ebből feltételezhetjük, hogy míg a névmások használatában valóban a depressziós állapot, és akár öngyilkossági veszély tükröződhet, a negatív szavak használata inkább az aktuális érzelmi hangulatot tükrözi, mintsem tartós depressziós állapotot.

4.2. Scatter/gather módszer

Amellett, hogy milyen tényezőkre érdemes fókuszálni a depressziós fórumok bejegyzéseinek klaszterezésénél, döntést kell hozni arról is, hogy pontosan milyen klaszterezési algoritmust érdemes használni a probléma kezelésére. A k-közép klaszterezésnek és a hierarchikus klaszterezésnek is megvannak az előnyei és a hátrányai, melyeket már korábban ismertettem (3.2.2. fejezet). A két módszer előnyös tulajdonságainak ötvözésére Cutting, Karger, Pedersen és Tukey (1992) mutattak be egy megoldást, melyet scatter/gather módszernek neveztek.

A scatter/gather módszer először hierarchikus agglomeratív klaszterezési algoritmussal keres ideális klaszterkezdőpontokat, majd ezekből a kezdőpontokból indít a teljes mintán klaszterezést. A kezdőközéppontok meghatározásához használt hierarchikus klaszterezést (un. Buckshot eljárást) a teljes mintának egy kisebb részén futtatja, ezzel kiküszöbölve azt, hogy a sokszoros elemtávolságok megállapításával túlzottan lelassuljon a hierarchikus algoritmus. Ehhez a szerzők javaslata alapján $\sqrt{n \cdot k}$ méretű mintát érdemes használni, ahol *n* a teljes minta elemszáma, *k* pedig a kívánt klaszterszám. Tekinthetünk erre úgy is, hogy ezzel a megoldással nem véletlenszerűen választunk ki *k* db kezdőpontot, hanem helyette $\sqrt{n \cdot k}$ darab kezdőpontot választunk, és ezeket csökkentjük vissza *k* darabra agglomeratív módon klaszterezve őket. Egy ekkora almintán megfelelően futtatható hierarchikus klaszterezés ahhoz, hogy a teljes mintára jellemző struktúrát fedezhessünk fel, ugyanakkor az alminta elég kicsi ahhoz, hogy gyorsan fusson rajta a hierarchikus klaszterezés. A kívánt klaszterszámnál

kapott klaszterek középpontjai pedig robosztus kezdőközéppontokat adnak a k-közép algoritmus futásához, hiszen így értelmes kezdőpontokból indíthatjuk az algoritmust, ezzel csökkentve annak a hatását, hogy a kezdőpontok kiválasztásának esetlegessége miatt adott kezdőpontválasztásoknál nem sikerülhet megtalálni a valódi klaszterstruktúrát. Bár az alminta véletlenszerű kiválasztása miatt ez az eljárás sem determinisztikus, de a különböző almintákon futtatott hierarchikus elemzés kvalitatívan jellemzően hasonló felbontást ad (Cutting et al., 1992).

A kiválasztott kezdőpontokból már a teljes mintán futtatható k-közép klaszterezés, hiszen ez az algoritmus jelentősen gyorsabban és kisebb memóriaigénnyel képes futni ugyanazokon az adatokon, mint egy hierarchikus klaszterezési algoritmus. A teljes mintás k-közép klaszterezés eredményét azonban még tovább javíthatjuk a scatter/gather módszerrel. Egyrészt tovább-bonthatunk egyes klasztereket, hogy részletesebb klasztereket kaphassunk. Ehhez egy kismintás hierarchikus klaszterezést végezhetünk a klaszter elemein, majd a kétklaszteres megoldásból eredő középpontokból indítva a klaszter összes elemén végzünk k-közép klaszterezést. Olyan esetekben érdemes ezt alkalmazni, ha a klaszter tartalma kevésbé koherens, nehezen interpretálható. A szerzők a klaszterkoherencia méréséhez olyan mutatókat javasolnak, melyek az elemek hasonlóságát nézik a klaszterközépponthoz és egymáshoz. Emellett, ha klasztereket túlzottan hasonlónak ítélünk meg az őket leginkább jellemző, legnagyobb gyakorisággal vagy súllyal rendelkező szavaik alapján, akkor össze is vonhatjuk őket egy közös klaszterbe.

A módszert eredetileg olyan célból ismertették a szerzők (Cutting et al., 1992), hogy böngészési keresések találatát rendezze klaszterekbe a lehető leggyorsabban, majd a felhasználónak lehetőséget adjon kiválasztani az őt legjobban érdeklő klasztereket, melyeket így újrarendezett az algoritmus, annak érdekében, hogy pontosabb találatokat adjon. Látható, hogy eredetileg is szöveges adatokra tervezték a módszert, és a cél épp az volt, hogy a sokdimenziós adatok ellenére mégis gyors eredmények legyenek kaphatóak, miközben még valamilyen mértékű hierarchia is kialakul a klaszterek összevonásával és tovább-bontásával. Később igyekeztek továbbfejleszteni a módszert, hogy még gyorsabban működhessen online keresési találatok rendezésére. Liu, Mostafa és Ke (2007) amellett érveltek, hogy érdemesebb lenne futtatni előzetesen a teljes mintán egy hierarchikus agglomeratív klaszterezést, majd ennek az eredményeiből építeni fel a különböző számú rendezéseket, összevonásokat,

tovább-bontásokat. Míg Cutting és munkatársai (1992) szerint, ha csak az elemek egy részén futtatunk újra klaszterezést, az más és pontosabb eredményt ad, mint ami a teljes klaszterezésből levonható lenne, Liu és munkatársai (2007) szerint ez egy nem alátámasztható állítás, ezért felesleges újraklaszterezést végezni, különböző sűrűségű felbontásokért elég csak a hierarchikus klaszterezés struktúrája alapján dönteni. Ugyan mindkét szerzőcsoport bemutat a saját módszerükre jól működő példát, Liu és munkatársai (2007) közvetlenül nem cáfolják meg munkájukban az újraklaszterezések hasznosságát. Csak a saját módszerüket mutatják be, de nem vetik össze a példájukban a Cutting és munkatársai (1992) által javasolt módszerrel elkészített klaszterezéssel, hogy alátámasszák azt az állításukat, hogy a két módszer eredménye érdemileg nem különbözik egymástól.

4.3. Saját kutatási kérdések

Hipotézis: A klaszterek egy részének újrabontásával kapott eredmények eltérnek a teljes mintából létrehozott hierarchikus klaszterezéstől. Az egyes klaszterek újrabontásával jobban interpretálható eredményeket kapunk.

Cutting és munkatársai (1992) szerint érdemes újrabontani pár kiválasztott klasztert, mert így pontosabb felbontást kaphatunk, ha lecsökkentjük a teljes mintát olyan klaszterekre, amelyek valamilyen szempont mentén hasonlóak. Liu és munkatársai (2007) szerint ez az újrabontás azonban nem hoz jelentősen eltérő eredményt attól, mint amit a teljes mintán futtatott hierarchikus klaszterezésből kapnánk, amelyből egy teljes klaszterstruktúrát kaphatunk meg.

Kutatási kérdés: Milyen tartalmi különbségek fedezhetőek fel a kialakult klaszterekben? Milyen kapcsolatban állnak ezek az azonos bejegyzéseken létrehozott topikmodell eredményével?

A depresszió témájú fórumok elemzésekor hasonló, de nem megegyező tartalmi struktúrákat tártak fel. Fontos kérdés, hogy az általunk feltárt struktúra mennyiben egyezik meg a szakirodalomban leírt struktúrákkal. Ugyanakkor érdemes összevetni a topikmodellezésből származó eredményekkel (Németh et al., 2021), hiszen valamelyest hasonló témáknak a klaszterelemzésben is fel kéne merülnie, mint amik az azonos bejegyzéseken kialakult topikokban is vannak.

Hipotézis: A létrejövő klaszterekben eltérés mutatkozik a névmáshasználat tekintetében. Súlyosabb állapotra utaló tartalommal rendelkező klaszterekben magasabb lesz az egyes szám, első személyű („I") névmás használata, és alacsonyabb az egyes szám, harmadik személyű („he", „she"), valamint a többes szám első személyű („we") névmás használata.

Ezek a különbségek abból fakadnak, hogy a depressziós személy inkább befelé fordul, és visszahúzódik a társas kapcsolatoktól, amely a névmások használata által is detektálható korábbi kutatások alapján.

5. Elemzés és eredmények

5.1. Adatok és előfeldolgozásuk

Kutatási kérdéseim vizsgálatához a RC^2S^2 (Research Center for Computational Social Science) kutatócsoport által létrehozott szövegkorpuszt használtam fel (ld: Németh et al., 2020). A korpusz a jelen dolgozatban felhasznált formájában 66 295 fórumbejegyzés szövegét tartalmazta. Ezek a bejegyzések a legnépszerűbb angol nyelvű egészségfórumokról kerültek begyűjtésre (pl. healthunlocked.com, depressionforums.org) SentiOne használatával. Az adatok gyűjtése megfelelt a GDPR szabályozásoknak. A korpuszban olyan szabadon hozzáférhető bejegyzések szerepelnek, melyek 3 éves időtartamban, 2016. február 15. és 2019. február 15. között születtek. A begyűjtött posztok mindegyikében szerepel legalább egy depresszióval kapcsolatos kifejezés a következők közül: depression, depressed, bummer, desolation, desperation, moody, upset, gloom, hopelessness, depressant, melancholia, sorrow, unhappiness, feeling blue, depressive, depressive disorder, unipolar depression, bipolar, bipolar depression, major depression, mdd, persistent depressive disorder, pdd, cyclothymia, mood disorder, adjustment disorder, chronic fatigue syndrome, cfs, premenstrual dysphoric disorder. A korpuszban csak olyan bejegyzések szerepelnek, melyek legalább 20 szóból állnak. A kutatásomhoz kapcsolódó elemzéseket Python 3.7-ben végeztem el.

Németh és munkatársai (2020) nyomán a bejegyzésekhez létrehozásra került egy olyan feldolgozás, melyben eltávolításra kerültek a duplikátumok, törlésre kerültek az URL-ek és email-címek, lemmatizálták a szavakat és megkeresték a szignifikáns bigramokat. Ebből a feldolgozási szintből kiindulva több verziót is létrehoztam az adatbázisból. Fő különbségként

a stopszavak eltávolításánál időztem el. Mivel a szakirodalmi bevezetőben ismertetett eredmények alapján a névmások jelentős információt hordozhatnak magukban a depresszió súlyosságára nézve, készítettem egy olyan verziót, melyben az NLTK (Natural Language Toolkit) stopszólistából kiindulva csak azokat a stopszavakat távolítottam el, amelyek nem névmások. Készítettem azonban egy olyan verziót is emellett, amelyben eltávolítottam a stopszó listán szereplő névmásokat is, mivel a túlzottan magas gyakoriságuk miatt lehetséges, hogy zavarják az elemzéseket. A két verziót még tovább variáltam a tekintetben, hogy a kutatócsoport eljárása nyomán elkészítettem belőlük egy olyan verziót, amelyben minden olyan szót eltávolítottam, amely a bejegyzések kevesebb, mint 1/10000-ben (azaz kevesebb, mint 7 bejegyzésben) szerepel, emellett pedig készítettem az olyan verziót is, amelyben azok a szavak is el lettek távolítva, amelyek kevesebb, mint a bejegyzések 1/1000-ben (azaz kevesebb, mint 67 bejegyzésben) szerepelnek. Azért éltem ezekkel az eltávolításokkal, mert a nagyon ritka szavak nem segíthetik lényegileg a bejegyzések eltávolítását egymástól, azonban jelentős zajjal járhatnak az elemzésekben. A teljes szótár stopszavazás után 133 ezres nagyságrendű volt(a stopszavazás módjától függően), míg az 1/10000-es eltávolítás után 18 ezres nagyságrendűre csökkent a korpusz szótárának mérete. Bár már ez is jelentős dimenziócsökkentést eredményezett, érdemes lehet megvizsgálni, hogy a még nagyobb mértékű, 1/1000-es eltávolítással kapott 4 ezres nagyságrendű szótáron végzett elemzéseken már érződik-e információveszteség hatása. Amennyiben ezen a kisebb szótáron sem kapunk rosszabb eredményeket, érdemes lehet inkább ezt használni, mivel a további dimenziócsökkentés jelentősen gyorsíthatja a további elemzések futási idejét. A bejegyzések számszerűsítéséhez, a klaszterezések futtatásához tf-idf súlyozással hoztam létre adatmátrixot.

5.2. Kismintás hierarchikus klaszterezés

A klaszterezést a korábban ismertetett scatter/gather módszer lépésein végighaladva végeztem el. Elsőként létrehoztam egy kis méretű almintát, melyen hierarchikus klaszterezést tudtam futtatni. Cutting és munkatársai (1992) alapján ezt a Buckshot hierarchikus klaszterezést $\sqrt{n * k}$ méretű mintán kellene elvégezni (*n*: elemszám, *k*: klaszterszám). Azonban mivel nincsen előzetes elképzelésünk arról, hogy hány klaszterünk lehet pontosan a korpuszban, inkább csak egy felső becsléssel élünk ebben a helyzetben a klaszterszámra. Így minden-bizonnyal megfelelően nagy mintán futtatjuk a Buckshot klaszterezést, és majd a

klaszterezés eredménye alapján kereshetünk ideális klaszterszámot a későbbiekhez. Mivel Németh és munkatársai (2021) a korpusz topikmodellezésénél 18 topikot különítettek el, és 20 csoportnál több értelmezése korlátokba is ütközhet a feladat kvalitatív jellegéből adódóan, ezért felső becslésként 20 klasztert állapítottam meg, mivel ennél jelentősen több klaszter technikai korlátok miatt semmiképp sem lehet majd. Így mivel 66295 bejegyzésből áll a korpusz, ezért 1152, véletlenszerűen kiválasztott bejegyzésen végeztem el a Buckshot klaszterezést. A klaszterezésben úgy jártam el, hogy az teljes mintából megalkotott tf-idf súlyozásos vektorokból választottam ki az alminta elemeihez tartozókat. Azért tettem így, és nem külön az almintán alkottam újra a tf-idf súlyozást, mert egyrészt az almintában kisebb a szótár mérete, így kisebb dimenziójú vektorokat kapnánk, másrészt a vektorértékek sem egyeznének meg az eredetiekkel. Így egy teljesen más vektortérben reprezentálnánk a bejegyzéseinket, mint ami a teljes mintából származik.

A Buckshot klaszterezést agglomeratív hierarchikus klaszterezésként definiáltam. Mivel szövegek klaszterezésénél a kutatási tapasztalatok alapján a koszinusz hasonlóság használata sokkal hatékonyabb, mint az euklidészi távolságé, ezért az elemek közötti távolságot koszinusz hasonlóságként határoztam meg, míg a klaszterek közötti távolságot átlagos csoportok közötti páronkénti távolságként számoltattam ki. Bár a Ward módszer hasznos eszköz lehet ahhoz, hogy egyenletes elemszámú csoportokat kapjunk, azonban nem alkalmazható együtt koszinusz hasonlósággal, csak euklidészi hasonlósággal interpretálható. Koszinusz hasonlósághoz azonban hasonlóan jó választás lehet az átlagos csoportok közötti páronkénti távolság Cutting és munkatársai (1992) nyomán.

Előzetes lépésként megvizsgáltam, hogy a 4 különböző feldolgozású adatbázisból melyiken teljesít legjobban a választott hierarchikus klaszterezési algoritmus. Úgy találtam, hogy a kofenetikus korreláció alapján azokon az adatbázisokon teljesít legjobban a hierarchikus klaszterezés, amelyeknél a stopszavazás a névmások megtartásával történt meg, ezen belül pedig valamelyest nagyobb a kofenetikus korreláció annál, amelynél a dokumentumok kevesebb, mint 1/10000-ében szerepelnek (1. táblázat). Ezek alapján tehát a névmások megtartásával létrehozott adatbázisokon jobban tudja reprezentálni a hierarchikus klaszterezés az eredeti távolságokat.

A klaszterezés eredményének megvizsgálásakor azonban felfedeztem, hogy bár ez inkább a legközelebbi szomszéd módszerénél kéne, hogy jellemző legyen, mégis az átlagos csoportok közötti páronkénti távolságmetrikánál is megjelent a láncalkotás (chaining) jelensége, vagyis a hierarchikus lépésekben az algoritmus egy nagy klaszterhez kapcsolt hozzá kisebb, akár csak egy elemből álló klasztereket még a dendogram csúcsán is. Emiatt más, koszinusz hasonlósághoz választható távolságmetrikákkal is leellenőriztem, hogy milyen eredményt kapunk a klaszterek elemszámát tekintve. A legközelebbi szomszéd módszere várható módon ugyanígy a láncalkotás problémáját mutatta, azonban a legtávolabbi szomszéd módszerével valamelyest egyenletesebb klaszterelemszámokat kapunk. Bár itt is megfigyelhető, hogy a dendogram tetején egy nagyobb elemszámú klaszterhez kerülnek fokozatosan hozzáadásra a kisebb elemszámú klaszterek, azonban itt többnyire nagyobb elemszámú klaszterekkel van dolgunk, mint az átlagos csoportok közötti páronkénti távolság használatánál. Ugyanakkor hiába tűnik az legtávolabbi szomszéd módszere jobban teljesítő távolságmeghatározásnak a klaszterek között, a kofenetikus korrelációs együtthatót nézve gyengébb eredményt ad, mint az átlagos csoportok közötti páronkénti távolság. Leellenőriztem mind a 4 feldolgozási verziójú adatbázison, és mindegyiken hasonló eredményeket kaptam mind a kofenetikus korreláció (1. táblázat), mind a láncalkotás tekintetében.

Koszinusz távolság	Névmások megtartása és < 1/10000-ben szereplő szavak eltávolítása	Névmások megtartása és < 1/1000-ben szereplő szavak eltávolítása	Teljes stopszavazás és < 1/10000-ben szereplő szavak eltávolítása	Teljes stopszavazás és < 1/1000-ben szereplő szavak eltávolítása
Átlagos csoportok közötti páronkénti távolság	0,6728	0,6445	0,5522	0,5312
Legtávolabbi szomszéd módszere	0,3846	0,4027	0,2186	0,2479

1. táblázat: Kofenetikus korrelációs együtthatók

Mivel azonban az semmiképp sem szolgál hasznos információval számunkra, ha olyan hierarchikus klaszterezésből próbálunk következtetéseket levonni a további klaszterezési folyamatra nézve, amelyben csak egy klaszter van folyamatosan növelve egy-egy következő elemmel, ezért inkább a legtávolabbi szomszéd módszerével meghatározott klaszterstruktúrából igyekeztem továbbépítkezni.

Tovább vizsgálva tehát a legtávolabbi módszer használatából származó dendogramokat, kitűnt, hogy az utolsó kapcsolási távolságok mind ugyanakkora (1-es) kapcsolási távolságokat adnak. Különben is megfigyelhető, hogy a kapcsolási távolságoknál elég magasak a távolságok már a hierarchikus összevonások kezdetén is, kevés olyan elem van, amelyek viszonylag közel lennének egymáshoz. Ez lehetséges azért is, mivel nagyon sok változónk van, több mint az elemszám, így a sok dimenzióban nehezen fordulhatnak elő olyan esetek, amelyek jelentősen hasonlítanánk egymásra. Azonban ez a probléma remélhetőleg a későbbiekben orvosolódni fog, amikor már teljes mintán kerül futtatásra a klaszterezés, hiszen akkor már magasabb is lesz az elemszám, mint a változók száma, és nagyobb eséllyel lesznek a sokdimenziós vektorok között olyanok, amelyek a sok dimenzió mentén is jelentősebb hasonlóságot mutatnak.

Felvetette bennem a kérdést az utolsó kapcsolási távolságok megegyezésénél, hogy mi alapján kerülnek kezelésre azok az esetek, amelyekben több klaszter is ugyanakkora távolságra van egymástól. Fernández és Gómez (2008) alapján ennek kezelésére többféle megoldás is rendelkezésre állhat, azonban nincsen egységes megoldási mód rá, valamint nincsen szakmai konszenzus sem a tekintetben, hogy milyen algoritmus tudna a legstabilabb, és hatékonyabb megoldást találni a problémára. Az általam használt scipy modul dokumentációját átvizsgálva nem találtam információt arra nézve, hogy ebben az általam alkalmazott szoftverben hogyan kerülnek kezelésre az ilyen „döntetlen” helyzetek. Bár magam sem tudok megfelelőbb módszert javasolni a szoftverbe az ilyen helyzetek kezelésére, megpróbáltam valamelyest megvizsgálni, hogy mennyire kapunk robosztus klaszterezést. Ehhez az 1152 elemű minta sorrendjét 10-szer random megkevertem, majd mindegyikhez kiszámoltam, hogy ha 20 klaszteres eredményt kérek, akkor koszinusz hasonlóság és legtávolabbi szomszéd módszere mellett mennyire robosztus az egyes elemek közös klaszterbe sorolása. Páronkénti χ^2 statisztikák alapján vizsgáltam meg, hogy a klaszterbesorolások függetlenek-e egymástól. A kapott eredményeket a 2. táblázat mutatja.

Szignifikáns p-értékek százaléka	Névmások megőrizve	Névmások eltávolítva
>0,0001 ritka szavak eltávolítása	57,77%	100%
>0,0001 ritka szavak eltávolítása	42,22%	100%

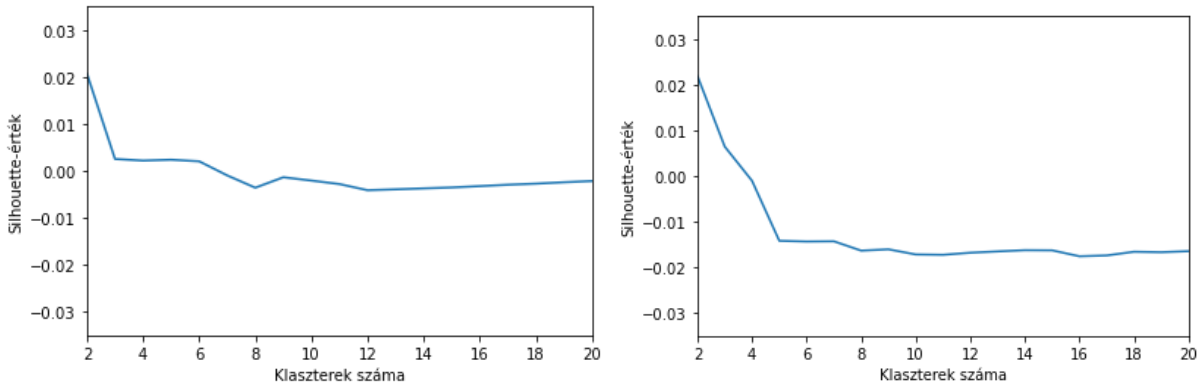
2. táblázat: Klaszterbesorolások robosztusságának ellenőrzése

Azt látjuk, hogy míg azoknál az adatbázisoknál, melyeknél a névmások nem lettek eltávolítva a szótárból a kapott p-értékeknek nagyjából a fele szignifikáns ($p<0,05$), azaz ezekben az esetekben a besorolások nem függetlenek egymástól, míg az esetek másik felében a tesztstatisztikák szerint az besorolások függetlenek egymástól. Annál az adatbázisnál, amelynél több ritka szó került megőrzésre, a besorolások 57,77%-a nem független egymástól, míg a másik adatbázisban 42,22%-a nem független a besorolásoknak. Ezen eredmények szerint a sokszor egyenlő klasztertávolságok miatt a hierarchikus klaszterezés nem ad kielégítően robosztus eredményt. Ugyanakkor azokban az adatbázisokban, amelyekben a névmások eltávolításra kerültek, nem volt olyan besorolás, amely független lett volna bármely másiktól. Ez érdekes eredmény, hiszen a dendogramok alapján arra következtethetünk, hogy ezeknél a klaszterezéseknél is meg kell küzdenie az algoritmusnak az egyenlő távolságok problémájával, valamiért itt mégis robosztusabb döntéseket tudott hozni az összevonások sorrendjéről. Bár a munkámban nem tudok kitérni arra, hogy javítsam az algoritmust ahhoz, hogy robosztusabb és helyesebb besorolásokat kapjunk, a fenti ellenőrzés hasznos információt ad ahhoz, hogy ismerhessük az általunk alkalmazott módszer korlátait. Mivel a Buckshot klaszterezés úgyis csak egy kezdő feltérképezésként használom, melynek eredményeit (klaszterszám és klaszterközéppontok) a következő lépések nagyban felül fogják írni, ezért megtartom ezt az eljárást is indítási céllal.

Ezen a ponton szerettem volna döntést hozni arról, hogy melyik adatbázissal haladjak tovább az elemzésben, azonban az eddigi eredmények alapján nem volt egyértelmű a választás. A kofenetikus korrelációs együtthatók tekintetében mindenképp a névmásokat megőrző adatbázisok bizonyultak alkalmasabb választásnak, azonban a névmásokat elvető adatbázisoknak sokkal robosztusabb klaszterbesorolásokat kaptunk. Így úgy döntöttem, hogy mindkét verziónál igyekszem kiválasztani egy ideális klaszterszámvágást, és lefuttatni a teljes mintás k-közép klaszterezéseket, hogy láthassam, hogy mennyiben térnek el a két adatbázison kapott eredmények. Azonban abban is különbséget tettem a két verzió között, hogy milyen mértékű legyen bennük a ritka szavak eltávolítása. Mivel úgy tűnik a kofenetikus korrelációk alapján, hogy a több ritka szó eltávolításán átesett adatbázisokon némileg magasabbak az értékek, ezért alapvetőleg ezeket az adatbázisokat tartottam szerencsésebb választásnak, hiszen így láthatjuk, hogy hatékonyan csökkenthetjük a dimenziók számát, miközben nem kapunk rosszabb teljesítményű klaszterezést. Ezért a névmások eltávolításán átesett

adatbázisokból azt választottam, melyben minden olyan szót eltávolítottunk, amely kevesebb, mint a bejegyzések 1/1000-ben szerepel. Ennek az adatbázisnak a szótára 4798 elemű. Ugyanakkor mivel a névmásokat megtartó adatbázisoknál robosztusabb eredményt mutatott a klaszterezés abban az esetben, ha magasabb dimenziós volt az adatbázis, ezért itt azt a verziót tartottam meg, ahol csak azok a szavak lettek eltávolítva, amelyek kevesebb, mint a bejegyzések 1/10000-ben szerepeltek. Ennek az adatbázisnak a szótára 18098 elemű. Így lehetőségem lehet összehasonlítani azt is, hogy két jelentősen eltérő dimenziószámú klaszterezés közül melyik hogyan teljesít.

Amiatt, hogy az utolsó lépésekben egyenlőek a klasztertávolságok az összevonásoknál, nehéz olyan klaszterszámot találnom, amely valamilyen szempont alapján ideális vágás lehet. Két szempontom marad, mely alapján döntést hozhatok: a különböző klaszterszámoknál kapott Silhouette-értékek és a klaszterek elemszámai. Idáig nem került említésre, de a Silhouette-értékeket mindegyik feldolgozású adatbázison megvizsgáltam, és elmondható, hogy nagyon alacsony, 0 körüli értékeket kaptam minden esetben, minimális különbségekkel a különböző klaszterszámoknál. Ugyanakkor az is észlelhető, hogy a Silhouette-értékek változása nem hozható összefüggésbe azzal, hogy mekkora méretű klaszter került a legutóbbi lépésben egybevonásra (ld. melléklet 1. és 2. táblázata). A Silhouette-értékek 0 közeli elhelyezkedése utalhatna arra, hogy nincsen valódi klaszterstruktúra az adatokban. Azonban ez a jelenség inkább annak tudható be a helyzetünkben, hogy sok a megegyező klasztertávolság, és ezt a Silhouette-érték nem tudja helyesen kezelni. Mivel a Silhouette-értékek még különböző klaszterszámok esetén egymáshoz viszonyítva is nagyon apró különbségeket mutatnak értékükben (ld. 1. ábra), ezért nem hagyatkozom rájuk az ideális klaszterszám kiválasztásakor.



1. ábra: Silhouette értékek ábrázolása különböző klaszterszámoknál (bal oldali ábrán a névmások megtartásával és a csak nagyon ritka szavak eltávolításával létrejött adatbázison, jobb oldali ábrán a névmások elvetésével és a több ritka szó eltávolításával létrejött adatbázison)

A megfelelő klaszterszám kiválasztását így inkább arra alapozom, hogy mekkora elemszámú klaszterek kerültek legutoljára bevonásra a nagyobb klaszterbe. Ez látható a mellékletek 1. és 2. táblázatában. A névmások megtartásával létrejött adatbázison ismertetem a döntési folyamatomat. 20 klaszteres megoldásnál a legnagyobb klaszterünk elemszáma 451, a többi klaszter elemszáma pedig 6 és 106 között váltakozik. Ezek fognak lépésként bevonásra kerülni a legnagyobb elemszámú klaszterbe. Ahogy a legtávolabbi szomszéd módszerének használatát is azért preferáltam, mert az valamelyes egyenletesebb számú klasztereket ad, most is az a cél, hogy inkább nagyobb elemszámú klasztereket tartsunk meg. A 19-es és 12-es klaszterszám között kisebb elemszámú klaszterek kerülnek bevonásra a nagy klaszterbe, sorrendben 10, 16, 12, 12, 10, 12, 12 és 8 elemszámú klaszterek. Ezek bevonásával a legnagyobb klaszter 550 eleműre nő. Ezután azonban először egy 48-as, majd egy 106-os elemszámú klaszter kerül bevonásra a nagy klaszterbe. Ez utóbbi két klaszternek bevonása rövid idő alatt nagy diszkriminálási veszteséggel járna, ezért még ezek bevonása előtt érdemes leállítani a klaszterezést. Így a 12 klaszteres megoldás mellett döntöttem. Bár ez a választás viszonylag önkényes alapon történik, mivel nincsen semmilyen precízebb mutató, mely rendelkezésemre állna a szoftverben, jól diszkriminálna és koszinusz távolság mellett is alkalmazható lenne, de ennek a választásnak hosszútávon nincsen olyan jelentős súlya, hiszen csak ahhoz szükséges, hogy valamilyen szempont szerint célszerű elemszámról indítsuk el a k-közép klaszterezést, azonban az abból létrejövő klaszterek megvizsgálása után szükség szerint úgyis növelni vagy csökkenteni fogom a klaszterek számát az összevonások, illetve tovább-bontások segítségével. Ugyanezen megfontolások mentén a névmások eltávolításával létrejött adatbázisnál 16 klaszteres megoldás mellett döntöttem.

Ezután a Buckshot klaszterezés végével már csak a 12, illetve 16 klaszterközéppontot kellett megkapnom, amelyekből majd indíthatom a teljes mintás k-közép algoritmusokat. Ehhez kiszámoltam azokat az egységvektorokat, melyekből indíthatjuk a k-közép klaszterezést az algoritmust koszinusz távolságra alapozva (számolás ismertetése a szakirodalmi bevezetőben).

5.3. Teljes mintás k-közép klaszterezés

A teljes mintás k-közép klaszterezést az nltk modulon keresztül oldottam meg, mivel a modulnak a KMeansClusterer függvénye lehetővé teszi, hogy euklidészi távolság helyett

koszinusz hasonlóság alapján történjen a klaszterezés, míg a gyakrabban alkalmazott klaszterezésre használt Python modulok (scikit-learn, scipy) nem engednek meg ilyen választást. Elsőként azon az adatbázison futtattam a k-közép klaszterezést, amelyben megőrzésre kerültek a névmások, és csak a nagyon ritka szavak lettek belőle eltávolítva. A klaszterezés eredménye a 8.3. mellékletben szerepel. A klaszterezés megvizsgálásakor odáig jutottam, hogy megvizsgáltam a klaszterekre leginkább jellemző (legnagyobb súllyal szereplő) 10 szót. Ezek alapján arra a következtetésre jutottam, hogy a klaszterezést legnagyobb mértékben a névmások irányították. Megfigyelhető, hogy a klaszterek többségében a legnagyobb súllyal szereplő szavak az azonos szám és személyhez tartozó névmásokból álltak, a többi szóból pedig többnyire nem lehet következtetni bármilyen más közös pontra.

Ezután elvégeztem azon az adatbázison is a k-közép klaszterezést, amelyben eltávolításra kerültek a névmások és több ritka szó került törlésre belőle. Úgy találtam, hogy a legnagyobb súllyal szereplő szavak alapján itt sokkal jobban azonosíthatóak különböző témák (8.4. melléklet). Mivel már csak a szavak alapján is sok klaszter értelmezhetőnek tűnt, ezért ezt a feldolgozású adatbázist sokkal alkalmasabban ítéltem meg a célunkhoz. Így a továbbiakban ennek használatával klaszteresztem a bejegyzéseket.

A klaszterek vizsgálatánál megvizsgáltam egyrészt, hogy milyen távol vannak a klaszterbe tartozó bejegyzések a klaszterközépponttól, valamint azt is, hogy milyen távol esnek az egy klaszterbe tartozó bejegyzések egymástól. Itt a korábbi hierarchikus elemzéssel szemben már nem koszinusz távolságot, hanem koszinuszhasonlóságot számolunk, amelyet úgy kaphatunk meg, hogy 1-ből kivonjuk a koszinusztávolságot. A koszinuszhasonlóság használatával tehát 1-es érték jelenti azt, hogy a két bejegyzés teljesen hasonló, és 0-ás érték azt, hogy teljesen különbözőek. A klaszterközéppontokhoz való átlagos koszinuszhasonlóság 0,2288 és 0,3115 között változtak klaszterenként, míg az azonos klaszterbe tartozó bejegyzések átlagos hasonlósága egymáshoz 0,05619 és 0,09683 között változtak. A későbbiekben látható, hogy ezek a mérőszámok nem feltétlen képesek előrejelezni azt, hogy a klaszterek mennyire könnyen interpretálhatóak, hiszen olykor alacsonyabb mutatókkal rendelkező klaszterekben sokkal egyszerűbben felfedezhető a közös pont, mint magasabb hasonlósággal rendelkezőknél, mégis képet adhat nekünk arról, hogy szavak használatában értelmezve mennyire kerültek hasonló bejegyzések az adott klaszterbe, még ha tartalmilag nehezebben is található hasonlóság közöttük. Emellett megvizsgáltam az is, hogy a különböző klaszterek

középpontjai mennyire esnek közel egymáshoz. Itt a koszinuszhasonlóságok 0,33 és 0,83 közé estek. Ez elég jó mutatónak bizonyult ahhoz, hogy előrejelezze két klaszter tartalmi hasonlóságát is egymáshoz. A pontosabb értékekre a következőkben a klaszterek ismertetésénél térek ki. A klaszterek tartalmának értelmezéséhez klaszterenként a legnagyobb súllyal szereplő szavakat, és a klaszterközéppontokhoz legközelebbi bejegyzéseket vettem segítségül.

5.3.1. Kapott klaszterek jellemzése

A következőkben a klasztereket olyan sorrendben mutatom be, amely szerint tematikailag is csoportosíthatóak. A legtöbb bejegyzés olyan klaszterbe került, melyek leginkább hosszabb beszámolókat gyűjtenek össze a bejegyzésíró saját helyzetéről, állapotáról vagy történetéről. Ezek a klaszterek valamelyest elkülönülnek egymástól témájuk, stílusuk alapján. Elsőként ezeket ismertetem.

Öt klaszterről elmondható, hogy mind hosszabb bejegyzéseket tartalmaznak, melyekben a saját történetüket mutatják be a személyek a pszichés küzdelmeikkel kapcsolatban. A 0-s klaszterbe 5780 bejegyzés került. Itt olyan bemutatkozó történetek szerepelnek többnyire, melyekben a személyek azt a bizonytalanságukat fejezik ki, hogy vajon szenvedhetnek-e valamilyen mentális zavarban. Az 1-es klaszterben olyan történetek szerepelnek, melyekben nagyobb fókusz esik az érzelmi életüket jelentősen befolyásoló társas kapcsolatokra, akár rokoni, akár párkapcsolati téren. Ebbe a klaszterbe 11891 bejegyzés került, ezzel ez a legmagasabb elemszámú bejegyzés. A 3-as klaszterbe 4670 bejegyzés került. Amentén kerülhettek ezek a bejegyzések különválogatásra, hogy a történetekben jobban hangsúlyozva van tartalmilag is és a szóhasználatban is az idő és a múlt tényezője. Az 5-ös klaszterben szereplő 7142 bejegyzésben valamelyest mozgalmasabb történetek szerepelnek. Ezekben többször szóbajön a munka témája is, és kevésbé érzelmi fókuszú beszámolókat olvashatunk. A bejegyzések azonban viszonylag kevésbé vannak közel a klaszterközépponthoz (0,2373) és egymáshoz (0,05619). A 6-os klaszterbe 4502 bejegyzés került. Kiemelendő, hogy ebben a klaszterben a legmagasabb a bejegyzések átlagos hasonlósága a klaszterközépponthoz (0,3115) és egymáshoz (0,09683), így ez lehet a legegységesebb klaszter a vektortérben. Az ebben szereplő bejegyzések kifejezetten érzelemfókuszúak. Feltűnően sok bennük az „én” szó

használata, nagyon énközpontú bejegyzések. Elkeseredettség és önostorozás tükröződik a beszámolókból.

A 6-os klaszter kivételével, mely nem mutat 0,7-nél nagyobb hasonlóságot más klaszter középpontjával, a másik négy klaszter erős hasonlóságokat mutat több klaszterrel is. A 0-s klaszter 0,7892, illetve 0,7395 értékű koszinuszhasonlóságban áll az 1-es és az 5-ös klaszterrel. Az 1-es klaszter középpontja a 0-sén kívül még szorosabb hasonlóságot mutat a 3-as (0,7883) és főleg az 5-ös klaszter (0,8362) középpontjával. Ez utóbbi a legnagyobb hasonlóság két klaszterközéppont között a klaszterezésünkben. A 3-as klaszter az előzőeken kívül az 5-ös klaszterrel mutatott szoros hasonlóságot (0,7968). Ebből látható, hogy a 0-s, 1-es, 3-as és 5-ös klaszterek között bár azonosíthatóak eltérések, értelmezhető többnyire, hogy milyen szervezőlogika mentén jöttek létre a különböző klaszterek, mégis a klaszterközéppontjaik nagyfokú hasonlóságot mutatnak, közel esnek egymáshoz.

A fenti klaszterektől valamelyest jobban különbözik két klaszter. A 4-es klaszterben csak 1629 bejegyzés került. Ebben valamelyest rövidebb bejegyzések szerepelnek, melyek általában nem első bemutatkozások, hanem már egy beszélgetésfolyamat részei lehetnek hangvételükből fakadóan. Gyakran beszélnek érzelmi tünetekről, azonban sokszor nem magukról, hanem depressziós rokonukról vagy barátjukról írnak. Kevésbé egységes klaszter, nehezen interpretálható, hogy mi a közös szervező pont. A bejegyzések viszonylag alacsony hasonlósága a klaszterközépponthoz (0,24) és egymáshoz (0,05703) is ezt tükrözheti. A klaszterközéppontja leginkább az 1-es klaszter középpontjához hasonlít (0,7452), mely magyarázható azzal, hogy abban a klaszterben is sokat beszélnek más személyekről, nem kifejezetten csak önmagukról.

Kevésbé egységes emellett a 11-es klaszter, melyben 3201 bejegyzés szerepel. Ebben tartalmilag az a meghatározó, hogy jellemzőbb a pozitív szavak használata, a bejegyzésírók a depresszióval való megküzdésükről, különböző praktikákról számolnak be. Itt esetenként enyhébb vagy akár javuló tünetekről számolnak be. Viszont szerepelnek a klaszterben olyan bejegyzések, amelyek a szóhasználatuk alapján tévesen kerülhettek a klaszterbe, hiszen az előzőekkel ellentétben nem tükröződik bennük bizakodás, megküzdés vagy javulás, hanem erősebb tünetek, és reménytelen érzés kerül leírásra. A klaszter bejegyzései alacsony hasonlóságot mutatnak átlagosan a klaszterközépponthoz (0,2390) és egymáshoz (0,05683).

Érdekes ugyanakkor, hogy erősebb hasonlóságot (0,7212) mutat a középpontja az 5-ös klaszter középpontjához. Lehetséges, hogy ez azért van, mert az 5-ös klaszterben is több cselekedet, aktív tett jelenik meg.

A klaszterek másik nagyobb hasonló csoportját akként lehet megragadni, hogy ebben nem önálló beszámolók, hanem inkább egymásnak adott válaszok szerepelnek. Nagyon könnyen interpretálható klaszter ebben a 14-es számú, melybe elég alacsony (1811) számú bejegyzés került. Ezek a bejegyzések leginkább kifejezetten rövid, egy-két mondatos köszönetek azért, mert mások válaszoltak nekik korábbi bejegyzésükre, tanácsot, vagy támogatást nyújtottak nekik. A 10-es klaszterbe 2826 bejegyzés került. Ez tartalmilag hasonlóképp egységes klaszter, leginkább régi tagok, vagy akár adminisztárorok válaszai, akik üdvözölnek egy első bejegyzést író tagot a fórumon, és rövid tanácsot adnak, vagy iránymutatást és biztatást a fórum használatához. A 9-es klaszterbe hasonló számú (2658) bejegyzés került. Ebben a bejegyzések átlagosan viszonylag közelebb vannak a klaszterközépponthoz (0,2843) és egymáshoz (0,08047) is. A klaszter bejegyzései nagyon hasonlóak abban, hogy szintén válaszok, melyek szinte ugyanúgy kezdődnek (Nagyon sajnálom, hogy… - I'm so sorry for…). Ezekben a válaszokban együttérzést fejeznek ki valakinek a leírtak miatt, tanácsot adnak neki, és biztatják, érzelmi támogatást nyújtanak.

Sokat foglalkozik még a támogatás témájával a nagyobb elemszámú (7030) 8-as klaszter. Az ebben szereplő bejegyzések valamilyen formában tartalmazzák a tanácsadás, támogatás fontosságát, azonban míg egy részük egy másik tagnak írott válasz és támogatás, addig a bejegyzések másik része nem másoknak szánt támogatás, hanem az önmagukról írt bejegyzésbe kerül bele a támogatás, segítségnyújtás témaköre. Ez a klaszter tehát nem tekinthető annyira egységesnek, hiszen a tanácsot adó bejegyzések kerülhetnének akár másik ilyen témájú klaszterbe is, elkülönülve a nem tanácsadó bejegyzésektől. Ez az inhomogenitás a klaszterközépponthoz (0,2431) és az egymáshoz (0,05895) való átlagos hasonlóságon is tükröződik valamelyest, valamint a más klaszterközéppontoktól való távolságon is, hiszen négy klaszterrel is magasabb, mint 0,7-es értékű hasonlóságot mutat. Ezek a 0-s, 1-es, 3-as és 5-ös klaszterek, melyek mind különböző fókuszú hosszabb beszámolókat tartalmaztak.

A klaszterek utolsó nagy csoportjába olyan klaszterek sorolhatók be, melyek testi tünetekről és orvosi kezelésekről szólnak. Megjelent a depresszió és egyéb mentális betegségek orvosi

kezelése több klaszterben. A 13-as klaszterbe 5045 bejegyzés került, így ez egy viszonylag nagy klaszternek mondható. A bejegyzések klaszterközépponthoz (0,2389) és egymáshoz (0,05689) való hasonlósága nem volt túl magas, mégis kifejezetten körülhatárolható, hogy egységesen a különböző depresszióra szedett gyógyszerekről és hatásaikról kerültek bele bejegyzések.

Emellett még megjelenik a gyógyszerek témája az alacsonyabb (1240) elemszámú 2-es klaszterben is, melyben a bejegyzések egymástól (0,07102) és a klaszterközéppontól (0,2679) vett átlagos távolsága viszonylag magasabb. Ebben leginkább olyan leírások vannak, melyeknek témája az úgynevezett kezelésrezisztens depresszió, melyben a depresszió nem javul a gyógyszeres vagy egyéb kezelések hatására. Gyakran megjelenik utolsó próbálkozásként az elektrokonvulzív kezelés (régi nevén elektrosokk-terápia) kérdésköre is.

Létrejött 2 olyan klaszter a fentieken kívül, amelyek bár szintén tünetekről és kezelésükről szólnak, de kevésbé közvetlenül a depresszióhoz kapcsolódóan. A 15-ös klaszterbe viszonylag alacsonyabb az elemszám (1880), azonban kifejezetten egységes klaszter, mely abban is megmutatkozik, hogy a bejegyzések közelebb vannak a klaszterközépponthoz (0,2957) is és egymáshoz (0,0869) is. Tartalom a pszichés állapotok az alvásról és az alvásproblémákról szól. A példaként elolvasott bejegyzések mindegyikének ez volt a fő témája, így sejthető, hogy egy nagyon egységes klaszterről lehet szó.

A 12-es klaszterbe 3063 bejegyzés került. Bár a bejegyzések viszonylag távolabb vannak a klaszterközépponttól (0,2288) és egymástól (0,052) is, mégis tartalmilag jól körüljárható klasztert kaptunk. Témája a különböző testi fájdalmak, tünetek, és egyéb, nem mentális betegségek (bár mentális okokat gyanítanak olykor a hátterében), valamint ezek kezelése.

Ezeknek a klasztereknek a középpontjai nem mutatnak nagyon erős, 0,7 feletti hasonlóságot egyik klaszterközépponthoz sem, de például az megfigyelhető, hogy a 12-as klaszterközéppont az összes többi közül leginkább a 13-es klaszterközépponthoz van közel (0,6671).

Külön tárgyalnám a 7-es klasztert, mivel bár ez az állapot más klaszterek bejegyzéseiben is felmerül olykor, ez a klaszter kifejezetten a bipoláris depresszió témakörével foglalkozik egységesen. A klaszterbe 1927 bejegyzés került. A középpontja nem áll nagyon közel más klaszterekéhez, a legnagyobb hasonlóságot (0,6102) az 1-es klaszterrel mutatja.

5.4. Klaszterek összevonása és tovább-bontása

Mivel nem találtam egységesnek a 8-as klasztert, amely ráadásul elég nagy elemszámmal rendelkezik, megpróbáltam ezt 2 klaszterbe tovább-bontani, mivel szerettem volna, ha külön tudom választani azokat a bejegyzéseket, amelyekben valóban segítségnyújtás zajlik valamilyen válasz formájában azoktól, amelyekben csak említi a bejegyzésíró a támogatás, segítség témakörét. Ehhez k-közép klaszterezést futtattam a klaszter bejegyzésein, 2 klaszteres megoldást kérve. Itt azonban már nem alkalmaztam Buckshot klaszterezést a klaszterközéppontok megállapításához, mivel nem voltam elégedett a hierarchikus klaszterezés eredményével. A kapott két klaszter elemszáma 2963 és 4067, a klaszterközéppontjaik 0,8412 hasonlóságúak egymáshoz. A két klaszter bejegyzéseibe beleolvasva megállapítható, hogy nem sikerült a kívánt szempont szerinti szeparálás, továbbra is vegyesen fordulnak elő tényleges támogatások és csak támogatást említők. Azonban a bejegyzések olvasása során úgy tapasztaltam, hogy többnyire inkább támogató válaszok szerepelnek a klaszterekben, és csak kevesebb olyan szerepelhet benne, amelyben csak elvétve említés szintén szerepelnek támogatásra, segítségre utaló szavak. Így a 8-as klaszter tekinthető egy olyan klaszternek, melyben főként támogató és a támogatás, segítségnyújtás fontosságát hangsúlyozó bejegyzések vannak. Így egyben tartom meg a 8-as klasztert, nem kezelem szétválasztva a továbbiakban.

Hasonlóan próbálkoztam szétbontani 11-es klasztert aszerint, hogy valóban pozitív személetű a bejegyzés vagy sem, valamint a 4-es klasztert aszerint, hogy másról beszél a személy, vagy saját magáról. A klaszterezések számszerű eredményei a 8.6-es és 8.7-as mellékletben láthatók. A 11-es klaszterből létrejött két klaszter középpontja 0,6738 koszinuszhasonlóságú egymáshoz. Ebből bár feltételezhetnénk, hogy valóban sikerült elkülöníteni két klasztert egymástól, azonban a bejegyzésekbe beleolvasva nem jött létre igazán értelmezhető eredmény. A kívánt szempont szerinti szétválasztás nem valósult meg semmiképp. A 4-es klaszterből létrejött két klaszter középpontja 0.8854 koszinuszhasonlóságú egymáshoz, elég közel helyezkednek tehát el. Ez a bejegyzéseket olvasva és a legnagyobb súllyal szereplő szavakat megnézve is érzékelhető: nem igazán fedezhető fel különbség a két klaszter között. Így a fenti klaszter-kettébontásokat elvetettem. Mivel a klaszterezés egy kevésbé irányítható elemzési módszer, ezért nem igazán lehet aszerint szétbontatni vele klasztereket, ami szerint mi szeretnénk. Megpróbálkozhatnánk olyan kezdőközéppontok kijelölésével, amelyet a

bejegyzések egy almintájának elolvasásával és csoportbasorolásával határoznánk meg, azonban ez már inkább klasszifikációs feladatmegoldás lenne, mintsem klaszterezési.

További újrabontásként még megpróbálkoztam azzal, hogy a bejegyzések egy részét újraklaszterezzem. Mivel az elsőként ismertetett hosszú bejegyzések jelentős hasonlóságot mutatnak egymáshoz, és olykor csak apróbb szempontok mentén térnek el egymástól, vagy nehezen interpretálhatóak, ezért megpróbálkoztam azzal, hogy csak ezen a szűkebb dimenziótéren készítsek el klaszterezést. Így tehát összevontam a 0-s, 1-es, 3-as, 4-es, 5-ös, 6-os és 11-es klasztereket. Ez a 7 klaszter összesen 38815 bejegyzésből áll. Nem az a célom, hogy növeljem, vagy csökkentsem a klaszterek számát, mivel úgy éreztem, hogy a klaszterek viszonylag jól értelmezhetően elkülönülnek egymástól. Inkább azt szerettem volna elérni, hogy pontosabb felosztást találjak ezen a csoporton belül, ezért 7 klaszteres megoldást vizsgáltam. Az újraklaszterezés eredménye a 8.8-es mellékletben látható, míg az eredeti és az új klasztertagságok összevetése a 8.9-es mellékletben.

Az eredetileg 0-s klaszter legtöbb eleme átkerült az 1-es klaszterbe. Ez a két klaszter azonosíthatóan a diagnózisban bizonytalanok leírása. Az eredetileg 1-es klaszter legtöbb bejegyzése a 6-os klaszterbe került. Mindkét esetben azonosítható, hogy ezek nagyon kapcsolati fókuszú leírások. Ugyanakkor jelentős része került a 2-es klaszterbe is, melyben úgy tűnik, hogy inkább mások mentális problémájáról számolnak be. Így ez egy sikeres megkülönböztetése lehet az eredetileg 1-es klaszternek. Hasonlóan azonosítható még a legtöbb bejegyzés átkerülésével az eredetileg 11-es klaszter az új 0-as klaszterrel, melyekben enyhébb tünetes beszámolók szerepelnek. Az eredetileg 3-as klaszter jellemzően időre jobban fókuszáló bejegyzéseket tartalmazott. Ezek legtöbbje az 5-ös klaszterbe került át. Ugyanígy az eredetileg 5-ös klaszter legtöbbje is az új 5-ös klaszterben szerepel. Az új 5-ös klasztert leginkább a mozgalmassága, kevésbé érzelmi fókusza dominálja. Így elmondható, hogy az időfókuszú klaszter többnyire egybeolvadt a mozgalmasabb témájú klaszterrel. Ez szerencsés, mivel az időfókuszúság különben is nehezen értelmezhető szempont volt. Kifejezetten nehezen interpretálható klaszter volt az eredetileg 4-es számú, melyben nehezen lehetett egyedi jegyeket felfedezni. Ez ahogy látható, feloldódott a többi klaszterben. A legtöbbje a kapcsolati fókuszú 6-os klaszterba került, de került egy kisebb részük az új 2-es klaszterbe is, mely mások problémáiról szól, úgyhogy vélhetően sikeresen elkülönítésre került az eredeti 4-es klaszter ebből a szempontból. Látható ugyanakkor, hogy az eredetileg 6-os klaszter két új

klasztert, a 3-as és 4-est töltötte fel a leginkább. Az eredetileg 6-os klaszter eredetileg is eléggé elkülöníthető volt a többitől, magas volt a bejegyzéseinek hasonlósága, és értelmezhető is volt erős énközpontúság és elkeseredettség mentén. Az új 3-as és 4-es klaszter szintén magas bejegyzéshasonlóságot mutatnak, azonban nem igazán interpretálható, hogy miben különböznek egymástól sem a legnagyobb súlyú szavak mentén, sem a bejegyzések olvasásával. Mivel a két klaszter középpontjuk alapján egymáshoz állnak a legközelebb, ezért a jobb interpretálhatóságért ezt a két klasztert egybevonom, és a továbbiakban egyként kezelem.

Összességében elmondható, hogy a 7 hasonlóságot mutató klaszter újrabontása eredményes volt. Jobban tisztultak a klaszterek értelemezhetőség szempontjából, és bár eggyel kevesebb klaszterünk lett, de így jobban értelmezhetőek a klaszterek tartalmukban. Így egybevonva az teljes mintás k-közép algoritmusból származó 2-es, 7-es, 8-as, 9-es, 10-es, 12-es, 13-as, 14-es és 15-ös klaszterekkel összességében 15 klaszteres végső megoldásunk született. A 15 klaszter összefoglaló táblázata a 8.10-es mellékletben látható. Az újrabontott klasztereket az átfedések elkerüléséért 20-szal kezdődően számoztam. Így kaptunk 20-as, 21-es, 22-es, 23-as (az új 3-as és 4-es összevonásából), 25-ös és 26-os klasztereket.

5.5. Nagy mintán hierarchikus elemzés

A klaszterezésben utolsóként megvizsgáltam, hogy nagyobb mintán milyen eredményt kapok hierarchikus klaszterezési módszerrel, hogy összehasonlíthassam az eredményét a scatter/gather módszerrel kapott eredménnyel. A klaszterezést a teljes minta felén, 33148 bejegyzésen futtattam le azért, hogy még kezelhető időn belül eredményt kapjak belőle. Agglomeratív módszerrel futtattam az elemzést, a korábbi tapasztalatokból kiindulva pedig csoportok közötti átlagos távolság és legtávolabbi szomszédok módszerével is megvizsgáltam, hogy milyen eredményeket ad.

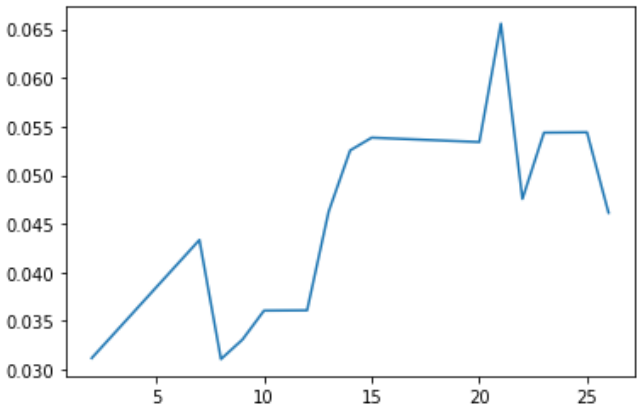
A korábbi kismintás hierarchikus klaszterezéshez hasonlóan a nagy mintán futtatott klaszterezésnél is erősen jelen van a láncalkotási jelenség. Megvizsgálva az utolsó 100 összevonás előtti állapotot úgy találtam, hogy csoportok közötti átlagos távolság használatával a 31652 elem egy közös klaszterbe került már, és csak a maradék 1496 bejegyzés oszlott el a többi 99 bejegyzésben. Ezen a problémán még a legtávolabbi szomszéd módszere sem tudott jelentősen segíteni, 100 klaszteres megoldásnál ott 28590 bejegyzés tömörült be a legnagyobb

klaszterbe. Azonban, ha nagyobb klaszterszámnál vizsgálom a hierarchikus összevonás eredményét, akkor látható, hogy nem történik láncalkotás a kezdetektől. 2000 klaszteres megoldásnál például már jól látható, hogy van több ezres nagyságrendű klaszter is, valamint sok százas nagyságrendű. Azonban ilyen nagy klaszterszám nem teszi lehetővé a klaszterek értelmezését, valamint az összevetést a scatter/gather módszer eredményével.

5.6. Névmások megjelenése a klaszterekben

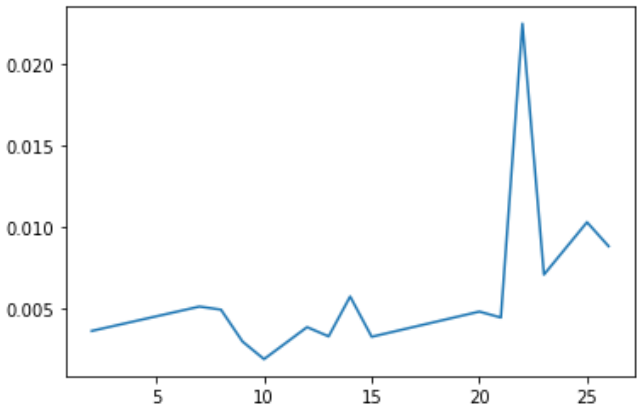
A névmások gyakoriságát az egyes klaszterekben olyan feldolgozású adatbázis segítségével vizsgáltam, melyben a stopszavak (és ezzel együtt a névmások) eltávolítása nem történt meg, csak a lemmatizáláson estek át a szövegek. A névmások megjelenésének eltérései a klaszterekben vizsgálhatók lettek volna varianciaanalízissel, azonban ehhez nem teljesül a megfigyelések függetlensége egymástól. Az adatgyűjtés módjából eredően az adatbázisba bekerülhetett egy bejegyzésírónak több bejegyzése is, és ezek a bejegyzések így nem tekinthetők függetlennek egymástól. Azonban nincs információ az adatbázisban a szerzőkről, ezért nem is detektálhatóak az azonos személytől származó bejegyzések. Így szignifikanciatesztek helyett csak az adatokra való ránézéssel és ábrázolásával vontam le következtetéseket. Azt vizsgáltam, hogy az „én” („I”) névmás melyik klaszterben fordul elő a leggyakrabban a klaszterben szereplő teljes szógyakoriság arányában, valamint ugyanezt vizsgáltam az „ő” („he” és „she”) névmások esetében is. A névmásoknak csak a ragozatlan formáját vizsgáltam.

Az egyes szám, első személyű névmás gyakorisága az összes szógyakorisággal arányosítva a 21-es klaszterben a legmagasabb (6,559%, 2. ábra), melyben olyan hosszú történetbemutatások szerepelnek leginkább, melyekben az a bizonytalanság fejeződik ki, hogy milyen zavarban szenvedhetnek valójában. Emellett még magas, 5 és 5,5% közötti gyakoriságot mutat a névmás a 25-ös, 23-as, 15-ös, 20-as és 14-es klaszterben is. Ezek a klaszterek egyrészt a hosszabb beszámolók közül kerülnek ki szintén (mozgalmasabb / elkeseredett/ enyhébb tünetek), míg a 15-ös klaszter az alvászavarokat tartalmazza, a 14-es pedig rövid köszöneteket. A legalacsonyabb arányban a 8-as és 2-es klaszterben fordul elő egyes szám első személyű névmás. Ezek a támogatás és a kezelésrezisztens depresszió témájában szerveződött klaszterek.



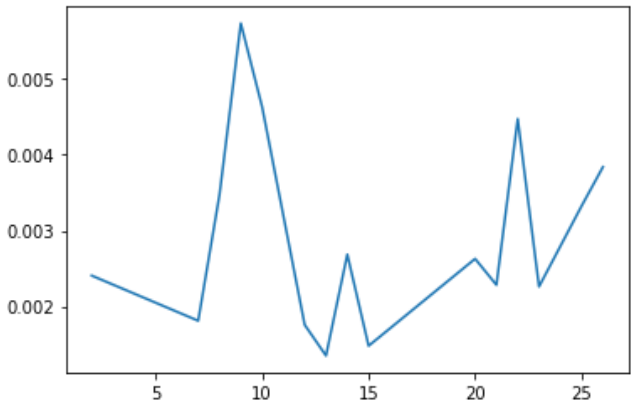
2. ábra: Egyes szám első személyű névmás aránya klaszterenként

Egyes szám harmadik személyű névmásokat kiugróan magas százalékban tartalmaz a 22-es klaszter (2,25%, 3. ábra), de gyakrabban jelenik meg még a 25-ös, 26-os és 23-as klaszterben is. A 22-es klaszterben azok a bejegyzések csoportosulnak, melyekben egy másik személy mentális betegségéről számol be a bejegyzésíró, míg a másik három is hosszabb beszámolókból áll össze (mozgalmasabb / kapcsolatfókuszú / elkeseredett). A legkevésbé a 10-es, 9-es, 15-ös és 13-as klaszterben fordul elő gyakran ilyen névmás (0,18 és 0,33% között). Ezek a klaszterek az új tagok üdvözlése, a biztató válaszok, az alvás és a gyógyszeres kezelés tematikájában szerveződtek.



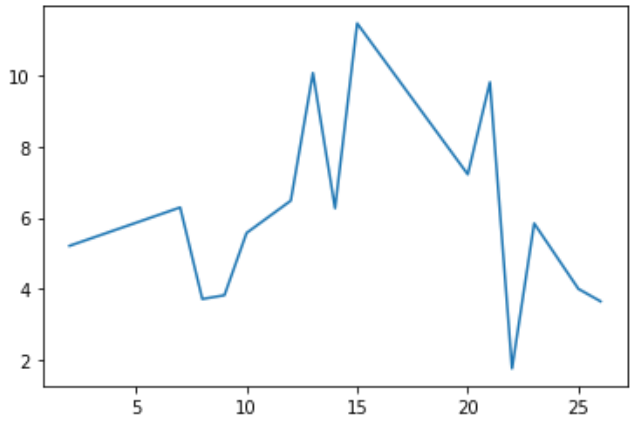
3. ábra: Egyes szám harmadik személyű névmások aránya klaszterenként

Többes szám első személyű névmások leggyakrabban a 9-es klaszterben jelennek meg (0,5733%, 4. ábra), de 0.45% körüli még a 10-es és 22-es klaszterben is. Ezek biztató válaszokat, új tagok üdvözlését, és más személy mentális betegségéről való beszámolókat tartalmaznak. A legalacsonyabb aránnyal a 13-as, 15-ös, 12-es, és 7-es klaszterben szerepelnek (0,1 és 0,2% között), melyek a gyógyszeres kezelés, alvásproblémák, testi fájdalmak, és bipoláris depresszió témájában íródtak.



4. ábra: Többes szám első személyű névmások aránya klaszterenként

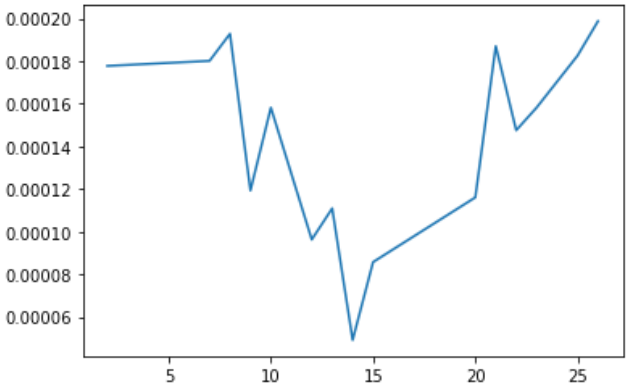
Összevetettem még azt is, hogy milyen eltérések mutatkoznak az egyes szám első személyű névmás gyakoriságában ahhoz képest, hogy mennyi egyes szám harmadik személyű és többes szám első személyű névmást használnak a klaszterekben (5. ábra), hiszen a szakirodalom szerint míg az egyes szám első személyű névmások használata fokozódhat súlyos állapotban, addig ezzel párhuzamosan az egyes szám harmadik személyű és a többes szám első személyű névmások használata csökkenhet, jelezvén a személy fokozott befelé fordulását, izolálódását. A 15-ös klaszterben van a többi vizsgált névmáshoz képest a legtöbb egyes szám első személyű, több, mint 11-szer annyi „én” („I”) szerepel benne, mint „ő” („he”, „she”) és „mi” („we”). A legmagasabb arányok közé tartozik még a 13-as és a 21-es klaszter is. Ezek a klaszterek az alvásproblémák, a gyógyszeres kezelés, és a diagnózisban való bizonytalanság témakörében szerveződtek.



5. ábra: Egyes szám első személyű névmások aránya az egyes szám harmadik személyű és a többes szám első személyű névmásokhoz képest klaszterenként

Ellenőrzésként vizsgáltam még az „öngyilkosság” szó arányát a klaszterekben. A 26-os klaszterben a legmagasabb a szó aránya (0,01987%, 6. ábra), mely kapcsolatfókuszú hosszabb

beszámolókat tartalmaz, de magasabb még a 8-as, 21-es, 25-ös, 7-es és 2-es klaszterben is (0,01778 és 0,01828% között). Ezek a támogatás, diagnózisban való bizonytalanság, mozgalmasabb beszámolók, bipoláris depresszió és kezelésrezisztens depresszió témájában íródtak.



6. ábra: Öngyilkosság szó aránya klaszterenként

6. Megbeszélés

6.1. Érvényesség

Az eredmények validálása nehézkes, mivel nem áll rendelkezésünkre a bejegyzések valamiféle felcímkézett változata. Így pusztán az interpretáláshoz használt olvasói benyomásokra hagyatkozhatunk, valamint arra, hogy az eredményeket mennyire támasztja alá az algoritmus. Ugyanakkor fontos, hogy a helyes eszközök használata is validálni képes az eredményeket. Korábbi kutatási ismeretek és tapasztalatok mentén olyan eszközöket használtunk fel, melyek használata bevett a szövegelemzés területén. Ilyenek a tf-idf súlyozás és a koszinusz hasonlóság használata. Ugyanakkor az eredmények stabilitását biztosítva, amennyiben sztochasztikus tényező is szerepet játszhat az elemzésben, biztosítottam a többszöri futások alapján vett eredmények vizsgálatát. Ez történt például a klaszterek egy részének újrabontásakor, mikor nem adtam meg a programnak a kezdőközéppontokat, hanem megengedtem, hogy válasszon, ugyanakkor 10-szer választtattam vele, ezzel kiküszöbölve a választás esetlegességét.

6.1.1. Eredmények értelmezésének korlátai

Mivel ez egy feltáró módszer, így valamelyest önmaga által irányított, azt képes megtalálni, ami az algoritmus számára láthatóvá válik. A k-közép klaszterezés nem tesz feltételezést az

eloszlásokról, pusztán csak a klasztereken belüli távolságokat igyekszik minimalizálni tekintet nélkül minden más struktúrára az adatokban. Ennek ellenére jól értelmezhető, és a korábbi kutatások alapján alátámasztható klasztereket talált.

Az olykor furcsa eredményeket (alacsony Silhouette értékek, láncalkotás hierarchikus klaszterezésnél) magyarázhatná az, hogy túl sok dimenzió van, azaz túl sok változó, amely szerint a klaszterezés történik, és ehhez nincs elég magas elemszám. Ebből a szempontból szerencsés volt az alacsonyabb dimenziós feldolgozású adatbázist választani a klaszterezéshez, ugyanakkor, ha túlzottan továbbcsökkentenénk a dimenziót, akkor pedig értékes információt veszítenénk a mintaelemeinkről, ezzel torzítva a klaszterezés eredményét. Olyan nagyságrendű dimenziónál, amelyet egy korpuszszótár ad, sajnos valójában nem tud elég sok bejegyzésünk lenni ahhoz, hogy betöltsék a sokdimenziós vektorteret. Emiatt a legtöbb bejegyzés távol fog egymástól kerülni, a vektortér széleire (ld. Hastie, Tibshirani és Friedman, 2009). Ez magyarázhatja, hogy miért volt olyan sok bejegyzés koszinusztávolsággal mérve abszolút távol egymástól.

Ezek a nagy távolságok ugyanakkor akadályozhatták a Silhouette-értékek mérőképességét, illetve a hierarchikus klaszterezés is problémákba ütközött ennek a kezelésénél. Mivel kevesebb bejegyzés volt a kis mintában, mint ahány dimenzió a szótárból adódóan, ezért nagyon nehezen kerülhettek be egymáshoz kicsit is hasonlító bejegyzések ennyi dimenzió mentén, főként, hogy a legtöbb dimenzióban minden dokumentum 0-s értéket vesz fel, hiszen a korpusz szótárának csak egy kis részét használja egy-egy bejegyzés. Épp emiatt a további klaszterbontásoknál már nem alkalmaztam a kezdőközéppontok kiválasztásához a kis mintás hierarchikus klaszterezést, hiszen látható volt, hogy nehezen tudja elkerülni a láncalkotási folyamatot még a legtávolabbi szomszéd módszerénél is, ezzel egyenetlen klaszterméreteket hozva létre. A láncalkotás problémáját azonban az sem tudta orvosolni, ha nagyobb mintán futtattam hierarchikus klaszterezést, ott is az utolsó összekapcsolásoknál már végig egy kiugróan nagy elemszámú klaszterbe vont bele sokkal kisebbeket. Amikor 2000 klaszteres megoldást vizsgáltam, azonban kiderült, hogy korábbi lépésekben még nem uralkodik ez a jelenség, ott még több nagyobb elemszámú klaszter van. Azonban ezek értelmezése a magas klaszterszám miatt nem kivitelezhető.

Látható, hogy a kezdeti összevonásoknál még egyenletesebben történik a klaszterhierarchia kialakítása, azonban a lépések vége felé közeledve már minden nagyobb elemszámú klaszter egybevonódik, és csak nagyon kicsi elemszámú klaszterek esnek rajta kívül, melyek utána lépésenként mind a nagy klaszterbe kerülnek bevonásra. Ezt az magyarázhatja, hogy ezek az utolsó lépésekig különálló, kicsi elemszámú klaszterek mind olyan bejegyzéseket tartalmaznak, amelyek nagyon erősen különböznek a legtöbb elemtől, és egyik nagyobb klaszterhez sem állnak valójában közel. Ezek úgynevezett outlier bejegyzések, melyek valamiért nagyon kilógnak az adatbázisból. Beleolvasva egy pár ilyen bejegyzésbe egy részük valóban elég sajátosnak tűnt (pl. idézet a természet szépségéről), míg mások látszólag nem tűntek ilyen mértékben kirívónak a korpuszból. Azonban mégis arra következtethetünk, hogy ezek a bejegyzések szóhasználatukban jobban különböznek, mint az összes nagyobb klaszter egymástól. Így szóhasználatukban outlierként működhettek, ezzel megzavarva a hierarchikus klaszterstruktúra felépítését. Ugyanakkor nem szerettem volna törölni ezeket a bejegyzéseket a jobb klaszterezés érdekében, mivel nehezen tudtam volna meghúzni egy határt, amelynél az addig valamennyi nagyobb elemszámú klaszterbe nem csatlakozott elemet kitörlöm, valamint túl sok elem törlésével járt volna ez az eljárás. Emellett mivel a k-közép klaszterezés ezen bejegyzések megtartásával is jól interpretálható klasztereket tudott azonosítani, nem szabotálták a bejegyzések a feladat végrehajtását. Amennyiben azonban a jövőben szeretnénk teljesen kiszűrni ezeknek az outlier bejegyzéseknek a hatását a klaszterstruktúrából, érdemes lehet DBSCAN (Density-based Spatial Clustering of Applications with Noise) klaszterezési módszer használatával kezelni a problémát (Ester, Kriegel, Sander és Xu, 1996). Ez az eljárás a vektortérbeli sűrűsödési pontok alapján hoz létre klasztereket, amely elemek pedig túl távol esnek a sűrűsödési pontokban lévő „magelemektől”, azokat outlierként kezeli az eljárás, és nem sorolja őket klaszterbe. Campello, Moulavi és Sander (2013) kiterjesztették a módszert hierarchikus klaszterezési eljárásra is. Ezzel az eljárással tehát akár ki lehetne szűrni az túlzottan egyedi elemek hatását a hierarchikus klaszterezésből.

6.1.2. Más eredményekkel összehasonlítás

A korábban ismertetett szakirodalom alapján több olyan kutatás is van, amely depresszió témájú fórumok bejegyzéseit igyekezett különböző módokon csoportokban rendezni, így ezekkel nagyon jól összehasonlíthatók a saját eredményeink. Különösen érdekes lehet a Németh és munkatársai (2021) eredményeivel való összevetés, mivel ők ugyanezekből a

bejegyzésekből álló adatbázison hajtottak végre topikmodellezést, így amennyiben jelentős hasonlóság mutatkozik a két módszer által kapott eredmény között, azok egymást is validálhatják. A névmások használatában megmutatkozó különbségek már több kutatás által is alátámasztottak, nem találtam olyan kutatást, amely ne talált volna összefüggést az egyes szám első személyű névmások fokozott használata és a súlyos depresszió vagy öngyilkosság között, így ez egy kevésbé vitatott eredmény. Ugyanakkor fontos tényező, hogy ezek a kutatások leginkább a vizsgálati környezetben létrehozott szövegeket vizsgáltak, melyeket irányított módon írattak meg a résztvevőkkel, vagy ha erre nem volt lehetőség, akkor is nagyon hasonló környezetből eredő szövegeket, ezzel is biztosítva, hogy még inkább csak a depressziós állapot okozhasson a névmáshasználatban különbséget (pl. esszé írása 20 perc alatt az egyetemre járással kapcsolatos legmélyebb érzéseikről és gondolataikról – Rude et al., 2004). A depresszió fórumokon azonban valójában nem egységesen ugyanolyan témában és stílusban íródnak bejegyzések, ezért érdekes kérdés, hogy itt is jól használható lesz-e a névmások különbsége a súlyosabb depressziós állapot jelzéséhez. A kutatás további kérdése még a scatter/gather módszer algoritmusának tényleges hasznossága. Ennek nem találtam teljesen pontos összevetését más klaszterezési módszerek eredményével, azonban több cikk is foglalkozik a témájával, melyekkel összehasonlíthatók a saját kutatásom tapasztalatai.

6.2. Értelmezés

6.2.1. Klaszterezés eredményének megvizsgálása

Németh és munkatársai (2021) eredményeivel kifejezetten jól összevethetők a jelen elemzésből származó eredmények, mivel ugyanazokon a bejegyzéseken történtek az elemzések. A topikmodellezésből származó eredményekhez hasonlóan a klaszterezésből is létrejöttek hosszú beszámoló jellegű klaszterek, és inkább az interakciós beszélgetésbe illeszkedő válasz jellegű bejegyzéseket tartalmazó klaszterek. A beszámolókban megjelentek kapcsolati vagy munka helyi fókuszok a topikmodellezéshez hasonlóan, ugyanakkor a párkapcsolati és családi problémák nem váltak külön klaszterbe. Találtam olyan klasztert is, amelyben jobbanlétről számolnak be. Az interakciós válaszokat ugyanakkor a klaszterezés során tartalmilag kevésbé érdekes szempontok mentén sikerült elkülöníteni, inkább szóhasználatban érződött némi tipikus, ugyanakkor tartalmat nem annyira befolyásoló különbség. Ugyanakkor a gyógyszer témája dominálta az egyik klasztert. Emellett megjelentek még olyan klaszterek, amelyek könnyedén körülhatárolhatóak voltak, mint például a bipoláris

fókuszú, vagy a kezelésrezisztens terápiáé. Ezekből az összehasonlításokból az tapasztalható, hogy a topikmodellezés képes volt arra, hogy szofisztikáltabb különbségű témákat azonosítson. A klaszterezésnél az ilyen finomabb különbségek ugyanakkor nehezen voltak interpretálhatóak. Ez lehetséges azért is, mivel a klaszterezés eredményének interpretálásához kevesebb fogódzó áll rendelkezésünkre, így nem ismerjük a szavak valószínűségét az egyes klaszterekben, csak a szavak súlyából tudunk következtetni. Ugyanakkor az is segítheti a szofisztikáltabb topikok felderítését, hogy a topikmodellezésnél egy bejegyzést több topik is jellemezhet különböző erősséggel. Ez finomabb bontást enged meg, mint a klaszterezés, amely olyan (elég gyakorta előfordulható) helyzetekben, melyekben kevésbé egyértelmű, vektortérben nézve klaszterek határán álló bejegyzéseket kezel, mindenképp csak egyik klaszterbe sorolhatja be azt. Ezzel azonban nehezíti a klaszterek interpretálását is. Ugyanakkor a klaszterezés elég jól teljesített abban, hogy felderítsen a tfidf-súlyozásos szavak alapján olyan klasztereket, amelyek specifikus jellemzőjük miatt jól interpretálhatóak. Ezek a klaszterek jellemzően kevésbé bonyolult bejegyzésekből állnak, melyek témája olvasva is könnyen azonosítható. Ugyanakkor jellemző lehet, hogy ezekben a klaszterekben a bejegyzésekben gyakran előfordulnak olyan szavak, amelyek egyértelműen utalnak a témára. Ilyen a kezelésrezisztens, a bipoláris, vagy a támogatás. Azonban nehézségbe ütközhetett a klaszterelemzés olyan csoportok felderítésénél, ahol a szervezőtematika nem jelenik meg a bejegyzések szóhasználatában annyira expliciten. Egyrészt ez is okozhatott értelmezési nehézségeket egyes klaszterezési megoldásoknál.

Érdekes eredmény ugyanakkor, hogy a Nimrod (2012) kutatásából kialakuló 9 fő depresszióforumon megjelenő témának (tünetek, kapcsolatok, megküzdés, élet, formális ellátás, gyógyszerek, okok, öngyilkosság és munka) legtöbbje jelen klaszterezés során is felderítésére került. Leginkább az öngyilkosság témája olyan, ami bár több klaszterben is felmerült, és a bejegyzések olvasásakor is többször detektálható volt, önálló klaszter mégsem szerveződött a téma köré. Sik (2020) online fórumokon végzett etnográfiájában a depresszió különböző keretezéseit vizsgálta. A három keretezési mód a kapott klaszterstruktúrában is tettenérhető. Míg a klaszterek egy része a gyógyszerekkel és orvosi kezeléssel, vagy más testi problémákkal foglalkozik, jelentős részük inkább a pszichés nehézségekre helyezi a hangsúlyt. A szociális keretezés is felfedezhető ezek mellett például a kapcsolati fókuszú klaszterben, melyben sokszor a múltba visszahúzódó családi vagy partneri kapcsolatokat is a mentális

zavaruk eredőjeként ismertetik a bejegyzésírók. Emellett még egybecseng a klaszterezés eredményével Feldhege, Moessner és Bauer (2020) eredménye is, akik úgy találták, hogy a depresszió fórumokon fellelhető hosszabb bejegyzések inkább múltbéli beszámolókból, történetekből, és a kapcsolatokról szólnak, míg a támogató céllal írt bejegyzések jellemzően valamelyest rövidebbek. Így tehát megállapítható, hogy sikerült olyan klasztereket létrehozni, melyek a korábbi depressziós fórumvizsgálatok alapján várhatóak voltak. Emellett felderítésre került némely olyan klaszter is, amely kevésbé jelenik meg ilyen csoportosítási vizsgálatokban. Ilyen például az alvásproblémák, vagy a bipoláris depresszió klasztere.

6.2.2. Névmáshasználat vizsgálata

A névmások használatából arra nézve szerettem volna levonni következtetéseket, hogy milyen klaszterekben vannak súlyosabban depressziós személyek. Ez a módszer azonban kevéssé vezetett eredményre. A klaszterek interpretálása alapján a 23-as klaszter az, amelyben a legsúlyosabb esetek fokozottabban szerepelhetnek. Ebben a klaszterben kifejezetten önmagukra fókuszáló, befelé fókuszáló, és elkeseredést, reménytelenséget tükröző bejegyzések szerepelnek a klaszterközépponthoz közeli bejegyzések elolvasása és a klaszterben legnagyobb súllyal szereplő szavak alapján. Ugyanakkor ez a névmáshasználat alapján nem jelent meg kiugró klaszterként. Ugyan viszonylag magasabb volt benne az „én” névmások aránya, de a diagnózisban való bizonytalanságot kifejező klaszterben kiemelkedően magasabb volt az „én” szó aránya, és a 23-as klaszterrel hasonló arányt mutatott az enyhébb tünetekről beszámoló, alvászavarokat tartalmazó, valamint a rövid köszönetekből álló klaszter is. Ugyanakkor az egyes szám harmadik személyű névmások arányában is elég magasan szerepelt a 23-as klaszter a többi hosszú beszámolóhoz hasonlóan, pedig a társas eltávolodást éppen ezeknek a névmásoknak az alacsonyabb használata jelezné. A többes szám első személyű névmások hasonlóan nem mutattak a 23-as klaszter súlyosságára mutató eredményt. Kiugróan magasnak az öngyilkosság szó sem mutatkozott a klaszterben.

A névmások megjelenésében az eredmények alapján inkább a klaszter tematikája játszott meghatározóbb szerepet, mint egyszerűen a depressziós állapot súlyossága. Ezt jól példázza, hogy mikor a társas kapcsolatokra utaló névmásokkal vetettem össze az „én” névmást, azt találtam, hogy az alvásproblémákkal foglalkozó klaszterben a legdominánsabb arányaiban az egyes szám első személyű névmás, amely könnyen megmagyarázható azzal, hogy az alvás nem

egy társas kapcsolódással összefüggő tevékenység, és még ha nem is teljesen egyedül zajlik, akkor is egy önálló, személyes állapot, mellyel kapcsolatban természetszerű, hogy a személy csak magára fókuszálva számol be. Hasonló arányt mutatott még a gyógyszeres kezelés klasztere is, mely hasonlóképp egy olyan témakör, amelyben a szövegírónak nincsen sok lehetősége és indoka a másokhoz való kapcsolódásáról beszámolni a téma jellege miatt. Hasonló tendenciák akkor is megfigyelhetők, ha az összes szóhoz viszonyítva vizsgáljuk a társas kapcsolódásra utaló névmások („he", „she" és „we") arányát.

Fokozottabb figyelmet érdemelhet viszont a diagnózisukban bizonytalanok klasztere a névmások szempontjából, hiszen itt a többi hosszú beszámolót tartalmazó klaszterhez (és így az énközpontú, elkeseredett klaszterhez) képest is kiugróan magas volt az egyes szám első személyű névmás használata. Így lehetséges, hogy ebben a klaszterben valamelyest fokozódhat a súlyos depressziós állapot megjelenése, amit magyarázhat is, hogy itt a legtöbb beszámolóban olyan személyek írtak, akik bár komoly tünetekkel rendelkeznek leírásuk alapján, de pontos diagnózissal nem rendelkeznek, és így célzott kezelést sem kaphatnak. Ennek a kezeletlenül hagyott depressziónak, vagy akár egyéb mentális zavarnak a veszélyességét jelezheti az „én" névmás fokozott használata, így erre a klaszterre fokozott figyelmet érdemes fordítani. Gould, Greenberg, Velting és Shaffer (2003) összefoglalója alapján az öngyilkosságot elkövetők több, mint 90%-a szenved valamilyen mentális zavarban, és ezeknek a jórésze kezeletlen. Így egy kezeletlen pszichés betegség fokozott rizikót képes magában hordozni, melyet helyesen detektálhat az egyes szám első személyű névmások fokozott megjelenése. Ezt alátámaszthatja az is, hogy ebben a klaszterben szerepelt szintén arányaiban az egyik legmagasabban az „öngyilkosság" szó. Valamint bár nem volt kiugróan alacsony, de a 3. és 4. ábráról leolvasható, hogy a többi klaszterhez képest inkább azok közé tartozott, ahol alacsony a társas kapcsolódásra utaló névmások aránya.

6.2.3. Scatter/gather módszer hasznosítása

A scatter/gather módszer alkalmazásával jól interpretálható klasztereket kaptam, melyek a fentiek alapján a korábbi kutatásokból kiindulva is várhatóak voltak többnyire. Azonban arra nem nyílt lehetőség, hogy összehasonlítsam az eredményeit a hierarchikus klaszterezéssel kapott eredményekkel, mivel a hierarchikus klaszterezés nem tudott viszonylag egyenletes elemszámú, interpretálható klaszterstruktúrát kialakítani az adatokból. Ugyanakkor a

scatter/gather módszer alkalmazása során megmutatkozott a módszer hasznossága. Mikor egy-egy klasztert bontottam csak tovább két új klaszterre, akkor egyik alkalommal sem sikerült jól értelmezhető kisebb klasztereket létrehozni, azonban mikor 7 egymáshoz hasonlóbb klasztert vontam egybe, és bontottam újra 7 klaszterbe, akkor nem ugyanazt a megoldást kaptam, mint ami az első megoldásban látható volt. Bár megfigyelhető volt, hogy a legtöbb azonos klaszterben lévő elem általában azonos klaszterbe kerül újra, az elemek jelentős része áthelyeződik másik klaszterbe, valamint nem is teljesen ugyanazok az értelmezésű klaszterek alakulnak ki. Így az megállapítható, hogy a klaszterek egy részének újrabontása hasznos megoldás, nem eredményezi ugyanazt, mint ami a teljes minta klaszterezéséből ered. Ugyanakkor mivel a hierarchikus klaszterezéssel nem sikerült megragadni az adatokban rejlő struktúrát, ezért a klaszterszám meghatározásához, és a kezdőpontok kiválasztásához nyújtott előnye nem mutatkozott meg a scatter/gather módszernek az elemzés során.

Bár az az eredeti csoporttagságok ismerete nélkül nehezen alátámasztható, hogy a klaszterek újrabontásával pontosabb eredményt kaptam, mint az eredeti klaszterekből származóval, a klaszterek értelmezhetősége ezt támasztotta alá. Tombros, Villa és Van Rijsbergen (2002) is úgy találták, hogy ha a klaszterek egy részén hajtunk végre újabb klaszterezést, az valamelyest pontosabb klaszterezést eredményezhet, mint ami a teljes minta klaszterezéséből kiolvasható lenne. Hearst és Pedersen (1996) alapján ez annak köszönhető, hogy a kevesebb klaszterre való leszűkítés átalakítja a dokumentumok terét, és így nagyobb az esély arra, hogy az egymáshoz hasonló elemek azonos klaszterbe kerüljenek. Ez annak köszönhető, hogyha kevesebb, és csak egymáshoz bizonyos mértékben hasonlóbb elemek helyezkednek el egy nagyon sok dimenziós vektortérben, akkor az elem legközelebbi szomszédai is máshogy alakulnak. Így többszörösen is érdemes volt a scatter/gather klaszterezés használata mellett dönteni, hiszen a hierarchikus klaszterezés nem hozott kielégítő eredményt, egy egyszerű k-közép klaszterezésnél pedig többet tudott nyújtani az, hogy egyes klasztereket újrabonthattam, ezzel jobb klaszterfelbontást kapva.

6.3. Kutatás eredményeinek felhasználhatósága

A kutatás egyik legerősebb motiváltsága volt, hogy amellett, hogy feltérképezhessem, hogy milyen csoportok tárhatók fel a fórumok bejegyzései közt, azt is megvizsgálhassam, hogy melyik csoportban lehetnek súlyosabban depressziós, vagy akár öngyilkossági veszélyben lévő

személyek. Ennek detektálása fontos segítséget nyújthatna a fórumokon és akár egyéb tereken is arra, hogy felismerhessük a súlyosabban érintett személyeket, így fokozottabb támogatást nyújthassunk. Mindennek öngyilkosság prevenciós szerepe is lehetne akár automatikus algoritmusok segítségével, melyek csoportokba sorolások által detektálják a veszélyeztetett személyeket, de akár egyszerűen azáltal is, hogyha felhívjuk a támogató személyek, fórumadminisztrátorok, és egyéb személyek figyelmét arra, hogy milyen témákról vagy milyen módon beszámoló személyek lehetnek potenciálisan súlyosabb állapotban.

Ehhez a névmások használatában fellelhető különbségeket szerettem volna felhasználni, mivel a szakirodalom alapján ez több kutató munkája által is egységesen alátámasztottan jól működő detektornak volt tekinthető. Azonban jelen kutatásban feltételezhetően mivel kevésbé hasonló szövegeken szerettem volna névmáshasználati eltéréseket detektálni, a névmások használata inkább a téma szóbeli sajátosságai által volt irányított, mint a depresszió súlyossága által, így nem bizonyult teljesen jól működő detektornak jelen esetben, bár sikerült vele elkülöníteni egy olyan klasztert, amely valószínűleg veszélyeztetett, akár öngyilkossági veszélyre utaló bejegyzéseket hordozhat magában fokozottabb mértékben. Érdemes lehet emellett azonban további detektorokkal bővíteni a kutatást, melyek sikeresebben tudják veszélyeztetett személyek egy csoportját elkülöníteni. Ilyen lehet például az Al-Mosaiwi és Johnstone (2018) által tárgyalt „abszolút” szavak köre. Az abszolút szavak olyan szavak, melyek kizárólagosságot hordoznak magukban, nem vesznek számba semmilyen valószínűsítő tényezőt. Ilyenek például a „mindig”, „senki” és a „teljesen” szavak. Mivel az ilyen típusú „abszolút” gondolkodás több mentális betegség, és az öngyilkossági gondolatok velejárója is, ezért a szerzők úgy vélték, hogy ezeknek a szavaknak a használata is jelezheti a szöveg írójának veszélyeztetett állapotát. A szerzők fórumokat vizsgálva úgy találták, hogy a szorongásos és depressziós fórumokon megnövekedett az abszolút szavak használata, az öngyilkossági fórumokon pedig még a két másik fórumnál is nagyobb mértékben megemelkedett. Emellett fontos lehet még más olyan detektorokat keresni az irodalomban, melyek segíthetik a veszélyeztetett személyek fellelését.

Érdemes lehet továbbá más szempontok mentén is megvizsgálni a kialakult klasztereket. Ilyen lehet például, hogy milyen keretezése jelenik meg inkább a depressziónak a különböző klaszterekben, mutatkozik-e eltérés abban, hogy milyen tényezőkkel magyarázzák a betegségüket a bejegyzésírók. Németh és munkatársai (2020) klasszifikációs módszerrel

igyekeztek csoportokba sorolni fórumbejegyzéseket az alapján, hogy orvosi-biológia, pszichológiai, vagy társadalmi keretezésbe ágyazzák a szövegekben a depressziós állapotot. Fontos, lehet ez a tényező, mivel az, hogy miben látják a depressziójuk okát, eredőjét, meghatározhatja, hogy hogyan viszonyulnak hozzá, milyen módon próbálják kezelni, hogyan tekintenek a gyógyulásukra. Így az is érdekes lehet, hogy a különböző klaszterekbe sorolt, így eltérő tematikában, stílusban íródott bejegyzésekben mutatkozik-e a különbség a keretezés tekintetében.

A mentális betegségekkel, így a depresszióval is foglalkozó fórumok vizsgálatát mindenképp érdemes tovább mélyíteni, hiszen ez a platform nyílt és könnyű hozzáférést ad a betegségben érintett személyek gondolataihoz, érzéseihez, vélekedéseihez. Ezek pontosabb megismerése segítheti a kezelési megközelítések módosítását, a betegség természetének alaposabb kiismerését, és a súlyos állapotban lévő személyek kiszűrését is.

7. Irodalomjegyzék

Aggarwal, C. C., & Zhai, C. (2012). A survey of text clustering algorithms. In C. C. Aggarwal & C. Zhai (Eds.), *Mining Text Data* (pp. 77-128). New York, USA: Springer.

Al-Mosaiwi, M., & Johnstone, T. (2018). In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. *Clinical Psychological Science, 6*(4), 529-542. https://doi.org/10.1177/2167702617747074

Bernard, J. D., Baddeley, J. L., Rodriguez, B. F., & Burke, P. A. (2016). Depression, language, and affect: an examination of the influence of baseline depression and affect induction on language. *Journal of Language and Social Psychology, 35*(3), 317-326. https://doi.org/10.1177/0261927X15589186

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python.* Sebastopol, CA: O'Reilliy Media, Inc.

Breault, K. D., & Barkey, K. (1982). A comparative analysis of Durkheim's theory of egoistic suicide. *The Sociological Quarterly, 23*(3), 321-331. https://doi.org/10.1111/j.1533-8525.1982.tb01015.x

Caliński, T., & Harabasz, J. (1974). A Dendrite Method for Cluster Analysis. *Communications in Statistics-theory and Methods, 3*, 1-27. doi:10.1080/03610927408827101.

Campello, R. J., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 160-172). Berlin, Heidelberg: Springer.

Cohan, A., Young, S., & Goharian, N. (2016). Triaging mental health forum posts. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology* (pp. 143-147).

Collins, M., & Singer, Y. (1999). Unsupervised models for named entity classification. In *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.*

Cutting, D. R., Karger, D. R., Pedersen, J. O., & Tukey, J. W. (1992). Scatter/gather: A cluster-based approach to browsing large document collections. In *ACM SIGIR Forum* (Vol. 51, No. 2, pp. 148-159). New York, NY, USA: ACM. https://doi.org/10.1145/3130348.3130362

Dhillon, I. S., & Modha, D. S. (2001). Concept decompositions for large sparse text data using clustering. Machine learning, 42(1), 143-175. https://doi.org/10.1023/A:1007612920971

El-Hamdouchi, A., & Willett, P. (1989). Comparison of hierarchic agglomerative clustering methods for document retrieval. *The Computer Journal, 32*(3), 220-227. https://doi.org/10.1093/comjnl/32.3.220

ESOMAR (2011). ESOMAR guideline on social media research. Retrieved from https://www.esomar.org/uploads/public/knowledge-and-standards/codes-and-guidelines/ESOMAR-Guideline-on-Social-Media-Research.pdf [Accessed 20 February 2021]

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining* (pp. 226–231). Portland, OR, AAAI Press.

Everitt, S. B., Landau, S., Leese, M., & Stahl, D. (2011). Cluster Analysis: 5th Edition. United Kingdom: John Wiley & Sons, Ltd.

Feldhege, J., Moessner, M., & Bauer, S. (2020). Who says what? Content and participation characteristics in an online depression community. *Journal of Affective Disorders, 263*, 521-527. https://doi.org/10.1016/j.jad.2019.11.007

Fernández, A., & Gómez, S. (2008). Solving non-uniqueness in agglomerative hierarchical clustering using multidendrograms. *Journal of Classification, 25*(1), 43-65. https://doi.org/10.1007/s00357-008-9004-x

Gould, M. S., Greenberg, T. E. D., Velting, D. M., & Shaffer, D. (2003). Youth suicide risk and preventive interventions: a review of the past 10 years. *Journal of the American Academy of Child & Adolescent Psychiatry, 42*(4), 386-405. https://doi.org/10.1097/01.CHI.0000046821.95464.CF

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2011). *Survey Methodology.* Wiley.

Hargittai, E. (2020). Potential biases in big data: Omitted voices on social media. *Social Science Computer Review, 38*(1), 10-24. https://doi.org/10.1177/0894439318788322

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. New York: Springer.

Hearst, M. A., & Pedersen, J. O. (1996). Re-examining the Cluster Hypothesis: Scatter/Gather on retrieval results. In *Proceedings of the 19th Annual ACM SIGIR conference* (pp. 76–84). Zurich, Switzerland. https://doi.org/10.1145/243199.243216

Hidaka, B. H. (2012). Depression as a disease of modernity: Explanations for increasing prevalence. *Journal of Affective Disorders, 140*(3), 205–214. doi:10.1016/j.jad.2011.12.036

Horne, J., & Wiggins, S. (2009). Doing being 'on the edge': Managing the dilemma of being authentically suicidal in an online forum. *Sociology of Health & Illness, 31*(2), 170-184. https://doi.org/10.1111/j.1467-9566.2008.01130.x

Huang, A. (2008). Similarity measures for text document clustering. In *Proceedings of the sixth New Zealand Computer Science Research Student Conference.* Christchurch, New Zealand.

Hubert, L. (1974). Approximate evaluation techniques for the single-link and complete-link hierarchical clustering procedures. *Journal of the American Statistical Association, 69*, 698–704. https://doi.org/10.2307/2286004

Ignatow, G., & Mihalcea, R. (2018). *An introduction to text mining: research design, data collection, and analysis.* California, USA: SAGE Publications, Inc.

Lester, D., McSwain, S., & Gunn III, J. F. (2011). A test of the validity of the IS PATH WARM warning signs for suicide. *Psychological Reports, 108*(2), 402-404. https://doi.org/10.2466/09.12.13.PR0.108.2.402-404

Liu, Y., Mostafa, J., & Ke, W. (2007). A fast online clustering algorithm for scatter/gather browsing. Sch. Inf. and Library Sci., Univ. North Carolina, Chapel Hill, NC, USA, Tech. Rep. TR-2007-06.

Kummervold, P. E., Gammon, D., Bergvik, S., Johnsen, J. A. K., Hasvold, T., & Rosenvinge, J. H. (2002). Social support in a wired world: use of online mental health forums in Norway. *Nordic Journal of Psychiatry, 56*(1), 59-65. https://doi.org/10.1080/08039480252803945

McCay-Peet, L., & Quan-Haase, A. (2017). What is Social Media and What Questions Can Social Media Research Help Us Answer?. In L. Sloan & A. Quan-Haase (Eds.), *The SAGE Handbook of Social Medai Research Methods* (pp. 13-26). London, United Kingdom: SAGE Publications Ltd.

McSwain, S., Lester, D., & Gunn III, J. F. (2012). Warning signs for suicide in internet forums. *Psychological Reports, 111*(1), 186-188. https://doi.org/10.2466/12.13.PR0.111.4.186-188

Müller, A. C., & Guido, S. (2017). *Introduction to Machine Learning with Python: A Guide for Data Scientists.* Sebastopol, CA: O'Reilly Media, Inc.

Németh, R., Sik, D., & Katona, E. (2021). The asymmetries of the biopsychosocial model of depression in lay discourses - Topic modelling online depression forums. *SSM – Population Health.* https://doi.org/10.1016/j.ssmph.2021.100785.

Németh, R., Sik, D., & Máté, F. (2020). Machine Learning of Concepts Hard Even for Humans: The Case of Online Depression Forums. *International Journal of Qualitative Methods, 19,* 1-8. doi: 10.1177/1609406920949338

Nimrod, G. (2012). From knowledge to hope: online depression communities. *International Journal on Disability and Human Development, 11(*1). doi:10.1515/ijdhd.2012.009

Nimrod, G. (2013). Online Depression Communities: Members' Interests and Perceived Benefits. *Health Communication, 28*(5), 425–434. doi:10.1080/10410236.2012.691068

Popat, S. K., Deshmukh, P. B., & Metre, V. A. (2017). Hierarchical document clustering based on cosine similarity measure. In *2017 1st International Conference on Intelligent Systems and Information Management (ICISIM)* (pp. 153-159). IEEE. doi: 10.1109/ICISIM.2017.8122166

Porter, M. F. (1980). An algorithm for suffix stripping. *Program, 14*(3), 130–137. https://doi.org/10.1108/eb046814

Rousseeuw, P. J. (1987). Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Computational and Applied Mathematics, 20*, 53–65. doi:10.1016/0377-0427(87)90125-7.

Rude, S., Gortner, E. M., & Pennebaker, J. (2004). Language use of depressed and depression-vulnerable college students. *Cognition & Emotion, 18*(8), 1121-1133. https://doi.org/10.1080/02699930441000030

Saif, H., Fernandez, M., He, Y., & Alani, H. (2014). On stopwords, filtering and data sparsity for sentiment analysis of twitter. In: *LREC 2014, Ninth International Conference on Language Resources and Evaluation. Proceedings.*, pp. 810–817.

Sik, D. (2020). From Lay Depression Narratives to Secular Ritual Healing: An Online Ethnography of Mental Health Forums*. Culture, Medicine, and Psychiatry,* 1-24. https://doi.org/10.1007/s11013-020-09702-5

Sloan, L., Morgan, J., Burnap, P., & Williams, M. (2015) Who Tweets? Deriving the Demographic Characteristics of Age, Occupation and Social Class from Twitter User Meta-Data. PLoS ONE 10(3), e0115545. doi: 10.1371/journal.pone.0115545

Sokal, R. R., & Rohlf, F. J. (1962). The comparison of dendrograms by objective methods. *Taxon, 11*(2), 33-40. doi: 10.2307/1217208

Spinney, L. (2009). European Brain Policy Forum 2009: depression and the european society. *European Psychiatry, 24*(8), 550-551. https://doi.org/10.1016/j.eurpsy.2009.04.001

Stirman, S. W., & Pennebaker, J. W. (2001). Word use in the poetry of suicidal and nonsuicidal poets. *Psychosomatic Medicine, 63*(4), 517-522. doi: 10.1097/00006842-200107000-00001

Subhashini, R., & Kumar, V. J. S. (2010). Evaluating the performance of similarity measures used in document clustering and information retrieval. In: *2010 First International Conference on Integrated Intelligent Computing,* pp. 27-31. IEEE. doi: 10.1109/ICIIC.2010.42

Tombros, A., Villa, R., & Van Rijsbergen, C. J. (2002). The effectiveness of query-specific hierarchic clustering in information retrieval*. Information Processing & Management, 38*(4), 559-582. https://doi.org/10.1016/S0306-4573(01)00048-6

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research, 37*, 141-188. https://doi.org/10.1613/jair.2934

VanderPlas, J. (2017). *Python Data Science Handbook: Essential Tools for Working with Data.* Sebastopol, CA: O'Reilly Media, Inc.

Williams, M. L., Burnap, P., & Sloan, L. (2017). Crime sensing with big data: The affordances and limitations of using open-source communications to estimate crime patterns. *The British Journal of Criminology*, *57*(2), 320-340. doi: 10.1093/bjc/azw031

8. Mellékletek

8.1. Névmásokat tartalmazó adatbázis Buckshot klaszterezés eredménye

Lépés	Kapott klaszterek száma	Legnagyobb klaszterbe utoljára csatlakoztatott klaszter elemszáma	Silhouette érték
1132. lépés	20	28	-0.002151
1133. lépés	19	10	-0.002413
1134. lépés	18	16	-0.002708
1135. lépés	17	12	-0.002940
1136. lépés	16	12	-0.003235
1137. lépés	15	10	-0.003521
1138. lépés	14	12	-0.003717
1139. lépés	13	12	-0.003922
1140. lépés	12	8	-0.004108
1141. lépés	11	48	-0.002797
1142. lépés	10	106	-0.002055
1143. lépés	9	60	-0.001350
1144. lépés	8	72	-0.003575
1145. lépés	7	35	-0.0009564
1146. lépés	6	36	0.002050
1147. lépés	5	16	0.002386
1148. lépés	4	6	0.002202
1149. lépés	3	23	0.002549
1150. lépés	2	191	0.020737

8.2. Névmások eltávolításán átesett adatbázis Buckshot klaszterezés eredménye

Lépés	Kapott klaszterek száma	Legnagyobb klaszterbe utoljára csatlakoztatott klaszter elemszáma	Silhouette érték
1132. lépés	20	5	-0.016448
1133. lépés	19	15	-0.016672
1134. lépés	18	5	-0.016577

Lépés	Kapott klaszterek száma	Legnagyobb klaszterbe utoljára csatlakoztatott klaszter elemszáma	Silhouette érték
1135. lépés	17	8	-0.017383
1136. lépés	16	22	-0.017564
1137. lépés	15	117	-0.016264
1138. lépés	14	12	-0.016233
1139. lépés	13	36	-0.016499
1140. lépés	12	25	-0.016789
1141. lépés	11	13	-0.017242
1142. lépés	10	8	-0.017167
1143. lépés	9	37	-0.016041
1144. lépés	8	28	-0.016327
1145. lépés	7	33	-0.014281
1146. lépés	6	10	-0.014328
1147. lépés	5	13	-0.014172
1148. lépés	4	39	-0.001024
1149. lépés	3	8	0.006461
1150. lépés	2	10	0.021928

8.3. Névmásokat tartalmazó adatbázis teljes mintán futtatott k-közép klaszterezés eredménye (1=közel, 0 = távol)

Klaszter sorszáma	Klaszter elemszáma	Legnagyobb súllyal szereplő 10 szó
0. klaszter	3459	she, her, my, it, me, you, say, go, get, we
1. klaszter	10359	my, me, it, year, get, go, feel, life, time, ve
2. klaszter	1179	ago, 'say', 'minute', 'hour', 'it', 'you', 'my', 'me', 'year', 'sober4life'
3. klaszter	11023	'you', 'it', 'your', 'feel', 're', 'help', 'get', 'know', 'yourself', 'hope'
4. klaszter	5516	'it', 'take', 'effect', 'my', 'med', 'you', 'side', 'me', 'depression', 'anxiety'
5. klaszter	4254	he, him, his, it, me, my, you, we, say, get

Klaszter sorszáma	Klaszter elemszáma	Legnagyobb súllyal szereplő 10 szó
6. klaszter	9440	'it', 'feel', 'me', 'my', 'like', 'you', 'get', 'go', 'thing', 'know'
7. klaszter	3694	im, my, dont, it, ive, me, feel, like, get, go
8. klaszter	6480	'your', 'you', 'it', 'help', 'get', 'depression', 'feel', 'my', 'go', 'hope'
9. klaszter	3077	'we', 'our', 'it', 'you', 'my', 'your', 'get', 'life', 'people', 'feel'
10. klaszter	2967	'sleep', 'day', 'it', 'my', 'night', 'get', 'go', 'today', 'feel', 'take'
11. klaszter	4847	'they', 'people', 'it', 'you', 'them', 'me', 'their', 'my', 'like', 'think'

8.4. Névmásokat nem tartalmazó teljes mintás k-közép klaszterezés eredményei (1=közel, 0 = távol)

	Klaszter elemszáma	Legnagyobb súllyal szereplő 20 szó	Klaszterbe tartozó bejegyzések átlagos közelsége a centroidhoz	Klaszterbe tartozó bejegyzések átlagos közelsége egymáshoz	Klaszter értelmezése
0	5780	'feel', 'like', 'know', 'make', 'get', 'go', 'well', 'really', 'want', 'time', 'thing', 'bad', 'think', 've', 'depression', 'depressed', 'even', 'people', 'way', 'life'	0.2681	0.07171	Hosszú beszámoló, bizonytalanság a diagnózisban
1	11891	'people', 'say', 'like', 'know', 'life', 'think', 'would', 'want', 'thing', 'make', 'one', 'time', 'love', 'go', 'get', 'feel', 'even', 'really', 'never', 'friend'	0.2469	0.06087	Hosszú történetek, kapcsolati fókusszal

	Klaszter elemszáma	**Legnagyobb súllyal szereplő 20 szó**	**Klaszterbe tartozó bejegyzések átlagos közelsége a centroidhoz**	**Klaszterbe tartozó bejegyzések átlagos közelsége egymáshoz**	**Klaszter értelmezése**
2	1240	'treatment', 'depression', 'ect', 'help', 'get', 'medication', 'work', 'go', 'well', 'doctor', 'therapy', 'resistant', 'take', 'year', 'time', 'know', 'try', 'say', 'also', 'would'	0.2679	0.07102	Kezelésrezisztens depresszió, gyógyszerek, elektrokonvulzív terápia.
3	4670	'year', 've', 'time', 'go', 'get', 'ago', 'depression', 'life', 'last', 'feel', 'old', 'like', 'know', 'work', 'since', 'start', 'still', 'anxiety', 'one', 'month'	0.2528	0.06372	Hosszú visszaemlékezések, hangsúlyozva az időtényező.
4	1629	'seem', 'like', 'get', 'thing', 'think', 'feel', 'depression', 'know', 'well', 'help', 'people', 'try', 'time', 'go', 'one', 've', 'say', 'really', 'make', 'life'	0.2400	0.05703	Kicsit rövidebb bejegyzések, érzelmi tünetekről, terapeutájukról (magukról, vagy beteg társukról)
5	7142	'get', 'go', 'work', 'time', 'job', 'thing', 'well', 'try', 'back', 'like', 'know', 'want', 'make', 've', 'feel', 'really', 'take', 'would', 'one', 'think'	0.2373	0.05619	Hosszú történetek. Mozgalmasabbak, kevésbé érzelmi fókusz, munka témája
6	4502	'im', 'dont', 'ive', 'feel', 'like', 'get', 'go', 'know', 'cant', 'want', 'time', 'think', 'really', 'try', 'say', 'thing', 'make', 'well', 'work', 'day'	0.3115	0.09683	Hosszú bejegyzések. Énközpontúság, elkeseredettség, önostorozás

	Klaszter elemszáma	Legnagyobb súllyal szereplő 20 szó	Klaszterbe tartozó bejegyzések átlagos közelsége a centroidhoz	Klaszterbe tartozó bejegyzések átlagos közelsége egymáshoz	Klaszter értelmezése
7	1927	'bipolar', 'diagnosis', 'diagnose', 'mania', 'bpd', 'depression', 'symptom', 'think', 'get', 'episode', 'people', 'like', 'also', 'disorder', 'psychiatrist', 'mood', 'know', 'say', 'one', 'time'	0.2461	0.06008	Bipoláris depresszió
8	7030	'help', 'depression', 'need', 'talk', 'know', 'support', 'get', 'thing', 'find', 'feel', 'go', 'also', 'may', 'try', 'well', 'see', 'think', 'hope', 'like', 'time'	0.2431	0.05895	Támogatás, segítség témája (egy része tényleges támogatás, másik része nem)
9	2658	're', 'feel', 'know', 'go', 'sorry', 'get', 'like', 'well', 've', 'think', 'hope', 'thing', 'say', 'help', 'good', 'll', 'time', 'people', 'really', 'need'	0.2843	0.08047	Válaszok „i'm so sorry" kezdettel (támogatás, tanács, biztatás)
10	2826	'post', 'forum', 'welcome', 'hi', 'hope', 'thread', 'depression', 'support', 'help', 'find', 'sorry', 'people', 'feel', 'glad', 'hello', 'see', 'like', 'well', 'good', 'know'	0.2442	0.05929	Régi tagok üdvözölnek újakat, tanács, vagy iránymutatás fórumhasználathoz
11	3201	'day', 'today', 'get', 'go', 'feel', 'good', 'well', 'take', 'one', 'every', 'time', 'like', 'work', 'bad', 'week', 'make', 'thing', 'try', 'know', 'think'	0.2390	0.05683	Megküzdési technikák, enyhébb tünetek (de egy részük tévesen szerepel itt)

	Klaszter elemszáma	**Legnagyobb súllyal szereplő 20 szó**	**Klaszterbe tartozó bejegyzések átlagos közelsége a centroidhoz**	**Klaszterbe tartozó bejegyzések átlagos közelsége egymáshoz**	**Klaszter értelmezése**
12	3063	'pain', 'symptom', 'test', 'depression', 'get', 'doctor', 'cause', 'take', 'also', 'go', 'blood', 'thyroid', 'level', 'low', 'feel', 'say', 'year', 'help', 'anxiety', 'know'	0.2288	0.0520	Testi fájdalmak, egyéb testi betegségek
13	5045	'take', 'med', 'effect', 'side', 'medication', 'work', 'dose', 'anxiety', 'doctor', 'antidepressant', 'week', 'depression', 'drug', 'get', 'try', 'go', 'start', 'help', 'well', 'time'	0.2389	0.05689	Gyógyszeres kezelés
14	1811	'thank', 'much', 'reply', 'help', 'feel', 'know', 'share', 'really', 'go', 'get', 'say', 'well', 'thanks', 'think', 'hope', 'appreciate', 'post', 'make', 'word', 'try'	0.26746	0.07099	Rövid köszönetek
15	1880	'sleep', 'night', 'take', 'get', 'hour', 'go', 'day', 'feel', 'help', 'wake', 'time', 'well', 'bed', 'work', 'try', 'depression', 'like', 'tire', 'much', 'good'	0.2957	0.0869	Alvásproblémák

Klaszterközéppontok közelsége a teljesmintás klaszterezésnél (1=közel, 0 = távol):

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
0	1	0,789259	0,484074	0,684859	0,669614	0,73946	0,610915	0,513135
1	0,789259	1	0,563272	0,788271	0,745245	0,836229	0,633055	0,610166

	0	1	2	3	4	5	6	7
2	0,484074	0,563272	1	0,563304	0,507982	0,575374	0,416266	0,508272
3	0,684859	0,788271	0,563304	1	0,658486	0,796803	0,582771	0,584498
4	0,669614	0,745245	0,507982	0,658486	1	0,707883	0,527911	0,536189
5	0,73946	0,836229	0,575374	0,796803	0,707883	1	0,628541	0,57169
6	0,610915	0,633055	0,416266	0,582771	0,527911	0,628541	1	0,4349
7	0,513135	0,610166	0,508272	0,584498	0,536189	0,57169	0,4349	1
8	0,703906	0,807519	0,616228	0,701919	0,693325	0,771696	0,57613	0,588623
9	0,601366	0,665346	0,438067	0,54756	0,55266	0,622167	0,440747	0,443824
10	0,500214	0,566786	0,440527	0,508915	0,496525	0,529208	0,416941	0,437696
11	0,63248	0,656709	0,44456	0,634436	0,572605	0,721233	0,519169	0,447033
12	0,576207	0,638165	0,567377	0,648797	0,575532	0,662786	0,476343	0,540154
13	0,571393	0,614172	0,5803	0,64652	0,579148	0,679116	0,484321	0,546113
14	0,492385	0,549512	0,382785	0,480249	0,46436	0,524745	0,425384	0,371965
15	0,481557	0,489367	0,38004	0,4951	0,450884	0,564418	0,409957	0,38285

	8	9	10	11	12	13	14	15
0	0,703906	0,601366	0,500214	0,63248	0,576207	0,571393	0,492385	0,481557
1	0,807519	0,665346	0,566786	0,656709	0,638165	0,614172	0,549512	0,489367
2	0,616228	0,438067	0,440527	0,44456	0,567377	0,5803	0,382785	0,38004
3	0,701919	0,54756	0,508915	0,634436	0,648797	0,64652	0,480249	0,4951
4	0,693325	0,55266	0,496525	0,572605	0,575532	0,579148	0,46436	0,450884
5	0,771696	0,622167	0,529208	0,721233	0,662786	0,679116	0,524745	0,564418
6	0,57613	0,440747	0,416941	0,519169	0,476343	0,484321	0,425384	0,409957
7	0,588623	0,443824	0,437696	0,447033	0,540154	0,546113	0,371965	0,38285
8	1	0,653554	0,673972	0,594536	0,64696	0,648311	0,552136	0,483725
9	0,653554	1	0,499894	0,496393	0,474874	0,477296	0,440023	0,378163
10	0,673972	0,499894	1	0,433196	0,447621	0,437642	0,461702	0,324971
11	0,594536	0,496393	0,433196	1	0,539378	0,572848	0,445973	0,51978
12	0,64696	0,474874	0,447621	0,539378	1	0,667195	0,418018	0,474418
13	0,648311	0,477296	0,437642	0,572848	0,667195	1	0,422099	0,536464

	8	9	10	11	12	13	14	15
14	0,552136	0,440023	0,461702	0,445973	0,418018	0,422099	1	0,335302
15	0,483725	0,378163	0,324971	0,51978	0,474418	0,536464	0,3353	1

8.5. 8-as klaszter kettébontása k-közép klaszterezéssel

	Klaszter elemszáma	**Legnagyobb súllyal szereplő 20 szó**	**Klaszterbe tartozó bejegyzések átlagos közelsége a centroidhoz**	**Klaszterbe tartozó bejegyzések átlagos közelsége egymáshoz**
0	2963	'help', 'get', 'depression', 'need', 'know', 'try', 'go', 'hope', 'feel', 'thing', 'well', 'think', 'find', 'people', 'also', 'like', 'time', 'anxiety', 'one', 'would'	0.2547	0.06453
1	4067	'help', 'support', 'talk', 'depression', 'need', 'may', 'know', 'find', 'thing', 'also', 'see', 'feel', 'go', 'well', 'get', 'like', 'time', 'think', 'try', 'would'	0.2519	0.06325

8.6. 11-es klaszter tovább-bontása k-közép klaszterezéssel

	Klaszter elemszáma	**Legnagyobb súllyal szereplő 20 szó**	**Klaszterbe tartozó bejegyzések átlagos közelsége a centroidhoz**	**Klaszterbe tartozó bejegyzések átlagos közelsége egymáshoz**
0	677	'today', 'day', 'good', 'go', 'feel', 'well', 'tomorrow', 'get', 'yesterday', 'yes', 'like', 'hope', 'bad', 'eat', 'work', 'one', 'take', 've', 'week', 'check'	0.2632	0.06791
1	2524	'day', 'get', 'go', 'feel', 'every', 'take', 'one', 'time', 'well', 'like', 'good', 'work', 'know', 'bad', 'week', 'make', 'try', 'thing', 'think', 'really'	0.2510	0.06264

8.7. 4-es klaszter tovább-bontása k-közép klaszterezéssel

	Klaszter elemszáma	**Legnagyobb súllyal szereplő 20 szó**	**Klaszterbe tartozó bejegyzések átlagos közelsége a centroidhoz**	**Klaszterbe tartozó bejegyzések átlagos közelsége egymáshoz**
0	842	'seem', 'get', 'like', 'help', 'know', 'feel', 'try', 'well', 'depression', 'go', 'think', 'people', 'ca', 'really', 'much', 'work', 'want', 'one', 'thanks', 'find'	0.2388	0.05592
1	787	'seem', 'like', 'thing', 'think', 'time', 'get', 'depression', 've', 'feel', 'one', 'people', 'know', 'life', 'well', 'say', 'go', 'try', 'make', 'would', 'even'	0.2562	0.06444

8.8. Hosszú klaszterek 7 klaszteres újraklaszterezése

	Klaszter elemszáma	**Legnagyobb súllyal szereplő 20 szó**	**Klaszterbe tartozó bejegyzések átlagos közelsége a centroidhoz**	**Klaszterbe tartozó bejegyzések átlagos közelsége egymáshoz**	**Klaszter értelmezése**
20	3408	'day', 'today', 'get', 'go', 'feel', 'well', 'take', 'one', 'good', 'every', 'time', 'like', 'work', 'bad', 'thing', 'week', 'try', 'make', 'know', 'think'	0.2385	0.05661	Enyhébb tünetek
21	5849	'feel', 'like', 'get', 'know', 'make', 'go', 'well', 'really', 'want', 'time', 'thing', 've', 'think', 'bad', 'depressed', 'depression', 'even', 'way', 'ca', 'people']	0.2655	0.07033	Bizonytalanság a diagnózisban

	Klaszter elemszáma	**Legnagyobb súllyal szereplő 20 szó**	**Klaszterbe tartozó bejegyzések átlagos közelsége a centroidhoz**	**Klaszterbe tartozó bejegyzések átlagos közelsége egymáshoz**	**Klaszter értelmezése**
22	2916	'say', 'ago', 'get', 'want', 'go', 'like', 'know', 'would', 'thing', 'minute', 'think', 'one', 'time', 'feel', 'people', 'make', 'hour', 'really', 'talk', 'friend'	0.2452	0.05980	Másik személy mentális betegségéről
23	3263	'im', 'ive', 'feel', 'get', 'go', 'like', 'dont', 'cant', 'know', 'time', 'try', 'want', 'think', 'work', 'well', 'make', 'thing', 'really', 'day', 'year'	0.3204	0.1024	Énközpontú, elkeseredett
24	2343	'dont', 'im', 'feel', 'ive', 'know', 'want', 'like', 'get', 'cant', 'go', 'think', 'really', 'time', 'people', 'thing', 'make', 'life', 'say', 'friend', 'even'	0.3075	0.09418	Énközpontú elkeseredett
25	10003	'get', 'year', 'go', 'work', 'time', 've', 'job', 'try', 'well', 'like', 'back', 'depression', 'know', 'thing', 'take', 'start', 'feel', 'want', 'really', 'one'	0.2387	0.05688	Mozgalmasabb, kevésbé érzelmifókusz. Munka témája gyakrabban.
26	11033	'people', 'life', 'like', 'think', 'thing', 'know', 'would', 'want', 'make', 'one', 'get', 'time', 'love', 'go', 'feel', 'even', 'really', 'way', 'never', 'friend'	0.2439	0.05939	Kapcsolatokról

Klaszterközéppontok egymáshoz való hasonlósága:

	20	**21**	**22**	**23**	**24**	**25**	**26**
20	1	0.645609	0.591599	0.498242	0.539233	0.730877	0.666175
21	0.645609	1	0.662304	0.563463	0.652778	0.752635	0.794226
22	0.591599	0.662304	1	0.514999	0.606643	0.741362	0.799651
23	0.498242	0.563463	0.514999	1	0.721974	0.596095	0.579158
24	0.539233	0.652778	0.606643	**0.721974**	1	0.661545	0.690853
25	0.730877	0.752635	0.741362	0.596095	0.661545	1	0.851890
26	0.666175	0.794226	0.799651	0.579158	0.690853	0.851890	1

8.9. Eredeti és újrabontás utáni klasztertagságok összevetése

Eredeti / új	**20**	**21**	**22**	**23**	**24**	**25**	**26**
0	3	**5221**	143	13	226	48	126
1	9	23	1885	3	387	385	**9199**
3	69	136	177	39	111	**3863**	275
4	86	202	137	28	60	396	720
5	227	251	462	45	215	**5258**	684
6	5	7	47	**3125**	**1285**	21	12
11	**3009**	9	65	10	59	32	17

8.10. Végső klaszterek

	Klaszter elemszáma	**Legnagyobb súllyal szereplő 20 szó**	**Klaszterbe tartozó bejegyzések átlagos közelsége a centroidhoz**	**Klaszterbe tartozó bejegyzések átlagos közelsége egymáshoz**	**Klaszter értelmezése**
2	1240	'treatment', 'depression', 'ect', 'help', 'get', 'medication', 'work', 'go', 'well', 'doctor', 'therapy', 'resistant', 'take', 'year', 'time', 'know', 'try', 'say', 'also', 'would'	0.2679	0.07102	Kezelésrezisztens depresszió, gyógyszerek, elektrokonvulzív terápia.
7	1927	'bipolar', 'diagnosis', 'diagnose', 'mania', 'bpd', 'depression', 'symptom', 'think', 'get', 'episode', 'people', 'like', 'also', 'disorder', 'psychiatrist', 'mood', 'know', 'say', 'one', 'time'	0.2461	0.06008	Bipoláris depresszió
8	7030	'help', 'depression', 'need', 'talk', 'know', 'support', 'get', 'thing', 'find', 'feel', 'go', 'also', 'may', 'try', 'well', 'see', 'think', 'hope', 'like', 'time'	0.2431	0.05895	Támogatás, segítség témája (egy része tényleges támogatás, másik része nem)
9	2658	're', 'feel', 'know', 'go', 'sorry', 'get', 'like', 'well', 've', 'think', 'hope', 'thing', 'say', 'help', 'good', 'll', 'time', 'people', 'really', 'need'	0.2843	0.08047	Válaszok i'm so sorry kezdettel (támogatás, tanács, biztatás)
10	2826	'post', 'forum', 'welcome', 'hi', 'hope', 'thread', 'depression', 'support', 'help', 'find', 'sorry', 'people', 'feel', 'glad', 'hello', 'see', 'like', 'well', 'good', 'know'	0.2442	0.05929	Régi tagok üdvözölnek újakat, tanács, vagy iránymutatás fórumhasználathoz

	Klaszter elemszáma	**Legnagyobb súllyal szereplő 20 szó**	**Klaszterbe tartozó bejegyzések átlagos közelsége a centroidhoz**	**Klaszterbe tartozó bejegyzések átlagos közelsége egymáshoz**	**Klaszter értelmezése**
12	3063	'pain', 'symptom', 'test', 'depression', 'get', 'doctor', 'cause', 'take', 'also', 'go', 'blood', 'thyroid', 'level', 'low', 'feel', 'say', 'year', 'help', 'anxiety', 'know'	0.2288	0.0520	Testi fájdalmak, egyéb testi betegségek
13	5045	'take', 'med', 'effect', 'side', 'medication', 'work', 'dose', 'anxiety', 'doctor', 'antidepressant', 'week', 'depression', 'drug', 'get', 'try', 'go', 'start', 'help', 'well', 'time'	0.2389	0.05689	Gyógyszeres kezelés
14	1811	'thank', 'much', 'reply', 'help', 'feel', 'know', 'share', 'really', 'go', 'get', 'say', 'well', 'thanks', 'think', 'hope', 'appreciate', 'post', 'make', 'word', 'try'	0.26746	0.07099	Rövid köszönetek
15	1880	'sleep', 'night', 'take', 'get', 'hour', 'go', 'day', 'feel', 'help', 'wake', 'time', 'well', 'bed', 'work', 'try', 'depression', 'like', 'tire', 'much', 'good'	0.2957	0.0869	Alvásproblémák
20	3408	'day', 'today', 'get', 'go', 'feel', 'well', 'take', 'one', 'good', 'every', 'time', 'like', 'work', 'bad', 'thing', 'week', 'try', 'make', 'know', 'think'	0.2385	0.05661	Enyhébb tünetek

	Klaszter elemszáma	**Legnagyobb súllyal szereplő 20 szó**	**Klaszterbe tartozó bejegyzések átlagos közelsége a centroidhoz**	**Klaszterbe tartozó bejegyzések átlagos közelsége egymáshoz**	**Klaszter értelmezése**
21	5849	'feel', 'like', 'get', 'know', 'make', 'go', 'well', 'really', 'want', 'time', 'thing', 've', 'think', 'bad', 'depressed', 'depression', 'even', 'way', 'ca', 'people']	0.2655	0.07033	Bizonytalanság a diagnózisban
22	2916	'say', 'ago', 'get', 'want', 'go', 'like', 'know', 'would', 'thing', 'minute', 'think', 'one', 'time', 'feel', 'people', 'make', 'hour', 'really', 'talk', 'friend'	0.2452	0.05980	Másik személy mentális betegségéről
23	5606	'im', 'dont', 'ive', 'feel', 'get', 'like', 'go', 'know', 'want', 'cant', 'think', 'time', 'really', 'try', 'thing', 'make', 'work', 'say', 'well', 'life'	0.2931	0.0857	Énközpontú, elkeseredett
25	10003	'get', 'year', 'go', 'work', 'time', 've', 'job', 'try', 'well', 'like', 'back', 'depression', 'know', 'thing', 'take', 'start', 'feel', 'want', 'really', 'one'	0.2387	0.05688	Mozgalmasabb, kevésbé érzelmifókusz. Munka témája gyakrabban.
26	11033	'people', 'life', 'like', 'think', 'thing', 'know', 'would', 'want', 'make', 'one', 'get', 'time', 'love', 'go', 'feel', 'even', 'really', 'way', 'never', 'friend'	0.2439	0.05939	Kapcsolatokról

Klaszterközéppontok egymáshoz való hasonlósága

	2	**7**	**8**	**9**	**10**	**12**	**13**	**14**
2	1	0.508272	0.616180	0.438067	0.440527	0.567377	0.580300	0.382785
7	0.508272	1	0.588585	0.443824	0.437696	0.540154	0.546113	0.371965
8	0.616180	0.588585	1	0.653546	0.673974	0.646929	0.648275	0.552132
9	0.438067	0.443824	0.653546	1	0.499894	0.474874	0.477296	0.440023
10	0.440527	0.437696	0.673974	0.499894	1	0.447621	0.437642	0.461702
12	0.567377	0.540154	0.646929	0.474874	0.447621	1	0.667195	0.418018
13	0.580300	0.546113	0.648275	0.477296	0.437642	0.667195	1	0.422099
14	0.382785	0.371965	0.552132	0.440023	0.461702	0.418018	0.422099	1
15	0.380040	0.382850	0.483710	0.378163	0.324971	0.474418	0.536464	0.335302
20	0.453332	0.455746	0.603934	0.503971	0.439545	0.548670	0.582926	0.451287
21	0.488479	0.516047	0.706537	0.603614	0.502236	0.581988	0.578329	0.492837
22	0.492319	0.522775	0.677712	0.576072	0.480919	0.569287	0.546444	0.488986
23	0.441598	0.462392	0.615014	0.472292	0.441613	0.505003	0.511575	0.449292
25	0.601283	0.606801	0.777246	0.614444	0.544809	0.692077	0.704759	0.527136
26	0.567178	0.614233	0.814878	0.668193	0.572156	0.640079	0.617187	0.547491

	15	**20**	**21**	**22**	**23**	**25**	**26**
2	0.380040	0.453332	0.488479	0.492319	0.441598	0.601283	0.567178
7	0.382850	0.455746	0.516047	0.522775	0.462392	0.606801	0.614233
8	0.483710	0.603934	0.706537	0.677712	0.615014	0.777246	0.814878
9	0.378163	0.503971	0.603614	0.576072	0.472292	0.614444	0.668193
10	0.324971	0.439545	0.502236	0.480919	0.441613	0.544809	0.572156
12	0.474418	0.548670	0.581988	0.569287	0.505003	0.692077	0.640079
13	0.536464	0.582926	0.578329	0.546444	0.511575	0.704759	0.617187
14	0.335302	0.451287	0.492837	0.488986	0.449292	0.527136	0.547491
15	1	0.527379	0.486717	0.440517	0.431291	0.563092	0.490636
20	0.527379	1	0.645609	0.591599	0.553474	0.730877	0.666175
21	0.486717	0.645609	1	0.662304	0.644764	0.752635	0.794226

	15	**20**	**21**	**22**	**23**	**25**	**26**
22	0.440517	0.591599	0.662304	1	0.593697	0.741362	0.799651
23	0.431291	0.553474	0.644764	0.593697	1	0.669370	0.671447
25	0.563092	0.730877	0.752635	0.741362	0.669370	1	0.851890
26	0.490636	0.666175	0.794226	0.799651	0.671447	0.851890	1