Eötvös Loránd Tudományegyetem

Társadalomtudományi Kar

MESTERKÉPZÉS

Facebook posztokra alapozott nyelvhasználati kutatás

Konzulens:

Dr. Kmetty Zoltán

Készítette:

Kelemen Sára Bernadett
QH6RJT
Survey statisztika szak

2021. április

Tartalmi kivonat

A szerzők tetszőleges demográfiai változó mentén való kategorizálása gyakori klasszifikációs probléma napjainkban. A nem, életkor vagy akár nyelvhasználat szerinti osztályozás a cégek esetében segítheti a fogyasztók hatékonyabb megismerését. Például a gyanús felhasználók kiszűrését is gyorsabbá és hatékonyabbá teheti az automatizált, nyelv alapú bűnügyi profilozás.

Online szövegelemzések esetében is gyakran elhangzó kritika, hogy a szövegek íróiról nincs semmilyen háttérinformációnk. Ha nyelvhasználatuk alapján képesek lennénk nemet becsülni, értékes plusz tudást szerezhetnénk. Ezért a saját kutatásomhoz hasonló, online szövegekre alapozott klasszifikációs kutatások fontosak, a kapott eredmények önmagukban vett hasznosíthatóságát nem szabad lebecsülni.

A közösségi média oldalakról elérhető strukturálatlan adatok elemzése azért is „kifizetődő", mert társadalmunk nagy része megtalálható ezeken a platformokon. Ezen kívül kis erőfeszítéssel hatalmas adatmennyiség állhat rendelkezésünkre, figyelembe véve például, hogy Facebook-ról akár néhány kattintással letölthető a valaha feltöltött összes adatunk.

Blogolás, chatelés, e-mail írás során a beszélt és írott kommunikáció mellett megjelent „hibrid" nyelvi variációt alkalmazzuk. Jellemzőit tekintve a Facebook-on születő bejegyzések is előbbiek közé sorolhatóak, ugyanis laza nyelvhasználat, kötetlen hangvétel, rövidítések gyakori használata, és egyszerűség, ami ismertetőjegyeik közt szerepel.

A nők és férfiak nyelvhasználata közötti különbségekre épülő kutatások egészen Robin Lakoff-ig vezethetőek vissza. Őt követően számos kutatás született a témában, melyek különféle blogbejegyzésekből (Argamon et al., 2003; Schler et al., 2006; Zhang & Zhang 2010), tweet-ekből (Fink et al., 2012; Markov et. al., 2017, Aragón & López, 2018) indultak ki. Facebook posztokra alapozott vizsgálatokkal, sokkal ritkábban, de szintén találkozhatunk (Rangel & Rosso, 2013; Sap et al., 2014). Noha az eredmények néha kissé ellentmondásosak, – például egyes kutatók szerint a nők, míg mások szerint a férfiak használnak több személyes névmást – úgy tűnik abban már mindenki egyetért, hogy van különbség a két nem közt nyelvhasználatuk terén.

Diplomamunkám egyik célja 110 nő és 37 férfi Facebook posztjainak felhasználásával annak a kérdésnek a megválaszolása volt, hogy az alkalmazott klasszifikációs modellek közül melyik az, amelyik a legsikersebben szét tudja választani a nemeket nyelvhasználatuk alapján. Ennek keretén

belül négy felügyelt gépi tanulási módszert vizsgáltam: Logisztikus Regresszió, Support Vector Machine, Random Forest és Naiv Bayes. A hiperparaméterek legjobb kombinációjának megtalálását követően hangoltam az egyes hiperparamétereket, majd a modellek teljesítményét a teszthalmazon összetett mutatók segítségével (precision, recall, f1-score) értékeltem. Abból kifolyólag, hogy kiegyensúlyozatlan osztályeloszlás esetén az accuracy rossz választásnak számít, a négy módszer eredményei közül a legeslegjobbat a teszthalmazon elért ROC görbe alatti értékek alapján választottam ki. Összességében, saját adataimon a legjobb osztályozónak a Bernoulli Naiv Bayes modell bizonyult. A rossz kategóriába sorolt emberek posztjait megvizsgálva, nem találtam olyan általános, minden egyénnél megjelenő domináns témákat, melyeket be lehetne azonosítani a férfiként klasszifikált női felhasználók esetében, sem a nőként klasszifikált férfiak esetében.

Kutatásom során arra is kitértem, hogy mely szófajok, illetve emoticonok és emojik a legnépszerűbbek a férfiak és nők körében. Azt tapasztaltam, hogy nemek szerint bontva nincs különbség, közel azonos arányban használják a négy leggyakoribb, (főnév, melléknév, ige és határozószó) és az összes többi szófajt. Leggyakoribb emoticonok mind a nők, mind a férfiak esetében a derűs hangulat kifejezést szolgálóak voltak. A nők azonban szignifikánsan jelentősebb arányban használták a 'mosolygósarc'-ként kódolt hangulatjelet. A leggyakrabban használt emoji aránya meg sem közelítette a leggyakrabban használt emoticonét. De itt is a jókedvet szimbolizáló „arcok" voltak a legelterjedtebbek. Az egy bejegyzésen belüli emoji csoportosítás inkább a nők esetében volt megfigyelhető.

Kulcsfogalmak

közösségi média, Facebook, írott beszélt nyelv, nyelvhasználat, nem, számítógépes szövegelemzés, felügyelt tanulás, klasszifikáció

Tartalomjegyzék

1 Bevezetés

Az internet térhódításával, valamint a digitális forradalomnak[1] köszönhetően egyre nagyobb mennyiségű elektronikus szöveg válik hozzáférhetővé. Míg a korábban elérhető szöveges adatforrások szerzői jelentős részben írók és újságírók voltak, mára már ezek jellemzően „egyszerű" emberek megnyilvánulásai, hiszen akárki által írt bármilyen témájú szöveg nyilvánosan elérhető az interneten. A befogadók tartalom előállítókká avanzsáltak, míg az olvasók írókká. Műveltségtől, anyagi helyzettől, szakértelemtől függetlenül bárki, a másikkal egyenlő félként leírhatja véleményét.

A Statista, piac- és fogyasztói adatokra szakosodott vállalat, 2020. júliusában közölt becslései[2] alapján 2021-ben 3.78 milliárdra tehető a közösségi média felhasználóinak száma világszerte. A közösségi oldalak, mint például Facebook és Twitter, felhasználói által hagyott „digitális lábnyomok"[3] elemzése egyre népszerűbbé válik, hiszen relatív egyszerűen elérhető, hatalmas adatmennyiségről van szó. Számos szövegelemzési módszert alkalmazó tanulmány készül az előbb említett platformok posztjainak felhasználásával – a számos közül egy témakört kiragadva – például depresszióban szenvedő egyének vizsgálata (De Choudhury, 2013 és Eichstaedt, 2018). A felhasználók posztjaiból kiinduló, az ő nyelvhasználati szokásaikat kutató, klasszifikációt alkalmazó cikkek száma is megnövekedett. Ezekben rendszerint Twitter-ről, vagy különböző blogokról származó szövegeket elemeznek (Fink et al., 2012; Argamon et al., 2003), úgy tűnik mintha a Facebook posztok háttérbe szorulnának. Egyrészt emiatt is vált témaválasztásom egyik mozgatórugójává éppen ez a platform. Kíváncsi voltam rá, vajon mennyire alkalmasak a Facebookról származó posztok ennek a témakörnek a tanulmányozására. A fellelhető szakirodalmak elsősorban angol nyelvűek, és jelentős részük angol vagy más, magyartól eltérő, idegen nyelvű posztokon keresztül vizsgálja a felhasználók nyelvhasználati szokásait. Ennek valószínű oka, hogy a magyart viszonylag nehéz nyelvnek titulálják. Tény, hogy például az angol

[1] A 20. század végétől kezdő, számítógépek és a digitalizálás által kiváltott áttörést értjük digitális forradalom alatt. A kifejezés a számítástechnikai és távközlési eszközök, a számítógép és a telefon elterjedésével járó hatások leírásával fejezhető ki. (Karvalics, 2012)

[2] https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

[3] Digitális lábnyomként szokás utalni minden digitális tevékenység után fennmaradó adatösszességre, amelyből következtetni lehet a tevékenységre, ami alapján be lehet az egyes felhasználókat azonosítani (Weaver & Gahegan, 2010).

nyelvvel összevetve, amelynek nagyon egyszerű a strukturáltsága, a magyar nyelv sokkal változatosabb, emiatt több eshetőségre kell felkészíteni a programot, mint más nyelvek esetében. Egyes sajátosságai miatt nehéz automatizált szövegelemzési módszerekkel vizsgálni. Például agglutináló-ragasztó jellege miatt, azaz, hogy a jelentést elsősorban a tőhöz kapcsolt toldalékokkal fejezzük ki. A netnyelvről, valamint arról, hogy a fiatalok csevegéssel töltött ideje, hogyan befolyásolja élőbeszédüket számos értekezlet elérhető magyar nyelven is. Előbbiek azonban inkább csak történeti összefoglalók, mindenféle saját kutatást nélkülözve. Utóbbiak pedig valamilyen kvantitatív interjús megkérdezésre és különböző iskolai feladatként írott fogalmazások analizálásra korlátozódnak. Az utóbbi években azonban egyre inkább találkozhatunk beszéltnyelvi leiaratozásokra alapozott analízisekkel (Vincze et al., 2021).

Dolgozatom célja egy klasszifikációs projekt keretein belül annak a kérdésnek a megválaszolása, hogyan lehet megkülönböztetni a nemeket nyelvhasználatuk alapján. Két főbb részre bontható: az elméleti háttér és a kutatás ismertetése. Az első tartalmi részben kitérek az írott beszélt nyelv fogalmára és jellemzőinek bemutatásával próbálom alátámasztani, hogy miért gondolom úgy, hogy a Facebook posztokat is ide sorolhatjuk. Bemutatom a klasszifikációs modellek illesztésének lépéseit, alátámasztva indokolt használatát néhány megelőző kutatás segítségével. A legrelevánsabb cikkeket és azok eredményeit ezt követőn részletesebben be is mutatom.

A második nagyobb tartalmi részben legelőször is bemutatom az általam használt adatbázist. Ismertetem az adatok előkészítésének lépéseit, a kiválasztott klasszifikációs modellek eredményeinek szemléltetése előtt pedig azoknak elméleti hátteréről is részletesebb leírást nyújtok. Legvégül pedig tárgyalom az általam felállított modellek teljesítményét, korlátait.

2 Virtuális írásbeliség

2.1 Írott beszélt nyelv – netnyelv

A 20. század utolsó negyedében a digitalizáció hatására a beszélt és írott kommunikáció mellett megjelent egy „hibrid” nyelvi variáció. A közvetítő közeg ebben az esetben a hálózat, a technika. Ez a nyelvi variáció látszatra formálisan írásbeli, de valójában a szóbeli közlésfolyamat funkcióját tölti be, átmenetet képez a két fő kommunikációs típus közt. Egyre több időt töltünk a számítógépek előtt, a szóbeli kommunikációt háttérbe szorítja az írásbeli. Az írott nyelv egyre inkább beszélt jellegűvé válik. A netnyelv funkcióját és tartalmát tekintve a szóbeliséghez közelebb áll. Ebben a tekintetben a chat és az e-mail az, amelyek legközelebb állnak, míg a weboldalak a legtávolabb. Az élőbeszéddel ellentétben itt a szó “nem száll el”, a párbeszéd korábbi részei visszakereshetőek. A netnyelv folyamatosan változik, nincsen végleges formába szorítva (Andó, 2010).

Chatelés során, ahogy a hétköznapi dialógusok esetében gyakran tapasztalhatjuk, nem találkozunk nagy horderejű, fontos témákkal. A lényegi mondanivaló eltűnik, funkciótlan mondatokat használunk. Előfordul, hogy mások gondolatait osztjuk meg, nem törődve az átfogalmazással és értelmezéssel, és ezek nélkül a szövegek híján vannak a logikai felépítésnek, nincs gondolati rendezettség.

Az elektronikus kommunikációként is emlegetett csevegés formát laza nyelvhasználat, kötetlen hangvétel, egyszerűség, rögtönzés, rövidítés és végletes korrigálatlanság jellemzi. Andó Évát idézve: „… kevesebb az elvont főnév, több a ragozott igealak, nagyobb arányban találunk személyes formákat (pl. névmások).” (Andó, 2010, 39. o.). Nem fordítunk akkora figyelmet a mondatszerkesztésre, helyesírásra. Papír alapú íráskor rá vagyunk kényszerítve arra, hogy papírra vetendő gondolatainkat előre átgondoljuk, a korrekció nehézkessége miatt. Ezzel szemben a számítógépes gépelés esetében annyiszor élünk a javítás lehetőségével ahányszor csak szeretnénk. Röviden, néha ködösen fogalmazunk, csupán a megértéshez éppen elegendő információt adjuk át, csak saját nézőpontunkat érvényesítjük. Nagyon ritkán találkozunk jól kidolgozott, több tagmondatból álló alárendelt összetett mondatokkal. A rövidségnek azonban lehetnek mögöttes okai is - egyrészről előfordulhat, hogy a begépelhető karakterek száma

korlátozva van, másrészt, ha lépést akarunk tartani beszélgető partnereinkkel, nincs időnk hosszas fogalmazásokat gépelni. Kommunikációnk felgyorsult a technika gyorsulásának köszönhetően. A tempó tartása sokkal fontosabbá vált, mint a minőség. Internetes kommunikációnkat egyszere jellemzi a közvetettség és közvetlenség, egyidejűleg személyes és nyilvános, monologikus és dialogikus. Alkalmazott stílusunkat nagyban befolyásolja, hogy kivel és milyen környezetben társalgunk. A mi feladatunk, hogy ismerjük az adott helyzethez kötődő adekvát nyelvhasználatot, hogy a kontextusnak megfelelő írásmódot válasszuk.

2.2 Nyelvi sajátosságok

Chateléskor, blogoláskor, kommentek vagy posztok írásakor, de szintúgy akár e-mail írásakor, jellemzően a feladó nem látja a vevőt. Olyan dolgokat is szavak által kell kifejezni, amit adott esetben, szemtől szembeni kommunikáció esetén nem tennénk.

A nyelvi jeleket kísérő nem nyelvi kifejezőeszközök, mint a gesztus és mimika hiánya ösztönözte a ma már mindennapjaink szerves részévé vált emoticonok és emojik megjelenését, melyektől elvárt, hogy képesek legyenek jól dekódolható érzelemkifejezésre. Gondolataink, érzelmeink kifejezésekor az emoticonokra és emojikra támaszkodunk. Ezek az írásjelekből vagy grafikus szimbólumokból álló hangulatjelek kontextusfüggők. Gyakori, hogy az egyén hangulata vagy az üzenet tartalma nem vág egybe az alkalmazott emoticonnal/emojival. Sokszor csak adott szövegkörnyezetben derül ki róluk, hogy pontosan milyen érzelmet, szándékot akart vele adott személy kifejezni. Manapság már funkcionális a megjelenésük, a nyelvhasználati formánk rögzült elemeivé váltak. A hangulatjelek mellett különböző, mára már szintén bevett módszerek jelentek meg a hangsúlyos szavak kiemelésére. Ilyen például mikor adott szó minden betűjét nagybetűvel írjuk le (ez a kiabálás, vagy az emelkedett hangnem jeleként is értelmezhető), idézőjelek közé tesszük a releváns részt, vagy épp betűtöbbszörözést alkalmazunk (Érsok, 2003).

Ha nem privát chat beszélgetést folytatunk, hanem például egy poszt alá kommentelünk, előfordulhat, hogy a tematikusan összefüggő beszélgetések nem egyben jelennek meg, megtörik a linearitás. Bár ez a probléma a privát csevegések esetében is előfordulhat, főleg, ha beszélgetőpartnerünk gondolatait nem koherensen begépelve küldi el, hanem gondolatait tördeli, egy-egy mondat, reakció után entert nyom. Ilyenkor előfordulhat, hogy a másik fél adott

pillanatban megjelenő részletre reagál, ami által a beszélgetés nehezebben követhetővé válik az esetlegesen oda nem illő, zavaróan ható kijelentések miatt. Többszereplős beszélgetések esetében, mint amikor egy poszthoz több személy is hozzászól, nehéz nyomon követni, hogy egyes komment épp melyik résztvevőnek szólt. Ilyen esetekben az egyértelműségre való törekvésként a felhasználót megszólítással próbálják evidensé tenni, azáltal, hogy az üzenet elé beszúrják a címzett nevét.

2.3 Nyelvhasználat

A fiatalok nyelvhasználata mindig is különbözött a felnőttekétől. Minden nemzedék esetében elmondható, hogy beszédszokásainkat folyamatosan alakítgatjuk korunk előrehaladtával. Nyelvhasználatunk stílusa nagyban függ életkorunktól. Figyeljünk meg egy gyermeket, egy középiskolást és egy felnőttet, mindannyiuknál más és más kifejezésmódot fogunk tapasztalni. Más módot találnak elképzeléseik közvetítésére, más szókinccsel, nyelvi korpusszal rendelkeznek. Az általános és középiskolákban egyre inkább háttérbe szorul az olvasás. A gyerekek a média által sugallt nyelvi viselkedésekből tanulnak, a nyelvi magatartást, nyelvi mintákat is innen veszik, amelyek azonban nem feltétlenül tartoznak a követendő példák közé. Noha a net nyelve keveri nyelvhasználatában a hivatalos, közéleti és szakmai stílust, a nyelvi igényesség nem mindig fedezhető fel.

Buda Zsófia 11 és 20 év közötti személyekkel végzett kérdőív alapú felmérésében többek közt azt vizsgálta, hogy az internetezéssel eltöltött idő vajon korrelál-e a fiatalok fogalmazási képességeivel. Eredményként azt kapta, hogy azok az egyének, aki 40 óránál többet töltenek internetezéssel rosszabb helyesírással és gyengébb nyelvi kifejezéskészlettel rendelkeznek. (Az online töltött órák csökkenésével azonban nem növekedett egyenes arányban a kapott pontozási átlag.) Azok, akik idejük jelentős részét online csevegéssel töltik nagyobb eséllyel alkalmazzák az erre jellemző nyelvi sajátosságokat élőbeszédük során is (Buda, 2011).

Az interneten keresztül történő egymással való csevegés során takarékoskodni akarunk a karakterekkel, ezzel is gyorsítva a kommunikációt, emiatt újabbnál újabb formákat találunk ki (Bódi, 2004; Balázs, 2005; Laczkó, 2007):

1) a szó második része hangsúlyos
2) szó eleje és vége alkotja az újonnan létrehozott szóalakot
3) szavak kezdőbetűiből álló rövidítések
4) szótagok vagy összetételi tagok kezdőbetűkkel való rövidítése
5) számok és betűkombinációk
6) kétjegyű mássalhangzók kihagyása
7) pongyola kiejtésű alakváltozatok
8) összeolvadásos formák
9) egy-egy szó vagy toldalék jellel való rögzítése
10) angol nyelvi szavak
    a) magyar toldalékolással
    b) elangolosodás
    c) angol szavak rövidítése betű és számkombinációval

Hogy a helyesírási szabályok gyakori figyelmen kívül hagyása nyelvi romlásnak tekinthető-e (Sík, 2001), vagy a szabad helyesírás-gyakorlathoz mindenkinek megvan a saját joga (Boda, 2007), a szakembereknek különbözik a véleménye. Személy szerint én azzal értek egyet, hogy ez tudatos nyelvi jelenség. Lehet, hogy valaki sietségből hagy ki, téveszt össze betűket, más lehet azért nem ír helyesen, mert ezt tartja populárisan követendőnek. Azonban mindenkinek megvan rá a lehetősége, hogy a közzétett szövegét átjavítsa. Hogy ezt megteszi vagy sem, az rajta múlik.

Másik, a nyelvészeket megosztó téma, az idegen szavak keveredése a magyar nyelvbe. Vannak, akik úgy vélekednek, hogy a nyelvet nem építik, hanem kifejezetten rombolják az idegen szavak, alkalmazásuk kizárólag hátrányt jelent (Molnár & Molnárné, 20009; Molnos, 2003). Velük szemben helyezkednek el azok, akik a szükségszerű idegen szavak használatát nem vetik meg, akik az új szavakat, kifejezéseket a nyelv természetes velejárójának tartják (Nádasdy 2003; Minya, 2003). Balázs Géza így fogalmaz: “Az idegen szavaknak sajátos stilisztikai értékük lehet. Bizonyos idegen szavak nélkül ma már meg sem tudnánk szólalni. [...] A legfontosabb, hogy ismerjük meg az adott idegen szavakat, a tudatosabb nyelvhasználók a tükörkifejezéseket, használjuk szabályosan őket.” (Balázs, 2005, 58. o.).

Tapasztalatom alapján inkább az idősebb generáció tagjai azok, akik ódzkodnak az idegen szavak használatától. A fiatalabb generáció már észre se veszi, ha nem magyar kifejezést használ (például random a véletlen helyett), ezek a kifejezések már mindennapi életünk részé váltak. A fiatalok jelentős része tanul valamilyen idegen nyelvet, így nagy eséllyel tisztában vannak adott szó jelentésével. Személy szerint a már megszokott, mindennapjaink szerves részévé vált szavak magyarosítását maradinak gondolom. Nem érzek rá igényt, hogy a rövid kifejezéseket hosszabb és bonyolultabb szavakra cseréljük le, csak mert magyarosabb hangzásúak lennének úgy (pl.: pizzéria helyett "olaszlepény-sütöde", mobiltelefon helyett "maroktávbeszélő", marketing helyett "hírvivő osztag"). Nem az mellett szeretnék érvelni, hogy cseréljük le minél több szavunkat idegen kifejezésre, csupán arra szeretnék rávilágítani, hogy ezek nyelvünk építőelemei is lehetnek.

3 Natural Language Processing

Natural Language Processing, röviden NLP, azaz Természetes Nyelv Feldolgozás. Aktív terület révén több különböző definíciója is fellelhető. A kifejezés olyan számítástudományi területet takar, mely során a számítógépek inputként, vagy outputként természetes nyelvet alkalmaznak. Az inputok lehetnek szóbeliek vagy írásbeliek, bármilyen nyelvből, stílusból, műfajból származóak. Az NLP a mesterséges intelligencia és a számítógépes nyelvészet határterületén helyezkedik el. Többféle módszer és technika közül választhatunk egy adott típusú nyelvelemzés elvégzésekor (Liddy, 2001). Bemeneti oldalról tekintve nem kis feladatot ró a gépekre. Gondoljunk csak a beszélt nyelv sokoldalúságára, a hordozott jelentés többszintűségére. Adott kontextusból kiragadott részletek megértése sokszor az egyének számára is nehézkes, adott témához illeszkedő háttértudást igénylő. Az NLP abban van a gépek segítségére, hogy azok ne csak megértsék az egyes szavak jelentését, hanem adott szövegkörnyezeten belül is képesek legyenek azok értelmezésére. Célja, hogy a bemeneti adat kisebb részekre való bontását követően összefüggéseket állapítson meg a komponensek között. Minél nagyobb szövegállomány áll rendelkezésre, annál könnyebb dolga van, annál pontosabb eredményt kaphatunk, hiszen a gépnek egyszerűbb kapcsolatokat találnia.

Alkalmazott eszköztárának egyik fontos eleme a Gépi Tanulás (Machine Learning, röviden ML). Algoritmusokat használ, hogy segítségükkel mintázatokat tárjon fel az adatokban, melyekből ezután adatmodelleket állít elő. Három fő alkalmazott megközelítés:

1. Felügyelt (supervised) tanulás: rendelkezésre áll egy már besorolt, címkékkel vagy struktúrával ellátott adathalmaz. Az algoritmus azt a mintázatát próbálja megkeresni a mögöttes magyarázó változónak, ami segíti a fel nem címkézett adatbázisnak a besorolását. Fajtái: klasszifikáció, regressziók.
2. Felügyelet nélküli (unsupervised) tanulás: nincsen előzetes információ, ami alapján a tanulást támogatni lehetne. Fajtái: sűrűségfüggvény becslés, klaszterezés, altér keresés, faktor analízis.
3. Megerősítő tanulás: használótól kapott visszajelzések alapján tanulja meg, hogy egyes környezetben hogyan kellene viselkednie. A felügyelt tanulástól eltérően itt nincs szükség címkézett bemeneti és kimeneti párok kialakítására (Szita & Szepesvári, 2010).

3.1 Szövegbányászat

A szövegbányászat célja a jellemzően strukturálatlan vagy kis mértékben strukturált szöveges állományokból történő rejtett, nem triviális információk felderítse, hozzáadott információk kinyerése. Az adatbányászat rokonának tekinthető, annak egyik részhalmazát képzi. Azonban adatbányászat esetében, az alkalmazott módszerek nem használhatóak közvetlenül a strukturálatlan szöveges adatokra, ezért van szükség a szövegbányászat szakterületére.

Szövegbányászat során az adatok szöveges formátumúak (nem mindig azok, lehetnek egyéb formájúak is), melyek kvalitatív információnak számítanak. Ezeket kvantitatívvá úgy alakíthatjuk át, ha a dokumentumot a benne szereplő szavak és kifejezések számával és egymáshoz viszonyított nyelvi pozíciókkal jellemezzük.

	Adat	Szöveg
Elemzés tárgya	Numerikus, kategorikus	Szabad formátumú, szöveges
Adatok jellege	Strukturált	Strukturálatlan, gyengén strukturált
Adatok tárolási helye	(relációs) adatbázis	Tetszőleges dokumentumgyűjtemény
Feladat	Összefüggések feltárása, jövőbeni szituációk előrejelzése	Információkinyerés, osztályozás, csoportosítás
Módszerek	Statisztikai modellek, döntési fák, neurális hálózatok, klaszteranalízis stb.	Számítógépes nyelvi eszközök, felügyelt és felügyelet nélküli gépi tanulók, szótárak

1. táblázat: Az adat- és szövegbányászat összehasonlítása
Az eredeti táblázat megtalálható Tikk (2007, old.: 21.) könyvében.

Sokféle problémára nyújthat megoldást különböző területeken:

- Ügyfélszolgálati tevékenység: “hangbányászat”, routing - email és hang, chatbot, virtuális asszisztens.
- Biztonság, bűnüldözés: entitásfelismerés, kapcsolatok azonosítása.
- Üzleti intelligencia és információszerzés.
- Egészségügy, gyógyszerkutatás.
- Web, szentiment elemzés.

Általános modellje öt lépésben írható fel: adatgyűjtés, előfeldolgozás, szövegbányászati eljárások, értékelés és tanulás. Az előkészítés egységesítési, formalizálási és normalizációs lépéseket tartalmazhat. Ebbe a munkafolyamatba soroljuk a felbontást (strukturális szegmentálás, mondatokra bontás és tokenizálás), a szótövezést (stemmelés, lemmatizálás), szófaj meghatározást (POS tagging) és a stopszó szűrést. A szöveganalitikai eszközök két további nagyobb csoportra oszthatóak – elemzés, feldolgozás és vizualizáció (Tikk, 2007).

3.1.1 Klasszifikáció

Az alábbi fejezet megíráskor az Adatelemzés nevű órámon elhangzottakra, illetve Németh, Katona és Kmetty 2019-es írására támaszkodtam.

Osztályozó módszer választásakor érdemes figyelembe venni annak gyorsaság, robusztusság, skálázhatóság, értelmezhetőség, skála-invariancia és pontosság tulajdonságait. A módszerek három fő lépésből tevődnek össze:

1. Modellkészítés

Meghatározzuk az osztálycímkéket. Ezek a megkülönböztetés alapjai lesznek, melyek véges számú különböző értéket vehetnek fel (annyifélét, ahány osztály van). Ezek értékeinek megfelelően próbáljuk meg elválasztani a mintákat. Feltételezzük, hogy a bemeneti ismérvek ezektől függnek (ha nem így lenne, nem tudnánk előre jelezni az új minták ismeretlen címkéjének értékét).

2. Modellellenőrzés

Az elkészített modell pontosságát mérjük. A tesztminták segítségével vizsgálható, hogy a modell az egyes mintákat jó osztályba sorolta-e, valamint, hogy milyen a minták helyes osztályba sorolásának pontossága. Iteratív folyamat, addig javítjuk a tanulás és tesztelés során a modellünket, amíg az az alkalmazáshoz mérten elfogadható pontosságú nem lesz.

3. Modell felhasználása

Amennyiben a téves osztályba sorolások száma nem túl nagy, úgy tekintjük, hogy az algoritmus által adott függvény a továbbiakban is használható az adott csoportok elkülönítésére, tehát sikerül elfogadható pontosságú modellt alkotni. Ekkor alkalmazhatjuk előrejelzésre. A mintaelemszám növelésével a függvény javulhat, de n → végtelen esetén sem várható tökéletes osztályba sorolás

Validálásra a túlillesztés (overfitting) probléma miatt is szükség van. Előfordulhat, hogy túlságosan hozzáillesztettük a tanító halmazhoz a modellünk becsült paramétereit. A modell már nem a populációról fog szólni, hanem a minta minden egyes kis sztochasztikus különbségéhez illesztünk egy kontrasztot, ami miatt csökken a modell predikciós ereje. Egy külső validációs adathalmazon tudjuk megítélni, hogy jelen van-e a túlillesztés, emellett az algoritmus tényleges hatékonyságát is. A tanulóhalmaz nagyságának minél optimálisabb meghatározása emiatt is nagyon fontos, ha mérete túlontúl nagy, akkor nagyobb eséllyel merülhet fel túlillesztés. A tanuló- és teszthalmaz felosztása az adatok időbeli eloszlásától és minőségétől is függ, de a leggyakrabban alkalmazott a ⅔ és ⅓ arányban történő szétosztás.

3.1.2 Felügyelt tanulás

Szövegelemzés során gyakori, hogy azt szeretnénk eldönteni egyes szöveget milyen kategóriába tudunk besorolni. Az osztályozás nem csak azért a legfontosabb formája a szövegelemzésnek mert képes a jelenségek leírására, hanem azért is, mert előrejelzésre is használható.

Osztályozási feladat során egy olyan 'f' célfüggvény megtanulása a célunk, amely attribútumértékek minden egyes 'x' halmazához előre definiált osztálycímkék valamelyikét (Y) rendeli hozzá. A program feladata lesz a megfelelő csoport kiválasztása, mely kategóriák a folyamat során állandóak, és előre adottak. Ez különbözteti meg a felügyelet nélküli tanulási technikáktól - például csoportosítás -, ahol a közös ismérvek nem ismertek előre. Klasszifikáció tekintetében beszélhetünk "puha" és "kemény" osztályozásról. Előbbi alatt azt értjük, mikor bináris döntést akarunk hozni, tehát azt akarjuk eldönteni, hogy egy adott halmazba tartozik-e a vizsgált szöveg/szövegrészlet. Utóbbi ezzel szemben skálaszerű értelemmel bír - egy skálán mérhetjük, hogy egy adott szöveg/szövegrészlet milyen valószínűséggel sorolható adott csoportba.

Az ismert osztályokba rendezésnek két fő módszerét különböztethetjük meg. Szótáralapú megoldás esetében egy előre rendelkezésünkre álló szószedetből indulunk ki, a szövegeket/szövegrészleteket a szószedet elemeinek felbukkanása alapján szortírozzuk. Szótáralapú osztályozás tekintetében beszélhetünk kétkategóriás és többkategóriás besorolásról. Viszonylag egyszerű, de nagyon időigényes feladat. Az összes lehetséges szó, szófordulat és kifejezés összegyűjtése nehézségeket okozhat, főleg amiatt, hogy figyelni kell arra is, hogy ezek ne rendelkezzenek olyan jelentéssel, ami akár több csoportra is illik. Emellett azzal is számolni kell, hogy a program nem képes átvitt értelemben gondolkodni, sem arra, hogy megkülönböztesse az azonos alakú szavakat. Nem ismeri fel a metaforákat, iróniát, sem a gúnyt, hiszen nem képes a szövegkörnyezet értelmezésére. Ezek mellett korlátja még a nehézkes validálás (Molnár, 2016).

A második módot, a felügyelt tanulási módszerek jelentik, amelyek valószínűségalapú besoroláson alapszanak. Ekkor nincs szükség szótár építésre, mert a korpusz részleteinek szavait használja fel működéséhez, és a modell validálása is könnyebb a statisztikai módszerekkel, mint szótáralapú kategorizálás esetében. A felügyelt tanulási megoldások statisztikai modellezésnek

tekinthetőek, ugyanis a kategóriába tartozást az egyes elemek esetében, a megfigyelési egységek bizonyos jellemzői (dokumentumokban szereplő szavak számai és ezek pozíció) felhasználásával akarják megmagyarázni. A metódusok során a megfigyelések felől haladunk az általánosítások felé, hogy elérjük célunkat: az új adatokat be tudjuk illeszteni a már meglévő egységrendszerbe (Kubik, 2016). A felügyelt tanulási módszer nem teljesen gépi folyamat. A procedúrát három részre bonthatjuk, ebből az első feladat a tanító halmaz létrehozása. A véletlen mintavétellel kiválasztott szövegminta konstruálása kézi kódolást és kellő szakmai hozzáértést igényel. Fontos, hogy a halmaz reprezentatív legyen a többi adatunkra nézve, mert a taníttatás akkor lesz a leghatékonyabb, ha hasonló adatokkal dolgozunk. Második körben tanítást hajtunk végre. "Megmutatjuk" a számítógépnek, hogy melyik csoportnak mik a jellegzetes elemei, milyen szavak és kifejezések. A tanuló halmazon kívül létrehozunk egy teszthalmazt is, mely a kapott eredmények jóságának megállapítására hivatott (accuracy, precision, f1-score, recall).

Felügyelt tanulás esetén négy alapkérdésre kell tudnunk válaszolni:

- Tanuló fázisban (training):
    1) Milyen magyarázó változókat használunk?
    2) Mi a klasszifikációs modellünk?
    3) Mik a modell paraméterek?
- Következtetési fázisban (validation):
    4) Mi a várható hatékonysága a modellnek?

4 Korábbi kutatásokból származó eredmények

Azzal, hogy miben különbözik a nemek közötti nyelvhasználat, a kutatók már jóval azelőtt elkezdtek foglalkozni, minthogy a közösségi médiák platformjai megjelentek volna. Később, miután az interneten egyre inkább elszaporodtak a különböző blogok, megjelentek a közösségi média platformjai, permanensen nőtt az igény és az érdeklődés az előbb említett felületekről származó hatalmas adatmennyiség felhasználhatóságának vizsgálatára a témában. A leggyakoribbak a különféle blogbejegyzésekből kiinduló kutatások (Argamon et al., 2003; Schler et al., 2006; Zhang & Zhang 2010), de elterjedtek a tweet-eket boncolgató elemzések (Fink et al., 2012; Markov et. al., 2017, Aragón & López, 2018) is. Facebook posztokra alapozott vizsgálatokkal,

sokkal ritkábban, de szintén találkozhatunk (Rangel & Rosso, 2013; Sap et al., 2014). Kiterjedt teóriák ellenére a tényleges empirikus vizsgálatoknak még nem sikerült egy koherens képet alkotni a nemek nyelvhasználatáról, melynek legfőbb oka, hogy a szakértők közt nincs egyetértés a legjobb elemzési mód kapcsán. Hozzá kell tenni azonban, hogy ez a kijelentés csak az általam vizsgált nagyrészt Nyugat-Európai nyelvek esetében igaz. A mai napig léteznek példának okáért olyan kisebb-nagyobb bennszülött törzsek, melyek esetében teljesen szétválik és ebből kifolyólag tökéletesen megkülönböztethető férfi és női nyelvhasználat.

Kutatásomban írott szövegeket elemzek, ezért a továbbiakban olyan szakirodalmak eredményeit mutatom be, melyek szintén olyan adatbázisokra építkeztek, melyek írott szövegekből állnak. A teljesség igénye nélkül csak arra szeretnék példákat mutatni, hogy az idő során milyen forrásokból, milyen módszereket alkalmazva kutattak ebben a témában.

Robin Lakoff volt az első, aki a nyelv és a társadalmi nem összefüggéseit vizsgálta. Lényegében a nyelvhasználatban észlelhető eltérésekre koncentrált. Tanulmányában a nők nyelvhasználatán és a nyelv nőkkel kapcsolatos használatán keresztül vizsgálta a nők alárendelt társadalmi pozícióját. Azt a következtetést vonta le, hogy a nők által használt nyelv számos aspektusban - szóhasználat, hanglejtés és szintaktika - különbözik a normának vett verziótól. Emellett több töltelékszót és érzelmek kifejezésére alkalmas szót használnak. Számára a női nyelv általános konvencionális udvariasságot testesít meg (Lakoff, 1973). Nemi és nyelvi elméletei azt sugallják, hogy a nők passzív nyelvet használnak (félénkség, kevesebb önbizalom jele) (Lakoff, 1975). Ezzel egyező eredményeket közölt Eckert és McConell-Ginet 2003-as tanulmányában.

Newman és munkatársai 2008-as tanulmányukban összegyűjtötték, hogy milyen, gyakran ellentmondásos eredmények születtek a korábbi kutatások során. Kihangsúlyozták, hogy a szavakra alapozott szöveganalitikák természetüknél fogva nem képesek megragadni azt a kontextust, amelyben a szavakat használják. A nemek közötti különbségek interpretálása árnyalt ügy, a társadalmi célok, szituációs igények és szocializáció komplex kombinációja.

A tanulmányban saját kutatásukat is ismertetik. Elemzéseikben kicsi, de szisztematikus különbségeket találtak női és férfi nyelvhasználat között. Ezt négy aspektusból vizsgálják: mikor, hol, miért és hogyan választható szét a nyelvhasznált mind abban a tekintetben, hogy mit, mind abban, hogy hogyan mondják. Esetükben a hangsúly funkcionális szavakon volt. Adataik

támogatják és igazolják, sem, mint ellentmondanak a korábbi kutatások eredményeinek. Azt sugallva, hogy a szó-számolási stratégiák életképes, nagyhatékonyságú alternatívái az emberi kódokon alapuló nyelvi elemzésnek.

Az alábbi táblázat munkájuk alapján a szakirodalmak kontrasztosságát hivatott szemléltetni.

Eredmények	**Irodalmak**
Nincs különbség a férfiak és nők nyelvhasználata között.	Bradley, 1981; Weatherall, 2002
Van különbség a férfiak és nők nyelvhasználata között.	Brownlow, Rosamon, & Parker, 2003 Colley et al., 2004
A nők többet kérdeznek, és átlagosan hosszabb mondatokat írnak. Kerülik a feszültséget. Nagyobb valószínűséggel köszönik meg a dolgokat, és sűrűbben kérnek bocsánatot, valamint zavarja őket az udvariassági formák megsértése. A nők intenzívebb mellékmondatokat, több kötőszót, személyes névmást és modális segédigét használnak. Több pozitív érzelmet mutatnak.	Mulac,Weimann,Widenmann, & Gibson, 1988 Mulac & Lundell, 1994 Savicki, 1996 Biber, Conrad, & Reppen, 1998 Thomas & Murachver, 2001 Mulac et al., 2001
A férfiak inkább véleményeket fogalmaznak. Több szót és “fordulatot” alkalmaznak. Személytelen, tényorientált nyelvezetet használnak és kevésbé foglalkoztatja őket az udvariasság, ám magabiztosabbak. A férfiak többet esküdöznek, hivatkoznak helyszínekre. Többször hivatkoznak haraggal kapcsolatos érzelmekre.	Mulac & Lundell, 1986 Dovidio, Brown, Heltman, Ellyson & Keating, 1988 Savicki, 1996 Herring, 2000 Mehl & Pennebaker, 2003

2. táblázat: Megelőző kutatási eredmények összefoglalása
Newman et al., 2008

Huffaker és Calvert 2005-ös írásukban tinédzserek webblogjain keresztül vizsgálták, hogy hogyan fejezik ki a serdülők magukat nyelvileg, milyen érzelmi kódokat használnak. Egyes platformokkal ellentétben a blogok nem követelik meg, hogy a felhasználók névvel azonosítsák magukat. A szerzők teljesen azonosítatlanul, névtelenül posztolhatnak. Ennek ellenére azt tapasztalták, hogy a blogok szerzői nem élnek ezzel a lehetőséggel. Az írópáros úgy vélte, hogy a személyes adatok, mint a név, életkor és egyéb privát információk megosztása mögött az az indok lapulhat, hogy így az egyes írások jobban tükrözik az ént, azt, ahogy a blogírók láttatni akarják saját magukat másokkal. Tanulmányukban elemezték az emoticonok használatát is. Arra a megállapításra jutottak, hogy ezek alkalmazása jobb benyomást alkot a szerzőről. A két nem kommunikációs szokásai gyakran eltérőek. A férfiak direkt és erőteljes stílust, míg a nők

indirektebb, intimebb interakciós kifejezésmódot alkalmaznak. Ezek párhuzamba hozhatók azokkal a hagyományos nemi szerepekkel, miszerint a férfiak az anyagiasságot (cselekvés, önfejlesztés, egyéniség), a nők pedig a törődést (érzelmi kifejezőkészség, mások szükségleteire való összpontosítás) testesítik meg. A nők online bejegyzéseik esetében olyan nyelvi stílust alkalmaznak, ami azt sugallja, hogy félénkek, kisebb önbizalmúak. Például vonakodnak nyíltan visszautasítani valamit, elkötelezni magukat valami mellett: "Ó, sajnálom, akkor épp az orvoshoz van időpontom." (Eckert & McConell-Ginet, 2003). Huffaker és szerzőtársa egy tartalomelemző szoftvercsomag segítségével értékelte ki a dokumentumokat a szavak száma, tartalomtípus és nyelvi hangnem szempontjából. A használt nyelv értékelését is a tartalomelemző szoftverre bízták, amely figyelembe vette a nyelvi kontextust, valamint a szavak gyakoriságát. A férfi és női blogok összehasonlítására khi-négyzet próbákat és független t-próbákat futtattak.

Lee eredményei alapján a nők több grafikus emoticont használtak érzelmeik kifejezésére, ami egybevág azzal a feltételezéssel, hogy a serdülőkorú férfiak hajlamosabbak érzelmeik letagadására. Azt tapasztalta, hogy férfi-férfi kommunikáció esetén nagyon ritka az emoticonok használata, férfi-nő kommunikáció esetében azonban már gyakoribb. A nők egyenlő arányban használják őket beszélgetéseik során, akár férfival, akár nővel csevegnek (Lee, 2003). Ezzel ellentétben a szerzőpáros nem talált nemi különbségeket a hangulatjelek használatának gyakoriságában. Sőt meglepő módon azt tapasztalták, hogy azok közt, akik használnak emoticonokat több a férfi. A nemi szerepeknek megfelelően ők is azt tapasztalták, hogy a férfiak nagyobb önbizalommal rendelkeztek. Nem tapasztaltak azonban több agressziót férfiak esetében, vagy nagyobb passzivitást a hölgyeknél. A korábbi eredményekkel ellentétben nem találtak a nekem között különbséget törődés, illetve együttműködés terén. A kutatásuk alapjául szolgáló blogbejegyzésekben Savickihez és Herringhez hasonlóan arra az eredményre jutottak, hogy a férfiak aktívabb, határozottabb és rugalmatlanabb nyelvet használtak. A nők ellenben nem használtak passzívabb, kooperatívabb, vagy alkalmazkodóbb nyelvet. E mögött meghúzódó lehetséges ok szerintük, hogy a nyelv és az interneten történő szociális interakciók változnak, talán pont amiatt, mivel a résztvevők is változnak. Az internetes nyelvet a hagyományos nyelvvel összevető tanulmányok gyakran túlegyszerűsítettek, abból kifolyólag, hogy az online interakciók nagyobb szabadságot és flexibilitást kínálnak. Az általuk vizsgált női bloggerek más nemi

szerepekkel rendelkezhetnek, mint az előttük vizsgált generációk, például az, amelyiket Lakoff vizsgált. Végeredményben, az általuk elvártakkal is ellentétben, azt tapasztalták, hogy a férfiak és nők nyelvhasználata inkább hasonló, sem mint különböző (Huffaker & Calvert 2005).

Rangel és Rosso (2013) spanyol anyanyelvű emberek nyelvhasználatát vizsgálta, hogy milyen grammatikai kategóriákat használnak Facebook posztjaikban. Azokra a kognitív tulajdonságokra összpontosítottak, amik nem és kor szerint különbözővé tesznek minket. Különböző csatornákon (Wikipedia, hírek, blogok, fórumok, Twitter és Facebook) vizsgálták a grammatikai kategóriák használatának alakulását. Mivel számomra a Facebook eredmények relevánsak, ezért továbbiakban csak az erre a csatornára vonatkozó eredményeiket foglalom össze. Tapasztalataik szerint legnagyobb arányban itt főneveket, igéket és mellékneveket használnak. Módbeli segédigékkel nem találkoztak és az indulatszavak aránya is igen alacsony volt. Nemekre lebontva eredményeik alapján a férfiak több előszót használnak, talán amiatt, hogy megpróbálják a környezetükben lévő dolgokat hierarchikusan besorolni. A nők ezzel szemben több determinánst, névmást és közbeiktatást alkalmaznak, valószínűleg amiatt, mert több érdeklődéssel vannak a társas kapcsolatok iránt. Adatkészletük nagyszámú névtelen szerző posztjából állt, ahol a posztok mellett szerepelt az író neme és kora. Adataik nem szerint igen, de kor szerint nem voltak kiegyensúlyozottak. SVM (Support Vector Machine) osztályozási eljárást alkalmaztak. Végeredményben pedig arra a megállapításra jutottak, hogy az általuk alkalmazott stilisztikai feature-ök jobban teljesítenek a kor, mint a nem azonosítására, esetleg annak köszönhetően, hogy az írásstílust inkább a szerző korától és nem a nemétől függ. A nem meghatározását elég nehézkes feladatnak találták. (Range & Rosso 2013).

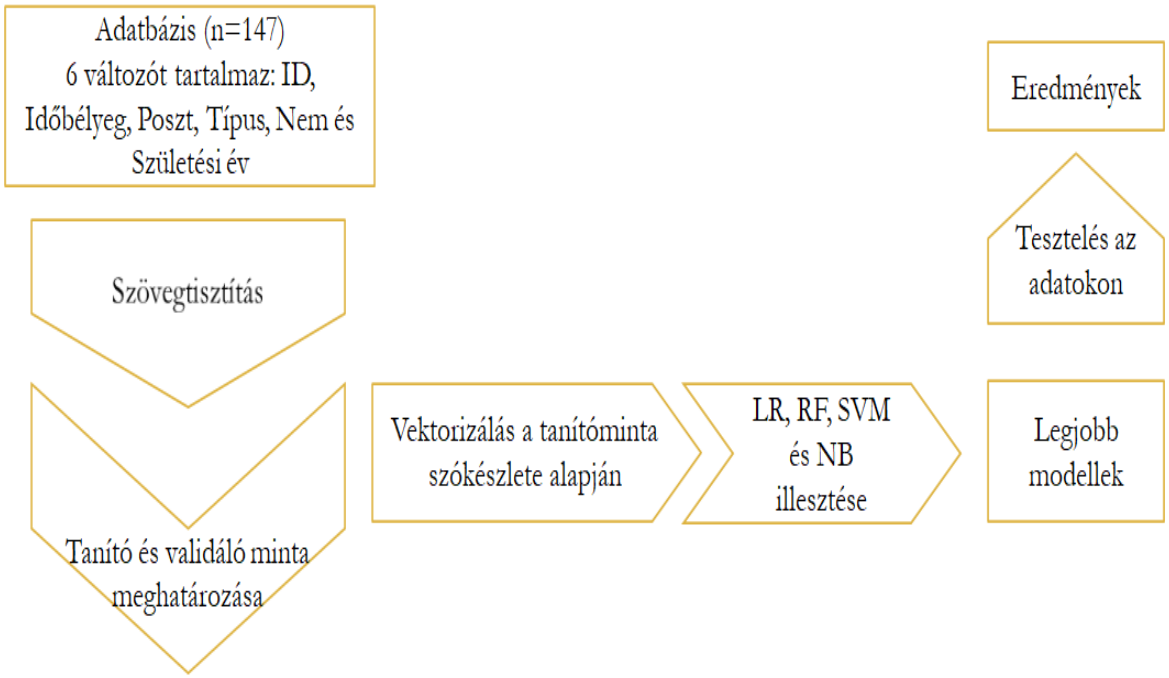
Sap és munkatársai (2014) abból indultak ki, hogy a férfiak és nők jelentősen különböznek érdeklődési és munka béli preferenciáikban. Az életkor előrehaladtával a személyiség fokozatosan változik. Emellett a közösségi média nyelve a kor és nem függvényében változik. Olyan Facebook posztokat elemeztek, melyek esetében az érintettek beleegyeztek állapotfrissítéseik megosztásába, emellett nemük és koruk feljegyzésbe is. Egy súlyozott szótárt alkalmaznak, amelyet a lineáris többváltozós regressziós és osztályozási modellek együtthatóinak felhasználásával hoztak létre. A nyelv kollinearitás miatt többváltozós lexikon fejlesztési megközelítést alkalmaztak, mely figyelembe veszi a kovarianciát. A nem előrejelzéséhez SVM

osztályozó eljárást használtak. A tanító adathalmaz felhasználásával több modellt is teszteltek a legjobb modell megtalálásának érdekében. Mind a kor, mind a nem esetében az általuk elért pontosság lényegesen magasabb, mint az alapérték (életkor esetében 23 év, nem esetében nő). Azonban azt tapasztalták, hogy minél kevesebb bejegyzés áll rendelkezésre egy-egy felhasználótól, annál kevésbé pontos a nem és az életkor előrejelzése. (Sap et al., 2014).

Garimella és Mihalcea (2016) a tipikus felületi szintű szövegosztályozási megközelítéseken túlmenően (szó előfordulások és szavak száma) szemantikai és pszicholingvisztikai szóosztályokat határoznak meg, amelyek tükrözik a férfiak és nők közötti szisztematikus különbségeket. Webblogokból álló nagy adathalmazon egy szóértelem elkülönítő keretrendszert használnak arra, hogy azonosítsák a nemeknek megfelelő szó előfordulását, azonosítsák a kiemelkedő szóosztályokat. Céljuk annak bemutatása, hogy a nemek alapján történő szétválasztás lehetséges, és hogy valóban vannak különbségek abban, ahogy bizonyos szavakat a férfiak és ahogy a nők használják. Korpuszalapú modellezés és szemantikai elemzésük eredményeként azt tapasztalták, hogy a két nem különbözik magán és nyilvános beszéd terén. Emellett Naív Bayes osztályozó módszerrel különbségeket kerestek a nemek között, abban ahogy ugyanazt a fogalmat használják. Az általuk megfigyelt nemi különbségeket nem különböző szójelentések használata okozta, sokkal inkább annak tudható be, hogy a férfiak és nők egyes szavakat különböző módon használnak, nagymértékben eltér a körülöttük lévő világról alkotott felfogásuk (Garimella & Mihalcea, 2016).

5 Kutatás

Az általam végzett elemzés célja választ kapni azokra a kérdésekre, hogy el lehet-e különíteni a nemeket nyelvhasználatuk alapján, és ha igen, akkor hogyan, milyen mértékben lehet megkülönböztetni őket? Célkitűzéseim közé tartozik még, hogy feltérképezzem, hogy a nők és férfiak hogyan használják a magyar nyelvet a Facebookon. Milyen szófajú szavakat, ezen kívül milyen emojikat, emoticonokat használnak leggyakrabban? Van-e különbség a nemek közt ezek használatában? Többféle klasszifikációs modell futtatásának segítségével szeretnék választ kapni arra, hogy melyik felügyelt tanulási algoritmussal lehet leghatékonyabban helyesen besorolni a saját nemüknek megfelelő osztályba a férfiakat és a nőket.



1. ábra: Az elemzés leegyszerűsített folyamatábrája

5.1 Adatok bemutatása

A kutatásomhoz nagy időintervallumra kiterjedő Facebook adatokat használtam fel.

A 2018-as Cambridge Analytica botrány előtt a legtöbb Facebook-ra alapozott kutatás a platform által kínált API-kat (Application Programming Interface, Alkalmazásprogramozási felület) használta fel. A közösségi média vállalatok adatok mintavételi eljárásainak generálására használt algoritmusai azonban eléggé átláthatatlanok. Nem lehet tudni, hogy az API-k révén gyűjtött adatok reprezentatív mintának tekinthetőek-e (Németh & Koltai, 2021). A botrány után ezek hozzáférhetőségének csökkentését követően, az API-alapú kutatás szinte lehetetlenné vált. Ennek a problémának egyik lehetséges megoldásaként olyan adatszerzési módszert kellett találni, ahol a felhasználók saját maguk osztják meg a Facebook-ról exportált személyes adataik egy részét a kutatókkal.

Egy Magyarországon zajló 2019-es kutatás során a résztvevőket egy önkitöltős kérdőív kitöltése után megkérték, hogy töltsék le a kitöltésre használt számítógépre Facebook adataikat. A minta

nem valószínűségi[4] kvótás[5] (nem és kor) kényelmi[6] minta volt. Ez korlátozza a kutatás megállapításainak általánosíthatóságát, ugyanis a minta fiatalabb volt, mint a magyarországi átlagos Facebook-felhasználók, valamint nagyobb volt a női résztvevők aránya (75%).

Egy fontos kritériuma volt a mintába kerülésnek, hogy adott résztvevőnek rendszeres Facebook felhasználónak kellett lennie. A rendszerességet ebben az esetben úgy definiálták, hogy legalább heti gyakorisággal használja a platformot. Az adatok a résztvevők Facebook használatának teljes időtartamát lefedik, azaz a regisztrációjuktól kezdve a letöltés pillanatáig a felületen végrehajtott összes tevékenységüket, beleértve azokat is, amiket később eltávolítottak. Néhány esetben ez több mint tíz évnyi adatot jelent. A letöltött adatprofilok nem tartalmaznak privát üzeneteket, sem keresési előzményeket vagy audiovizuális tartalmakat (fotók és videók). Valamint kizárták az olyan tevékenységeket - mint például a Marketplace - melyeket a felhasználók ritkán, vagy szinte soha nem használtak (Breuer et al., 2021).

Milyen tevékenységeket kaptak meg végül a kutatók? Az általam is elemzett cselekvések 20 típusba sorolhatóak. Gyakoriságukat tekintve az első három legnagyobb mértékben előforduló cselekedet típusok a posztok/bejegyzések valakinek az idővonalára, az állapotfrissítések és a csoportba írt bejegyzések voltak. De ezek mellett még sűrűn előforduló tevékenységek voltak a fotók, illetve videók feltöltése, (csak a feltöltött tartalomhoz írt szöveg, egyes esetekben a videó/kép url-je) különböző linkek megosztása, saját idővonalon bejegyzések közzététele. Ritkábban, de az adatbázisban találhatunk többek között emlékmegosztásokat, vagy akár geotag-eket (hol járt a felhasználó) is.

A mintát 110 nő és 37 férfi teszi ki. Az átlagos életkor 30 év, a legfiatalabb egyén 18 éves, a legidősebb pedig 71. Az adatbázisban 149.471 tevékenység szerepel, egy sor egy tevékenységnek felel meg. Minden poszt szövege mellett szerepel az egyoldali kulccsal anonimizált szerző. Mindegyik résztvevőt egy egyéni azonosítóval láttak el. Ebből a poszt íróját nem lehet visszakövetni, de segítségével tudjuk, hogy melyik tevékenységek származnak ugyanattól a felhasználótól. Emellett fel van tüntetve a felhasználó neme, és születési éve. Valamint szerepel

[4] Az elemek kiválasztása nem véletlenszerűen történik az alapsokaságból, a populáció tagjainak nem azonos az esélye a mintába kerülésre. (Babbie, 2003)

[5] Előre meghatározott összetételű minta a cél. Ehhez a populációt először egymást kölcsönösen kizáró alcsoportokra bontják, majd az egyes szegmensekből az arányoknak megfelelően választanak alanyokat. (Babbie, 2003)

[6] A mintát a populáció legkönnyebben elérhető tagjaiból alakítják ki. (Szokolszky, 2004)

a tevékenység típuskódja is, hogy hova lett bejegyezve a poszt (saját idővonalra, egy ismerőse idővonalára, csoportba, eseményhez). Ezek mellett a posztok időbélyeggel (timestamp) is el lettek látva, ami alapján másodpercre pontosan meg lehet mondani, hogy mikor született a bejegyzés. A posztok átlagosan 118 szóból állnak. A leghosszabb poszt 13.610 mondatalkotó elemből áll. A nők terjedelmesebb része 50 és 90 közötti szóból álló bejegyzéseket tett közzé, velük szemben a férfiak nagyobbik zöme 45 és 125 közötti lexémát használ. A leghosszabb bejegyzések valamilyen csoportba történő posztoláskor születtek.

5.2 Adatelőkészítési eljárások

Ha a nyers szövegkorpusz már rendelkezésünkre áll, a következő lépés a szövegekből kvantitatív adat, azaz egy elemzésre alkalmas numerikus adatbázis előállítása, melyekre könnyen integrálhatóak a különböző algoritmusok, és az adatbázist a problémának megfelelően reprezentálják. Az előfeldolgozás (preprocessing) több elemből is állhat, de minden eset egyedi, hogy melyeket alkalmazzák egyrészt az elemzés céljától, másrészt a korpusz műfaji sajátosságaitól is függhet. Az egymást követő lépések egymástól függőek, hogy melyik lépést hagyjuk el, illetve, hogy az alkalmazottakat milyen sorrendben hajtjuk végre, nagyban hatást gyakorolnak a kutatás eredményére (Németh, Katona, Kmetty, 2019).

Lehetséges feladatok:

- Tokenizálás (Tokenization)
- Szótövesítés (Stemming, Lemmatization)
- Stopszó szűrés (Stop word removal)
- Szófajok és más nyelvészeti kategóriák meghatározása (Part of Speech tagging)
- Tulajdonnevek vagy más névelemek felismerése (Named entity recognition)

A következő fejezetekben ismertetett adatelőkészítési eljárásokat és modellfuttatásokat a Python nevű, nyílt forráskodú programozási nyelv és egy-egy kifejezetten szövegelemzést segítő csomagjának segítségével hajtottam végre. Előfordult, hogy egyes előfeldolgozási módszereket (mint speciális karakterek eltávolítása, vagy duplikáció szűrés) az előrehaladás során ismételten elvégeztem, melyekre az adatok kissé mélyebb tisztítása miatt volt szükséges.

Tokenizálás előtt megtisztítottam a korpuszt minden olyan tartalmi elemtől, melyeknek a szöveg mondanivalójára nézve nincs hozzáadott értéke, nem erősítik a műveletek eredményességét. Legelőször is a teljes szövegben lecseréltem a nagybetűket kisbetűs megfelelőjükre. Ez elsősorban az emoticonokkal való munkafolyamatomat könnyítette meg jelentős részben. Így ugyanis az írásban használt betűkarakterekнél figyelmen kívül hagyhattam, hogy egyes emoticonokat valaki kis-, míg mások nagybetűs formájában használnak.

Az egyes típusú tevékenységek - poszt/bejegyzés valakinek az idővonalára - elhanyagolható részt tekintve csupán születésnapi és névnapi köszöntések különböző formáit, valamint egyéb keresztény ünnepekkel kapcsolatos jókívánságokat fedtek le. Mivel úgy éreztem a számtalan féle, gyakran ismétlődő gratulációk magához a klasszifikáció folyamatához nem tesznek hozzá semmilyen plusz értéket, ezért következő lépésként ezt a típust eltávolítottam az adatbázisból 61.142 sorosra csökkentve ezáltal a korpuszomat. Nem lettek azonban teljesen “száműzve” az elemzésből. Érdekesnek találtam annak szemügyre vételét, hogy vajon melyik születésnapi köszöntési forma elterjedtebb a nőknél és melyik a férfiaknál, így a későbbiek során ezeket egy külön elemzésben felhasználtam.

Ezt követően eltávolítottam minden linket, és e-mail címet a posztokból. Abban az esetben, ha egy bejegyzés csak magát a linket/e-mail címet tartalmazta a teljes sort töröltem az adatbázisból. Ha a link/e-mail cím szöveg közé volt beágyazva, csupán csak magát a linket/e-mail címet vágtam ki a posztból. Az adatbázisom ekkor 57.479 sort számlált. Abból kifolyólag, hogy születésnapi köszöntések, többek között mint “boldog születésnapot”, “boldog szülinapot”, “happy birthday”, “isten éltessen sokáig”, valamint egyéb jókívánságok, mint “boldog karácsonyt”, “kellemes ünnepeket”, “sikeres új évet”, “boldog húsvétot” nem csak azok között a posztok között voltak megtalálhatóak, amiket valaki másnak címeztek, hanem a mintában szereplők saját üzenőfalán is rengeteg közzé lett téve, ezért ezeket a többi benmaradt típusból is kiszűrtem, ami 3.364 sornyi csökkenést eredményezett az adathalmazon.

A posztokban szereplő személyek anonimizálásakor a neveket egy ‘@’-jelet követő szám és betűkombinációval helyettesítették. Mivel plusz információval ezek az egységek sem szolgáltak, viszont később az esetlegesen feleslegesen bent maradt speciális karakterek problémát okozhattak volna, ezért hasonló logika mentén, mint amit az url-ek esetében is használtam, ha

csak személyjelölések szerepeltek egy posztban, akkor a teljes bejegyzést eltávolítottam, egyéb esetben csak az elmaszkolt részt vágtam ki.

5.2.1 Emoticonok és emojik

A posztokban használt hangulatjeleket két nagy csoportra lehet szétbontani. Emoticonokról abban az esetben beszélhetünk, mikor szigorúan csak írásban használt - gyakran speciális - karakterekből álló szimbólumokat alkalmazunk. Az emojik ellenben azok a szövegbe ágyazott apró képek, amelyeket nem tudunk billentyűzettel bevinni, hanem egy rendelkezésünkre álló készletből választjuk ki őket. Utóbbiaknak a platformokon megjelenő kinézete szolgáltatásonként és eszközönként eltérő lehet, de a mögöttes jelentésüket tartalmazó unicode-ok alapján beazonosíthatóak. Mind az emoticonok, mind az emojik használata információt tartalmaz(hat) a szerzőről, ezért nem távolítottam el őket. Azonban ahogy fentebb is utaltam rá, a modellek futtatásakor a speciális karakterek problémát okozhatnak, ezért mind az emoticonokat, mind az emojikat átkódoltam.

Az emoticonok esetében közel 30 különböző előfordulásról lehet beszélni. A felhasználók sok esetben egy “eredeti” emoticonnak rengeteg más kissé eltérő alakját alkalmazták. “Eredeti” emoticonoknak az ismétlődések, felesleges szóközök és plusz karakterek nélküli megfelelőjét vettem egy-egy hangulatjelnek (például: :)). Viszont mivel az egyéb előfordulások, úgy, mint például: :), :-), :--) ugyanazon hangulat helyettesítését szolgálják, ezért ugyanazt a kódnevet kapták - ebben az esetben mosolygósarc -.

Példák a leggyakoribb előfordulásokra átkódolt megfelelőjükkel:

- :) = mosolygósarc
- :d = nevetősarc
- :(= szomorúarc
- :’(= sírósarc
- :o = meglepettarc
- ^_^ = derűsarc
- *-* = izgatottarc
- :@ = mérgesfej
- <3 = szívjel

Emojikból rengeteg félét használtak a mintában szereplő felhasználók. Közel 100 különböző szimbólum lett átkódolva, de ezeken kívül rengeteg volt, amit csupán egy-egy egyén használt egyetlen egy posztjában. Emiatt, ha valamelyik emojinak 10 alatti volt az össz előfordulása, akkor nem lett átkódolva, hanem törlésre került. Hogy a későbbiekben meg tudjam különböztetni az emoticonokat és emojikat - hiszen arra is választ szerettem volna kapni, hogy melyik hangulatjelet használják gyakrabban a nők és melyeket a férfiak - az emojik kódolásakor mindegyiket egy 'u' betűvel kezdtem, utalva arra, hogy ezek unicode-ok voltak (az adatbázisban ugyanis így jelentek meg, nem képes alakjukban).

Példák a leggyakoribb előfordulásokra átkódolt megfelelőjükkel:

- <u+0001f600> = unevetősarc
- <u+0001f642> = umosolygósarc
- <u+0001f60b> = unyelvnyújtóskacsintós
- <u+0001f60a> = uszivecskésszem
- <u+0001f44d> = ulike
- <u+0001f60e> =unapszemüvegesarc
- <u+0001f457> = uruha

Általánosságban elmondható, hogy ugyan emojiból sokkal szélesebb palettát fedtek le a felhasználók, ám az emoticonok nagyobb számban fordultak elő az adatbázisban. Előbbiek átkódolása viszont kisebb kihívást jelentett, ugyanis az interneten számtalan oldal[7] segítséget nyújt abban, hogy az unicode-okat saját magunknak dekódolni tudjuk. Ellenben az emoticonok (és különösen az összetetteb emoticonok) esetében sokszor csak maga a szerző tudja, hogy miért épp azt a hangulat jelet használta, amit. Ezeknél a kódnevek kialakításakor saját tapasztalataimra hagyatkoztam, ám mivel magam sem vagyok túlontúl jártas az emoticonok alkalmazásában sokszor nehézkes volt rájönnöm, hogy egy-egy szimbólum milyen érzelmet is takarhat. Diplomamunkámnak ellenben nem is képezi részét az, hogy a szimbólumok és a szavak hangulat töltetét vizsgáljam.

[7] Általam használt oldal: http://www.unicode.org/emoji/charts/full-emoji-list.html

Előfordultak olyan unicode-ok, melyek feltehetően valamely régebbi kódolással voltak megadva, vagy esetlegesen valamilyen egyedi telefon márka saját unicode-jait képezik, ami miatt nem találtam meg a megfelelőjüket az internet segítségével sem. A dekódolás sikertelensége miatt nem tudtam megfejteni, hogy milyen emóciót helyettesítenek, illetőleg milyen funkcióval szerepelnek adott bejegyzésben, ezért ezeket eltávolítottam az adatbázisból. Komplett sorok is törlésre kerültek abban az esetben, ha tartalmilag semmit nem tettek volna hozzá az elemzéshez. Például: "¦¦¦¦¦-¦---------- <u+258c> ¦¦¦--- <u+258c> --------_¦_¦_¦ <u+2590> --- <u+258c> <u+2590> - <u+258c> - <u+feff> ---- kamion ---_¦_¦_¦_¦_¦_¦ <u+2590> ------- <u+258c> ---------- <u+258c> ¦_¦_¦ <u+2590> - (@) ----(@) (@) ---------- (@)----" ahol az unicode-ok és a speciális karakterek a kamion kerekeit, illetve ablakait helyettesítik.

A fentebb említett kétféle hangulatjel típus mellett meg kell említeni egy harmadik nem túl sok variánst magában foglaló, de relatíve gyakori verziót még. Ezekben az esetekben a képi megfelelők mögött zárójelek között egy betű szerepel. Hat darab előfordulását kódoltam át: (h) = napszemüvegesfej, (a) = angyalfej, (y) = likeujj, (n) = dislikeujj, (b) = sörjel, (l) = pirosszív.

Az emoticonok és emojik kezelése után töröltem az adatbázisból minden speciális karaktert, valamint eltávolítottam a felesleges szóközöket a bejegyzések elejéről és végéről, illetve a szavak közötti dupla/tripla/… szóközöket.

5.2.2 Tokenizálás, lemmatizálás és stopszó szűrés

A szöveg tisztítási lépéseket követően 46.514 sornyi szöveget kellett tokenizálnom. Ennek a folyamatnak a során megkülönböztetjük egymástól a szövegeket alkotó egységeket. A tokenek, melyek feladattól függően lehetnek szavak, kifejezések, az adott posztban szereplő karaktersorozatok, együttesen jelentéstani elemzési egységet szolgálnak (Naveenkumar-Kiran-Reddy-Raju, 2015). Az általam elemzett posztállomány tokenjeinek össz-száma kissé meghaladja 237.000-et. Ezek közül 72.504 egyedi tokent tudtam megkülönböztetni.

Szavak csonkolásakor két megközelítést kell megkülönböztetni: ragozás eltávolítása/szótövesítés (stemming) és lemmatizálás. Előbbi esetében az elemzéshez a szóról mindennemű toldalék egyszerű levágását követően megmaradt szótövet használjuk fel. Lemmatizáláskor, ezzel szemben, a toldalékolt szó normalizált vagy szótári alakját (lemmáját) keressük meg, így a két eljárás akár jelentős eltéréseket is produkálhat, de az sem

elképzelhetetlen, hogy azonos eredményt kapunk. A lemmatizálás mindig értelmes szóalakot állít elő (ami többelemű is lehet, pl.: kapunk - kap, kapu), ellenben a szótövezés eredményéül kapott szó csonk gyakran nem értelmes szótári alakú (Tikk, 2007).

A lemmatizálási lépés azért is ajánlott, mert így a modellünk mérete csökkenthető, standardizálható és könnyebben értelmezhetővé tehető. A Python lemmatizáláshoz felhasználható algoritmusai néhány perc alatt képesek több tízezer szavas szöveg feldolgozására. A hibaarány általában tűrhető határon belül van, de ez természetesen a nyelvtől és szövegtől is függ. Különösen igaz ez, egy olyan morfológiailag gazdag nyelvre, mint a magyar, amely nem csak toldalékolásokban és összetett szavakban bővelkedik, hanem képzett alakokban is. "A különböző nyelvi osztályozási rendszerek szerint egy főnévnek 16–24 különböző esete lehet, amit, ha a birtokos, személyes ragokkal kombinálunk, akár 1400 különböző szóalakot is kaphatunk. Melléknevek esetén ez a szám a fokozás miatt akár 2700 is lehet, míg igéknél lényegesen csak kevesebb, 59 alak lehetséges" (Tikk 2007, 48. o.).

A tokenizálást követően Orosz György spaCy fejlesztését felhasználva[8], ami kifejezetten a magyar nyelvű szövegeken való alkalmazásra lett készítve, lemmatizáltam a szövegeimet. Majd Part of Speech (Pos) tag-gel láttam el a tokenjeimet, azaz szófaj szerint azonosítottam őket. Mivel a későbbiek során meg kívántam vizsgálni, hogy mely szófaj a legelterjedtebb a Facebook posztok esetében, illetve, hogy mely a legnépszerűbb a nemekre bontva, ezért nem hagytam ki egy szófajt sem az elemzésből.

A tiltólistás szavak eltávolítására a Python Natural Language Toolkit[9] (NLTK) nevű csomagját választottam, azon okból kiindulólag, hogy már magyar nyelvű stopszólistával is rendelkezik. A módszer nem kezeli a szavak sorrendjét, csupán a gyakoriságát. Nem ismeri fel a hétköznapi nyelvhasználatot sem, így a helyesírási hibákat, elgépeléseket nem tudja kezelni, ezek problémát jelenthetnek. A sokat használt, illetve nyelvtani funkciót betöltő szavak törlése javíthatja az elemzésünk végeredményét, hiszen a leggyakrabban használt lexémák vizsgálatakor így nem az 'a', 'az' vagy épp az 'egy' szavakat fogjuk visszakapni. Esetemben azonban, mivel feltételezhetőleg információvesztéssel járna minden stopszó eltávolítása, amit az NLTK stopszó listája tartalmaz,

[8] https://github.com/oroszgy/spacy-hungarian-models
[9] https://www.nltk.org/

hiszen lehet, hogy a személyes névmások használata is olyan sajátosságot jelent, amely alapján el lehet különíteni a férfiakat és nőket nyelvhasználatuk alapján. Ezért a fentebb említett listán egy kisebb módosítást eszközöltem, és csak a kötőszókat, névelőket, határozószókat, valamint néhány olyan általam megadott szót távolítottam el, melyekről a leggyakrabban használt unigramok vizsgálatát követően kiderült, hogy a posztokban gyakran alkalmazták, de érdemi információt nem hordoznak, nem adnak többletinformációt a kutatáshoz (pl.: debrecen, delon). A tiltólistára rakott kifejezések eltávolítását követően az adatbázis üres sorokat is tartalmazott, ezek eltávolításával a végleges tevékenységeim száma, amelyekre a modelleket is futtattam 46.514 volt.

Az előkészítés folyamán tulajdonnevek vagy más névelemek felismerésével nem foglalkoztam.

5.3 Egyéb eredmények a korábbi kutatások alapján

Az előzetes eredmények vizsgálatakor viszonylag gyakran találkoztam azzal, hogy egyéb dolgokat is megvizsgálnak klasszifikálás során. Voltak olyan szerzők, akik azt vizsgálták, hogy mely szófajok gyakoriak általánosságban a Facebook posztokban, és olyanok is, akik az emoticon és emoji használatot tanulmányozták. Részben ezekből az elemzésekből motivációt merítve, részben saját érdeklődés miatt az adatbázisom esetében én is megvizsgáltam az alábbiakat:

1. Egyes szófajok milyen gyakorisággal fordulnak elő. A négy leggyakoribb esetében milyen arányban használják azokat a nők és milyen arányban a férfiak.
2. Melyek a leggyakrabban használt emoticonok és emojik. Milyen arányban használják ezeket a férfiak és milyen arányban a nők.
3. Melyek a két nem által egyaránt használt leggyakoribb szavak.
4. Hogyan alakul az „én” személyes névmás használatának gyakorisága összességében és nemekre lebontva.
5. Milyen arányban használtak a férfiak és nők születésnapi köszöntéseket.

1. A négy leggyakoribb szófaj a főnevek, melléknevek, igék és határozószók voltak. Vincze Veronika és munkatársai (2021) alapján elmondható, hogy bizonyos szövegtípusokban eltérő a szófaji arány. Eredményük alapján kétféleképp lehet csoportosítani a korpuszokat. A főnevek és melléknevek dominanciája a leíró jellegű korpuszokra jellemző. Ezek általában jogi szövegek,

újsághírek, üzleti, számítástechnikai szövegek, melyek célja az olvasó tényszerű informálása. Az igék és mellettük a határozószavak sűrű előfordulása pedig az interaktív szövegek sajátossága – mint például egy iskolai fogalmazás vagy egy irodalmi szöveg, regény - melyek esetében a szerzőnek határozott szándéka, hogy megszólítsa, illetve párbeszédet folytasson az olvasóval. Azt tapasztalták, hogy sok főnév sok melléknévvel, míg sok ige sok határozószóval jár együtt. Utóbbi azzal magyarázható, hogy a határozószavak azok, melyek kifejezik az ige adott minőségét, ahogy a főnevek mellett a melléknevek tudják megjeleníteni ugyanezt a kvalitást. A Facebook-szövegeket középen helyezték el. Ezt erősítik meg saját eredményeim is.

Nemek szerint bontva nincs jelentős különbség, közel azonos arányban használják mind a négy (és az összes többi) szófajt, bár a nők minden esetben egy-két százalékponttal ugyan, de magasabb használati arányt mutattak, mint a férfiak. Utóbbi esetben figyelembe kell venni azt is, hogy az összes token - melynek be lett azonosítva a szófaja – 66 százalék származik a hölgyektől és 34 százaléka az uraktól.

2. Az emoticon és emoji használati tendenciák alapján az adatbázisban szereplő felhasználókra az emoticonok nagyobb számú használata, ellenben az emojik változatosabb alkalmazása volt jellemző.

A leggyakoribb emoticonok a derűs hangulat kifejezését szolgálóak voltak. Ezek közül is a legnépszerűbbek a ’mosolygósarc’-ként és ’nevetősarc’-ként kódolt érzelmet kifejező ábrák voltak, melyek együttesen az emoticonokat tartalmazó posztok 30 százalékában fordultak elő. Míg utóbbit mind a két nem közel azonos arányban alkalmazta, addig előbbi használata a nők esetében szignifikánsan jelentősebb volt. E kettő mellett gyakori volt még a ’szívjel’, az ’xD’, és a ’kacsintósfej’ alkalmazása.

Emojik esetében a halmozás gyakrabban megfigyelhető esemény volt, azaz mikor egy poszton belül több különböző emojit alkalmazott adott felhasználó. A leggyakrabban használt emoji aránya meg sem közelítette a leggyakrabban használt emoticonét. De gyakran alkalmazták a ’unevetősarc’, ’uszivecskésszem’, és ’ukacsintás’ piktogramokat. Az egy bejegyzésen belüli emoji csoportosítás inkább a nők esetében volt megfigyelhető. Leginkább az egyes ruhadarabokat, körmöst, fodrászt szimbolizáló kis képek követték rendre egymást, de a jókedvet kifejező arcok, a szeretetet kifejező szívjelek halmozása is sok esetben megfigyelhető volt. A férfiak esetében

pedig inkább az olyan rájuk talán jobban jellemző emojik voltak megfigyelhetőek, mint a söröskorsó vagy az autó.

3. A férfiak és nők által leggyakrabban alkalmazott 15 szót a lentebbi táblázatban összefoglalva szemléltetem. Gyakoriság szerint rendezve az első 15 találatban szerepelt három emoticon is, azonban mivel szavakra vonatkozóan szerettem volna megnézni a gyakoriságot, ezeket kivettem a listából. Személyes névmások szintén nem szerepelnek.

Helyezés	Férfiak	Nők
1	tud	tud
2	mond	szeret
3	ma	nap
4	év	mond
5	nap	szép
6	isten	megy
7	szeret	ma
8	jön	tesz
9	úr	ember
10	ember	fog
11	megy	élet
12	apa	holnap
13	élet	ír
14	magyar	kis
15	ad	jön

3. táblázat: A nemek által leggyakrabban használt 15 kifejezés az adatbázisban

4. Több szerző, mint Mulac és munkatársai (2001) vagy Mehl & Pennebaker (2003) alapján a nők több személyes névmást használnak. Az általam elemzett adatbázisban a személyes névmások használata igen gyakori volt. Az egyes szám második és harmadik személyt kifejező ’te’ és ’ő’ szavak a 15 leggyakoribb kifejezés közt szerepeltek mind a két nem esetében. Az E/3-t kifejező személyes névmás használata férfiak esetében magasabb volt, és az ’ők’ személyes névmást is – bár nagyon elhanyagolható mértékben – de ők alkalmazták többször. Az egyes szám első személyt kifejező ’én’ szó gyakrabban szerepelt a nők által írt bejegyzésekben, ahogy a ’te’ és ’ti’ kifejezések is. A T/1-et kifejező személyes névmás esetében nem volt különbség férfiak és nők esetében.

5. Az egyes típusba tartozó tevékenységek – poszt/bejegyzés valakinek az idővonalára - 85%-a a születésnapi köszöntések valamilyen formája volt. Az ismerőseik felköszöntése ebből az alkalomból a hölgyek körében sokkal elterjedtebb tevékenység volt, a posztok több mint 4/5-e tőlük származik. A leggyakoribb kifejezés, amivel jókívánságaikat kifejezték a 'szülinap' volt, ami elé nagyon sok esetben oda kerültek a 'nagyon sok boldog' szavak. Általuk kedvelt kifejezés volt még a 'szülcsi napcsi', a férfiak ezt szinte nem is használták. Azonban az 'isten éltessen' szókapcsolat esetében pont ennek ellentetje figyelhető meg. A férfiak nagyobb arányban köszöntötték fel ismerőseiket angolul, és gyakrabban kezdték köszöntésüket csupán a 'sok boldog' szavakkal.

5.4 Modellek elméleti háttere

Annak érdekében, hogy minél pontosabb klasszifikációt kapjunk érdemes a modellek és változók kombinációjával kísérletezni. Többféle gépi tanulási modell tesztelése abban is segít, hogy kiderítsük melyik illeszkedik jobban az adatokra, melyikkel tudjuk a legmegfelelőbben megfogni a pontok és címkéjük közötti kapcsolatot. Ebből kifolyólag többféle osztályozó teljesítményét vizsgáltam kutatásom során: Support Vector Machine (SVM), Bináris Logisztikus Regresszió (LR), Random Forest (RF) és Naiv Bayes (NB). Az eljárások kiválasztásakor motivációt jelentett még az is, hogy a megelőző, hasonló témát körüljáró kutatások során ezek voltak a legnépszerűbbek és legjobban teljesítők nem csak blogbejegyzéseken, illetve tweet-eken futtatva, de Facebook posztok esetében is.

A Support Vector Machine szövegbányászati feladatok során (is) nagy népszerűségnek örvendő modell, stabil teljesítmény nyújtása miatt. Kiválasztása melletti indok még, hogy a nemek osztályozásakor - Logisztikus Regresszióval egyetemben - képes valamilyen fajta választ adni a címkéket meghatározó kategória sajátosságaira. A Random Forest pedig szintén gyakori klasszifikációs módszer, mely logikáját tekintve kissé másabb, mint a másik két (lineáris) modell, ám csakúgy, mint a Logisztikus Regresszió szintén képes valószínűségi becslésre.

A modellek futtatása során azonban a Random Forest modell semmilyen paraméter beállítással nem volt képes hatékony eredményeket hozni (erről később az Eredmények című alfejezetben részletesebben írok), így az első tervekkel ellentétben Naiv Bayes modelleket is futtattam.

Választásom azért esett erre az osztályozó algoritmusra, mert a korábbi kutatások során a többi módszer mellett ez volt még viszonylag gyakran előforduló, mikor Facebook posztokra alapozott vizsgálatot végeztek.

5.4.1 Support Vector Machine

Az SVM (magyarul sokszor tartóvektor-gépnek fordítják) algoritmus ígéretes tapasztalati eredményeket mutat sokan úgy gondolják, hogy ez számít az egyik legtöbb esetben jól teljesítő felügyelt tanulási algoritmusnak. Szinte bármilyen osztályozási problémát kezelni tud a megfelelő kernelfüggvény megválasztásával. Jól működik sokdimenziós adatokkal és elkerüli a dimenzió problémát. Azokban az esetekben hatékony, mikor a dimenziók száma nagyobb, mint a minták száma. Legjelentősebb akadálya a nagy memória igénye és a tanításához szükséges kvadratikus programozás nagy algoritmikus komplexitása. Nagy adathalmazok esetén a tanításhoz szükséges idő nagysága miatt nem teljesít túl jól, valamint a zajjal rendelkező adatkészletek is problémát jelentenek számára a target osztályok átfedése miatt. A Support Vector Machine olyan kernel gép, mely hasznosítja a statisztikus tanuláselmélet eredményeit is. Alapvetően lineáris szeparálásra képes, de kiterjeszthető nemlineáris szeparálásra és nemlineáris regressziós feladatokra is, a túlillesztést kontrollálva. Több osztályt tartalmazó változókra is használható, a lentebb írtak azonban azokra az esetekre igazak, ahol kétértékű változónk van, azaz bináris osztályozási feladattal állunk szemben. Az SVM modellek felépítése többlépcsős folyamat. A döntési határ kijelölésében nem vesz részt a minta összes eleme, a support vector-okat azok alkotják, amelyek a határt támasztják, azaz csak azok a pontok befolyásolják a hipersík paramétereit, amelyek rajta találhatóak (Ray, 2017).

A klasszifikálónak ki kell választania a hipersíkok valamelyikét a döntési határ reprezentálásához annak alapján, hogy várhatólag milyen jól teljesítenek a teszteseteken. Minden adatot pontként ábrázolunk az n dimenziós térben (n a meglévő feature-ök száma), úgy, hogy az egyes feature-ök értéke egy adott koordináta értéke. Ezután - lineárisan szeparálható adatok esetén - célunk, hogy megtaláljuk azt a hipersíkot, ami a lehető legjobban elkülöníti az egyes csoportokat. A hipersík döntési határ, amely segít az adatpontok osztályozásában. A valós számok halmazán lehet egy pont, egy sík esetén egy egyenes, többdimenziós tér esetén pedig egy sík. A döntési határ margóját szeretnénk maximalizálni, és a tanulóhalmazon mért hibát szeretnénk minimalizálni.

A hipersík egyenlete: $x^T\beta + \beta_0 = 0$

ahol x egy p dimenziós, β pedig egy p+1 dimenziós oszlopvektor.

Két döntési határ esetén mindkettő szét tudja választani a tanulóeseteket a megfelelő osztályokra, jól el tudják különíteni az adatokat. Mindegyik B_i döntési határhoz tartozik két hipersík, amelyeket b_{i1}-gyel és b_{i2}-vel jelölünk. b_{i1}-et úgy kapjuk meg, hogy addig tolunk el a döntési határtól egy vele párhuzamos hipersíkot, amíg az nem érinti a legközelebbi első csoport elemét, míg b_{i2}-t úgy kapjuk, hogy a döntési határral párhuzamos hipersíkot addig toljuk el, míg az nem érinti a legközelebbi második csoportbéli elemet. A két hipersík közötti távolságot nevezzük az osztályozó margójának. A nagyobbik margójú döntési határ lesz a tanulópéldányok maximális margójú hipersíkja, ugyanis a nagy margóval rendelkező döntési határok általánosítási hibája a legtöbb esetben megfelelőbb. A kis margós osztályozók hajlamosabbak a modell túlillesztésre (a konfidenciaintervallumokhoz hasonlóan), ami által az újonnan felbukkanó eseteket gyakran rosszul általánosítják.

Az előbbi példát folytatva, ha B_1 döntési határ hibásan osztályozza az új eseteket, ellenben B_2 helyesen, ez nem jár azzal, hogy B_2 jobb döntési határ lenne. Esetlegesen ugyanis az új esetek a tanulóadatokban lévő zajnak feleltethetőek meg. Zajos, illetve outlier értékek esetében nagy a túlillesztés esélye. Ennek kiküszöbölésére engedjük meg, hogy adott elemek a határ rossz oldalára essenek. Tanítsunk meg egy olyan döntési határt, amely tolerálja a kis hibákat a tanulóhalmazon (Tan, Steinbach, Kumar, 2006).

A tényleges Support Vector Machine lineárisan nem tagolható csoportokra vonatkozik. Az adataink p-dimenziós terének kiterjesztésében keresünk lineáris szeparáló hipersíkot. Vezessünk be egy paramétert (ξ =ξ_1, … ξ_n), amely méri egy elem távolságát a hibás irányban a döntési határtól:

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \xi_i$$

feltéve, hogy $y_i(wx_i) \geq 1 - \xi_i$ és $\xi_i \geq 0\ i = 1, \dots, N$.

Ha a hibatag értéke nagyobb mint 0, akkor a megfigyelési egység befolyásolja a margó nagyságát, és azon belül helyezkedik el, ha értéke az 1-et is meghaladja, akkor a hipersík rossz oldalán található.

A C költésgparaméterrel (felfogható a félreosztályzás költségeként is) hangolhatjuk, hogy inkább ||w||-t szeretnénk minimalizálni, vagy azt szeretnénk, hogy a hibásan besorolt elemek a lehető legközelebb legyenek a döntési határhoz (gondolhatunk rá regularizációs paraméterként). Minél nagyobb értéke minél kisebb pontot jelent a rossz oldalon. AZ SVM kernel egy olyan függvény, amely az alacsony dimenziós bemeneti teret áttranszformálja magasabb dimenziós térré, vagyis a nem szétválasztható problémát szétválaszthatóra alakítja. A skaláris szorzat általánosításaként is felfogható, mérhetővé teszi a hasonlóságot két megfigyelési egység között. A klasszikus SVM hinge veszteségfüggvényt alkalmaz, mely kicserélhető - ha például nem csak azt szeretnénk, hogy a margóra eső elemek határozzák meg a döntési határ illesztését - negatív binomiális log-likelihood veszteségre. Az SVM veszteségfüggvénye hasonló alakot ölt, mint a Logisztikus Regresszióé. (Rakovics, 2016).

5.4.2 Logisztikus Regresszió

A Logisztikus Regressziót Kovács Erzsébet (2014) műve alapján mutatom be. Az LR egy olyan klasszifikációs eljárás, mely során előre definiált, egymást kölcsönösen kizáró csoportok egyikébe soroljuk a megfigyeléseket a magyarázó változó(k)ból kinyert információ alapján. A függő változó ebben az esetben egy diszkrét változó, ami lehet bináris, azaz kétféle kimenetelű, vagy multinomiális, azaz többféle kimenetű. Az osztályozás során logisztikus illesztő függvényt használunk.

Ezt követően csak a bináris esetet mutatnám be részleteibe menően, abból kifolyólag, hogy kutatásom során számomra ez a fajta logisztikus regresszió releváns.

Bináris Logisztikus Regresszió esetében a függő változó (Y) kétértékű, 0 és 1-es értéket vehet fel. A magyarázó változóval/változókkal annak a bekövetkezési esélyét szeretnénk prediktálni, hogy a kimenet 1 lesz. Az esély (odds) a magyarázó változó(k)tól függő feltételes valószínűségek aránya megkapható:

$$odds = \left(\frac{p}{1-p}\right) = \exp\left(b_0 + b_1 x_1 + \cdots + b_p x_p\right) = e^{\underline{bx}}$$

Az esély logaritmusa a logit, ami a magyarázó változók lineáris függvénye:

$$\ln(odds) = ln\left(\frac{p}{1-p}\right) = \text{logit(p)} = \left(b_0 + b_1 x_1 + \cdots + b_p x_p\right)$$

ahol p az esemény bekövetkezésének valószínűségét, X pedig a magyarázó változó(k)at jelöli.

A valószínűség 0 és 1 közötti értéket vehet fel, az odds 0 és végtelen közöttit, míg ennek a logaritmusa] -∞; +∞[között mozoghat. Az odds 1-nél nagyobb értéke azt jelenti, hogy az 1 bekövetkezésének valószínűsége nagyobb, mint a 0 bekövetkezésének valószínűsége. A b_i becsült paraméter az x_i (i = 1, …, k) változó egy egységnyi abszolút változásának a logitra gyakorolt parciális hatását mutatja.

Az Y eredményváltozó kategóriáinak bekövetkezési valószínűsége Maximum Likelihood becslés alkalmazásával becsülhető meg.

5.4.3 Random Forest

A Random Forest előnyei közé sorolható, hogy pontos klasszifikációra képes, nagy adatok esetén is gyorsan lefut, valamint, hogy becsléseket ad arra, melyik változók bírnak nagy jelentőséggel.

Nagy mértékben a döntési fák elvén alapul, a bagging egy speciális alesete, melynél az egyedi döntési fák kvázi korrelálatlanok. Míg a döntési fák estén minden vágást úgy választunk meg, hogy az a lehető legjobb legyen az összes lehetséges magyarázó változónk szerinti kettéosztás közül, addig Random Forest esetén ez leszűkül a magyarázó változóink egy véletlenszerűen kiválasztott halmazára minden egyes bináris bontás esetében. (Általában m ≅ √p.) A két eljárás közötti legfőbb különbség a magyarázó változókból kiválasztott változóhalmaz nagyságában rejlik. (James, Witten, Hastie, Tibshirani, 2013). Több gyenge osztályozó átlagos teljesítménye alapján klasszifikál, több kisebb méretű döntési fát is épít. Addig folytatja a fák építését, míg az előre rögzített mélységet el nem éri. Az egyes erdők hatékonysága függ a generált fák számától és minőségétől, valamint a fák közötti korrelációtól (ha nő, az eredmény romlik).

Az algoritmus a következő lépésekre bontható:

1. b = 1, ..., B:
    a. Válasszunk egy N elemű véletlen mintát a tanító halmazból bootstrapping-gal. Ez lesz a teszthalmazunk;
    b. Illesszünk egy döntési fát, Tb-t a következő szabályok szerint, amíg el nem érjük a minimális levél-elemszámot:
        I. Válasszunk m változót véletlenszerűen p magyarázó változó közül;
        II. Ezek közül keressük meg a legjobb vágást, mely mentén a legjobban szétválaszthatóak az osztályok recursive binary partitioning elv alapján:

$$\Delta i(s, M) = i(M) - P_L(i(M_L) - P_R(i(M_R);$$

III. Végezzük el a vágást az ii. pont változója mentén.

2. Rögzítsük a fákat, majd döntési szabályként klasszifikáció esetén válasszuk a többségi becslést (Hastie et al., 2009).

5.4.4 Naiv Bayes

A Naiv Bayes féle osztályozók előnye, hogy robusztusok izolált zajos pontokra, illetve az irreleváns attribútumokra. Az NB osztályozók a Bayes tételen alapuló osztályozási algoritmusok gyűjteménye, egy algoritmuscsalád, ahol mindegyik algoritmus egy közös elven alapul - minden osztályozott feature pár független egymástól. Az osztályozás lényege a Bayes tételen alapszik: P(Y|X) = P(X|Y)P(Y) / P(X), melynek segítségével megkapjuk Y bekövetkezésének valószínűségét, feltéve, ha X bekövetkezik. Az attribútumok adott Y osztálycímke melletti feltételes függetlenségének feltételezésével becsüli meg az osztályra vonatkozó feltételes valószínűséget. x_i-k feltételes valószínűségét kell megbecsülni adott Y mellett, ami nem igényel nagy tanulóhalmazt jó valószínűségi becsléshez. A feltételes függetlenségi feltevés kifejezhető:

$$P(X|Y = y) = \prod_{i=1}^{d} P(X_i|Y = y)$$

ahol minden X= {X_1, X_2, …, X_d) attribútumhalmaz d attribútumból áll.

Az NB alapvető feltevése, hogy minden feature függetlenül és egyenlően (minden tulajdonság azonos súlyt vagy fontosságot kap) járul hozzá az eredményhez, ezért nevezik naivnak. A függetlenség feltételezése valós helyzetekben általában nem helytálló, de gyakorlatban gyakran jól működő. A szükséges paraméterek becsléséhez kisméretű tanító halmaz is elegendő, valamint a kifinomultabb osztályozókhoz képest sokkal gyorsabban képesek eredményt adni, különösen hasznos nagyon nagy adathalmazok esetén. Minden eloszlás függetlenül becsülhető egydimenziós eloszlásként, ami segít csökkenteni a dimenzionalitásból fakadó problémákat (Tan, Steinbach, Kumar, 2006).

Gauss Naiv Bayes esetén feltételezzük, hogy az egyes feature-ökhöz tartozó folyamatos értékek egy Gauss-eloszlás (normál eloszlásnak is nevezik) szerint oszlanak el, abból kifolyólag, hogy a prediktorok folytonos értéket vesznek fel, és nem diszkrétek.

Ekkor a feltételes valószínűség képlete az alábbira módosul:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

ahol σ_y és μ_y paraméterek maximum likelihood becsléssel történik.
Bernoulli Naiv Bayes során az osztályváltozó megjóslásához használt paraméterek bináris változók. Kifejezetten bünteti az olyan feature elő nem fordulását, amely egy osztály indikátora. A Bernoulli Naiv Bayes-re vonatkozó döntési szabály alapja:

$$P(x_i|y) = P(i|y)x_i + (1 - P(i|y))(1 - x_i)$$

Complement Naiv Bayes különösen alkalmas kiegyensúlyozatlan adathalmazok esetében. Az egyes osztályok kiegészítéséből származó statisztikákat használja a modell súlyok kiszámításához. Az osztályozási szabály:

$$\hat{c} = \arg\min_c \sum_i t_i w_{ci}$$

Egy dokumentumot ahhoz az osztályhoz rendel, amelynek a legszegényebb a bővítmény egyezése (Zhang, 2004).

5.5 Funkció tervezés, vektorizálás

Számunkra egy mondat megértése (általában) nem okoz problémát, megértjük a szavakat mivel ismerjük annak és a mondatnak a szemantikáját. A számítógépeknek azonban nincs ilyen egyszerű feladata. Mit tehetünk ennek kiküszöbölésére? Mivel a gép bármilyen adatról is legyen szó, azt csak számérték formájában képes megérteni, ezért vektorizáljuk a teljes szövegünket. Vektorizálás során célunk a korpuszt alkotó dokumentumok elhelyezése a magyarázó változók által kifeszített vektortérben, mely változók többnyire a szöveg valamilyen kisebb egységei, tokenjei. Jellemző azonban, hogy a korpuszban előforduló tokenek számottevő része csak egy-egy dokumentumban fordul elő. Emiatt érdemes valahogy a magyarázóváltozók halmazát lecsökkenteni, hogy redukáljuk a vektortér dimenzióinak számát, alapesetben ugyanis egy magyarázó változó a vektortér egy dimenziójának feleltethető meg (Tikk, 2007).

Vektorizálás során a tanító halmazból kinyert n-grammok határozzák meg a magyarázó változókat, a későbbiekben teszteléshez és validáláshoz is ezeket kell kinyerni a korpuszból. Emiatt még vektorizálás előtt felosztottam az adathalmazom tanító és teszthalmazra. A 46.514

bejegyzésből álló adatbázisom véletlenszerűen ⅔ és ⅓ arányban osztottam fel. A modellek futtatásakor azonban ez túl soknak bizonyult egyes algoritmusok esetében (kifejezetten SVM, de a logisztikus regresszió is lassan futott le), ezért, hogy ne ennyi emberről (egy sor egy egyénnek feleltethető meg, ha nem vesszük figyelembe, hogy egy id-hoz hány darab tevékenység tartozik) kelljen eldöntenie a modelleknek, hogy női vagy férfi szerzőről van szó, a tevékenységeket összevontam. Későbbiekben az adatbázis egy sorában az összes, adott azonosítóval rendelkező személy által írt poszt egyben, egy cellában szerepel.

A teszthalmaz elkülönítése azért is szükséges, mert későbbiekben ezen tudjuk mérni a modellek teljesítményét még nem látott adatok jóslásakor. A tanító halmaz pedig az algoritmus szempontjából nagy jelentőségű, megválasztásánál törekednünk kell a reprezentativitásra, hogy a halmaz a leghatékonyabb taníttatás érdekében minél jobban hasonlítson az összes többi adatunkra. Általánosságban a tanító halmaz arányának növelésével javítható a modell hatékonysága. Ha túl kis arányú tanító halmazunk van, nehezebben kaphatunk magas találati arányt, ha azonban túlontúl nagy, akkor fellép a túlillesztés veszélye.

A bejegyzések összevonását követően az adatbázisom már csak 147 sort számlált. Abból kiindulólag, hogy így nem volt túl számottevő a megfigyeléseim száma növeltem a tanító halmazom méretét. A módszerek teljesítményében ekkor azonban romlást tapasztaltam, emiatt visszatértem a ⅔ - ⅓ arányú felbontáshoz. Így végül a tanító halmazomba 98, míg a teszthalmazomba 49 egyén került.

A Feature Engineering (funkció tervezés) a gépi tanulási modellek inputjaként viselkedő adatok feature-ökké való alakításának folyamata. Mikor szöveg formátumú adatokkal foglalkozunk, számos mód létezik az adatokat reprezentáló feature-ök megszerzésére. Az alábbiakban ezek közül a leggyakoribbakat mutatom be Zafra (2019) cikke alapján. A később futtatott modellek a TF-IDF vektorizálást követően produkálták a legjobb értékeket, így ezt mutatom be részletekbe menőleg, a többit csak érintőlegesen.

1. Szószám-vektorok

A módszer használata során az adatkészlet egy mátrixnak feleltethető meg, amelyben minden sor egy dokumentumot képvisel, minden oszlop egy a korpuszból származó kifejezés, a cellák pedig az egyes kifejezések frekvenciaszámát jelenti az egyes dokumentumokban.

2. Szóbeágyazások

A szavak és dokumentumok reprezentálásának egy olyan formája, amely neurális háló alapú. A vektortéren belül egy szó helye a szövegből tanulható meg, és a szót használatakor körülvevő szavakon alapul.

3. Szöveg vagy NLP alapú feature-ök

Manuálisan létrehozhatjuk azokat a featreu-öket, amikről úgy gondoljuk, hogy fontosak lehetnek a kategóriák közötti megkülönböztetés során (ilyen lehet a szósűrűség, a karakterek vagy szavak száma stb.). Használhatunk NLP alapú feature-öket is a Part of Speech modellek segítségével, a POS tag-ek frekvenciaeloszlását felhasználva.

4. Topikmodellek

A Topikmodellezés olyan technika, mely során a szavak csoportjait egy dokumentum-gyűjteményből azonosítják, mely a gyűjtemény legjobb információit tartalmazza. Egyes módszerek, mint például a Látens Dirichlet-allokáció, megpróbálnak minden témát szavak szerinti valószínűségi eloszlás alapján ábrázolni (maguk a tokenek ugyanis értelem nélküliek ebben az esetben).

5. Term Frequency-Inverse Document Frequency vektorok

A TF-IDF egy pontszám, mely a kifejezés relatív fontosságát mutatja a dokumentumban, valamint az egész korpuszban. Két kifejezésből áll a TF a Term Frequency rövidítése, az IDF pedig az Inverse Document Frequency-é. Az előbbi a normalizált kifejezés előfordulási gyakoriságának meghatározását végzi. Nagymértékben függ a dokumentum hosszától, és a benne található szavak jellegétől. A normalizálásra amiatt van szükség, mert nem jelenthetjük ki, hogy a hosszabb dokumentumoknak nagyobb jelentősége lenne azáltal, hogy egy gyakori szó többször szerepel benne, mint egy rövidebb dokumentumban. 0 és 1 között vehet fel értéket. 0 esetén a kifejezés nem létezik a dokumentumban, 1 esetén pedig a dokumentumban szereplő összes szó megegyezik.

Minden dokumentum és szó esetében egyedi, de formálisan így tudjuk kifejezni:

$$tf(t,d) = f_{t,d}$$

ahol t a vizsgált kifejezést, d a dokumentumot jelöli, $f_{t,d}$ pedig a t előfordulási gyakoriságát a d dokumentumban.

Utóbbi pedig az inverz dokumentum gyakorisága, melyet a korpuszban található dokumentumok számának logaritmusának és azoknak a dokumentumoknak a számának elosztásával kapunk meg, melyekben az adott kifejezés szerepelt. A t kifejezés informativitását méri, relatív súlyként van jelen. Nagy korpuszok estén, az IDF érték "felrobbanna", ezért a hatás csillapítása végett a logaritmusát vesszük. A nullával való osztás problémájának elkerülése végett, mind a számlálóhoz, mind a nevezőhöz szokás egyet hozzáadni. Az egész kifejezéshez való 1-es érték hozzáadása pedig a 0-s súly elkerülését szolgálja:

$$idf_1(t, D) = log \frac{N + 1}{|\{d \in D : t \in d\}| + 1} + 1$$

ahol t a vizsgált kifejezést, D a vizsgált korpuszt, N = |D| a korpuszban szereplő dokumentumok számát, míg d a korpuszban szereplő dokumentumokat jelöli.

A TF-IDF pontszám az alábbi képlettel kapható meg:

$$tf - idf(t, d, D) = tf(t, d) * idf\ (t, D)$$

Értéke a dokumentum egyes szavainak megjelenési számával arányosan növekszik, és ellensúlyozásra kerül a korpuszban lévő dokumentumok számával, amik tartalmazzák adott szót. Ez annak arányosításában segít, hogy egyes szavak gyakrabban jelennek meg. Azt a tényt is figyelembe veszi, hogy egyes dokumentumok a TF kifejezés normalizálásával nagyobbak lehetnek, mint mások. A mondatok szavainak sorrendje nem kerül figyelembevételre - ahogy szószám-vektor esetében sem.

A korpuszban lévő dokumentumok reprezentálására, ahogy fentebb már utaltam rá, TF-IDF vektorizálást választottam, melyet a Python scikit-learn csomagjának erre használható implementációjának[10] segítségével hajtottam végre. A fearure-ök ezzel a módszerrel történő létrehozásakor az alábbi paramétereket módosítottam:

- ngram_range: meghatározhatjuk, hogy az n-gramok mely tartományát vesszük figyelembe (unigramok, bigramok, trigramok, ...). Beállíthatunk alsó és felső korlátot. Az n minden min_n <= n <= max_n értéke fel lesz használva.
- min_df: szókincs összeállításakor minden olyan kifejezést figyelmen kívül hagy, melynek dokumentumokban történő előfordulása szigorúan kisebb, mint adott küszöb. A

[10] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

szakirodalom cut-off-ként is emlegeti.

- max_df: szókincs összeállításakor minden olyan kifejezést figyelmen kívül hagy, melynek dokumentumokban történő előfordulása szigorúan nagyobb, mint adott küszöb (korpusz-specifikus stopszavak).
- max_features: olyan szókincset készít, mely csupán a korpuszban frekvencia szerint sorba rendezett legfelső maximum N feature-t veszi figyelembe.

Paraméterek	Tesztelt értékek
n_gram range	unigram és bigram, bigram és trigram, unigram, bigram és trigram
min_df	1, 2, 3, 5, 7, 10, 15
max_df	0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 1
max_features	None, 300, 500, 600, 700, 1000

4. táblázat: A TF-IDF vektorizálás során használt paraméterek lehetséges értékei

Ahogy a fenti táblázatból is látható többféle értéket és ezek kombinációit próbáltam ki. Arra számítottam, hogy a bigramok majd javítják a modell teljesítményét azáltal, hogy azokat a szavakat is figyelembe veszik így a modellek, melyek általában együtt jelennek meg. Első körben a rendkívül ritka szavaktól, melyek csak egy-egy bejegyzésben szerepelnek, nem szerettem volna megválni, hiszen lehet, hogy egyes nőket vagy férfiakat ezek tesznek egyedivé. Azonban az egyedi szavak zöme elgépelések eredménye. A modellek eredménye az alsó küszöb emelésével, míg a felső határérték csökkentésével javult - csak míg el nem érte az “optimális” értékeket. A maximális feature-ök számát legelőször nem korlátoztam, ellenben így a tanító megfigyelések számához képest túl sok feature-öm volt, ami gyakran túlillesztéshez vezethet, ezért a későbbiekben ezt módosítottam.

Végeredményben a modellek a legjobb előre jelző erővel abban az esetben bírtak, mikor unigramokat és bigramokat vizsgáltam, az 5 és afeletti előfordulási gyakoriságú és a dokumentumok 70%-nál kevesebben előforduló szavakra korlátoztam a szókincs összeállításakor felhasznált kifejezéseket, valamint a kifejezésgyakoriság szerint rendezett legfelső 500 feature-t vettem figyelembe.

5.6 Prediktív modellek

5.6.1 Hiperparaméterek hangolása

A kutatásom során futtatott algoritmusok mindegyike sokféle hiperparaméterrel rendelkezik, amelyeket be kell hangolni. Az egyes modellek tanításakor és a legjobb hiperparaméter-készlet meghatározásakor Zafra (2019) módszerét követve jártam el. Először is eldöntöttem, hogy mely paramétereket fogom hangolni. Majd definiáltam a lehetséges értékekből egy rácsot és Randomized Search[11]-öt hajtottam végre ötszörös keresztvalidálással, mivel így az egyes hiperparaméterek szélesebb értéktartományát tudtam lefedni nagy végrehajtási idő nélkül. Az ötszörös keresztvalidálás azt jelenti, hogy 5 darab halmazra osztjuk fel a rendelkezésre álló adatokat, majd 5 darab iteráció során mindig 4 darab halmazból fog állni a tanító halmazunk, a maradék 1-et pedig teszthalmazként értelmezzük. K értékét (5) önkényesen választottam meg. Miután megkaptam a legjobb hiperparaméterekkel rendelkező modellt - úgy, hogy leszűkítettem az egyes tartományokat - Grid Search[12]-öt végeztem szintén ötszörös keresztvalidálással, kifejezetten meghatározva a kipróbálandó beállítások minden kombinációját, hogy a hiperparaméter térben megtaláljam a legjobban teljesítő kombinációt.

Az egyes paraméterek SVM modell esetében[13]:

- C: a hibakifejezés büntetési paramétere, regularizációs paraméter.
- kernel: az algoritmusban használandó kerneltípust határozza meg.
- gamma: kernel együttható 'rbf' és 'poly' esetében.
- degree (fok): a 'pol' kernel funkció fokértéke.

Az egyes paraméterek LR modell esetében[14]:

- C: a regularizáció erősségének inverze. Kisebb értékek erősebb regularizációt jelentenek.
- multi_class: megadjuk, hogy a problémánk bináris.
- solver: az optimalizálási problémában használandó algoritmus.
- class_weight: az osztályokhoz tartozó súlyok.
- penalty: a büntetéskor használandó norma meghatározására szolgál.

[11] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.htm
[12] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
[13] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
[14] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

Az egyes paraméterek Rf modell esetében[15]:

- n_estimators: a fák száma az erdőben.
- max_feature: egy csomópont felosztásakor figyelembe vett feature-ök maximális száma.
- max_depth: a maximális szintek száma az egyes döntési fákban.
- min_samples_split: a belső csomópont felosztásához szükséges minimális számú minta.
- bootstrap: az adatpontok mintavételének módszere (visszatevéssel vagy anélkül).

Modell	Paraméterek	Tesztelt értékek
SVM	C	.0001, .001, .01, .1, 1, 10, 100
	kernel	linear, rbf, poly
	gamma	.0001, .001, .01, .1, 1, 10, 100
	degree	1, 2, 3, 4, 5
LR	C	0.001, 0.01, .1, 1, 10, 100
	multi_ class	ovr (bináris)
	solver	newton-cg, sag, saga, lbfgs
	class_weight	balanced, None
	penalty	l2
RF	n_estimators	1-től 100-ig ötösével lépkedve
	max_features	None, sqrt
	max_depth	None, 1-től 100-ig ötösével lépkedve
	min_samples_split	1, 2, 4
	bootstrap	True, False

5. táblázat: A TF-IDF vektorizálás során használt paraméterek lehetséges értékei

5.6.2 Teljesítmény mérése

Miután megtaláltuk a legjobb kombinációját a hiperparamétereknek, elvégeztük a hiperparaméter hangolást a tanító adatokkal, és ráillesztettük a modellt a tanítóadatokra, értékelnünk kell a teljesítményét a „látatlan" adatokon (teszthalmaz). Ha az osztályeloszlás kiegyensúlyozatlan, az accuracy (pontosság, helyes besorolási arány) rossz választásnak számít, mivel minden osztályt egyforma fontosságúként kezel, magas pontszámot ad azoknak a

[15] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

modelleknek, amelyek csak a leggyakoribb osztályt jósolják. Saját adataimra is igaz volt ez, ezért az alábbi összetett mutatókat választottam az accuracy helyett a modellek teljesítményének értékelésére. Az alternatív mérőszámok esetében a ritkább osztály (férfiak) szerepel egyesként, míg a többségi (nők) kettesként.

1. Recall (Felidézés) vagy Sensitivity (Szenzitivitás)

Megmondja, hogy az egyik osztály elemeinek hány százalékát jósolta ténylegesen abba az osztályba. Arra a kérdésekre válaszol, hogy mennyire érzékeny az osztályozó a pozitív esetek felderítésében, az összes releváns dokumentum közül mennyi szerepel a találatok között? Kiszámításakor az eredeti kategória elemszámait vesszük figyelembe. Hatékony rendszer esetén értéke magas. Nem veszi figyelembe, hogy adott dokumentum hol szerepel a találati listában.

$$Recall = \frac{True\ Positive}{Treu\ Positive + False\ Negative\ (Total\ Actual\ Positive)}$$

2. Precision (Pontosság/Megbízhatóság)

Azt méri, hogy a prediktív modellünk jóslatainak mekkora hányada igaz valójában, mekkora a helyesen besorolt elemek aránya. A modell annál jobban fog működni, minél kevesebb a zaj, az irreleváns találat. Hatékony rendszer esetén értéke magas. Precision értéke adott algoritmus mellett csak Recall rovására javítható. Nem veszi figyelembe, hogy adott dokumentum hol szerepel a találati listában, nagyon érzékeny az adathiányokra, sem az osztályok nagyságának arányait.

$$Precision = \frac{True\ Positive}{Treu\ Positive + False\ Positive\ (Total\ Predicted\ Positive)}$$

3. F1-Score

A Precision és a Recall parametrikus harmonikus közepe (súlyozott aránya). Kiegyensúlyozott F-mértéket használunk, mindkét tényező egyenlő súllyal szerepel. Nem veszi figyelembe, hogy adott dokumentum hol szerepel a találati listában. Legjobb értékét 1, legrosszabbat pedig 0 esetén éri el.

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Recall, Precision és F1-Score mutatókat Tikk (2007) alapján ismertettem.

4. Area Under the ROC Curve (AUC)

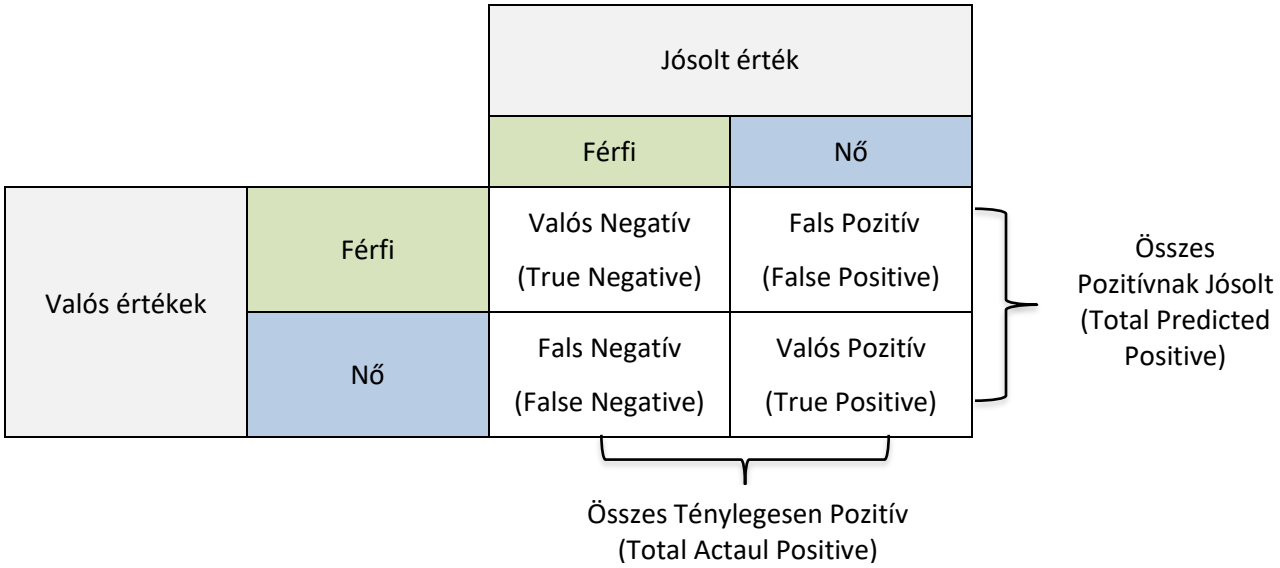
A ROC, azaz Receiver Operating Characteristic ábra a bináris osztályozók teljesítményének osztályozására szolgál. Kompromisszumot (trade-off) kínál az Igaz Pozitív Arány (TPR) és a Fals Pozitív Arány (FPR) között különböző klasszifikációs küszöbértékek esetén. A görbe mentén minden egyes pont egy az osztályozó által generált modellnek feleltethető meg. Nem érzékeny az osztályok kiegyensúlyozatlanságára. Illesztése mögött húzódó algoritmus bemeneti változói a valódi címkék és a jósolt valószínűségi értékek mellett a pozitív (esetemben férfiak) és negatív (esetemben nők) megfigyelések számai.

$$TPR = Sensitivity = \frac{Treu\ Positive}{True\ Positive + False\ Negative}$$

$$FPR = \frac{False\ Positive}{True\ Negative + False\ Positive}$$

A ROC görbe alatti terület (AUC-ROC) bináris osztályú problémák esetén a rangsorolás minőségének mérésére szolgál, megadja, hogy egy véletlen pozitív eset milyen valószínűséggel kerül a rangsorban előbbre. Tökéletes modell esetén értéke 1-gyel egyenlő. Ha random találgatást végez, értéke 0,5-öt vesz fel (Tan, Steinbach, Kumar, 20006).

A keveredési mátrix (confusion matrix) pedig az osztályozási modell teljesítményének leírásában van segítségünkre.



2. ábra: Keveredési mátrix

5.7 Eredmények

Az adatelőkészítési folyamtok elvégzését, és a tanítómintára illesztett vektorizálást követően a posztokból álló adathalmazomon mindhárom modellt teszteltem. Első körben SVM modelleket, ezt követően LR modelleket végül pedig RF modelleket futtattam. A modellek által a teszthalmazon elért legjobb eredményt az alábbi táblázatban mutatom be, ahol RF helyett már az NB három különböző fajtájával kapott eredményeket ismertetem.

Ahogy fentebb már utaltam rá a Random Forest módszer meglepő eredményeket hozott, ami később a „lecserélésre" késztetett. Az algoritmus a hiperparaméterek semelyik kombinációjában sem volt képes elkülöníteni a férfiakat és nőket. Csak nőket prediktált a tanító halmaz méretének növelését követően, ahogy keresztvalidáció végrehajtása után is. A modell előtt futtatott SVM és LR kapcsán azt tapasztaltam, hogy a két nem elkülönítése nem lehetetlen vállalkozás. Az RF „kudarca" mögött meghúzódó okot valószínűsíthetően a kis férfi mintaelemszámban kell keresni.

Modell	Mutató	Férfi	Nő
SVM	Precision	0.70	0.95
	Recall	0.78	0.93
	F1-Score	0.74	0.94
LR	Precision	0.47	0.97
	Recall	0.89	0.78
	F1-Score	0.62	0.86
CNB	Precision	0.47	0.94
	Recall	0.78	0.80
	F1-Score	0.58	0.86
BNB	Precision	0.50	0.97
	Recall	0.89	0.80
	F1-Score	0.64	0.88
GNB	Precision	0.64	0.95
	Recall	0.78	0.90
	F1-Score	0.70	0.92

6. táblázat: Az egyes modellek teszthalmazon elért eredményei

Modell	Jósolt Férfi	Ténylegesen Férfi	Jósolt Nő	Ténylegesen Nő
SVM	2	7	3	37
LR	1	8	9	31
CNB	2	7	8	32
BNB	1	8	8	32
GNB	2	7	4	36

7. táblázat: Az egyes modellek keveredési mátrixának eredményei

5.7.1 A legjobb osztályozó bemutatása

A futtatott modellek közül célom volt egy „legeslegjobb” kiválasztása. Értékelési metrikaként a ROC görbe alatti értéket választottam. Az AUC értéke 0.5 és 1 között mozoghat, és egy osztályozó annál jobb, minél magasabb értéket ér el. Ezen érték teszthalmazon mért eredménye alapján rendezve a modelljeim sorrendje az alábbiak szerint alakult:

Modell	Tanító halmaz	Teszthalmaz
SVM	1	0.85
BNB	0.89	0.84
GNB	0.94	0.84
LR	0.89	0.83
CNB	0.89	0.79

8. táblázat: Az egyes modellek AUC-ROC értékeik alapján sorba rendezve

A fenti táblázat alapján talán evidens lehetne az SVM modellt kinevezni „legeslegjobb”-nak, az elért első helyezése miatt. Azonban figyelembe kell venni a tanító halmazon elért nagyon magas eredményét – és az ezzel szembeni alacsonyabb teszthalmazon elért AUC-ROC értékét. Az egyes érték tökéletes modellt jelent, de kicsit árulkodó is, ez az optimális érték túlillesztésre utal(hat). Egy tanító adatokra túlontúl jól illeszkedő modell pedig rosszabb általánosítási hibával rendelkezhet, mint egy nagyobb tanítási hibával rendelkező. Emiatt a Bernoulli Naiv Bayes modellt választottam, mint az az osztályozási módszer, mely az adataimat a legjobban/legpontosabban el tudta különíteni. Döntésem mellett meghúzódó másik indok, hogy

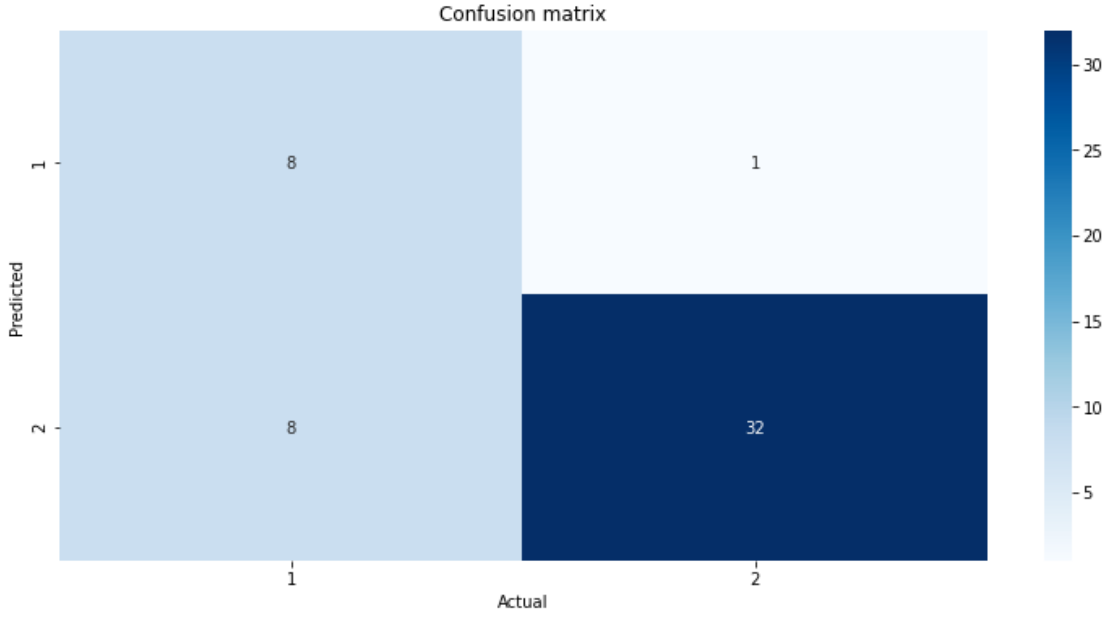
az adatokat Recall értékeik alapján is sorba rendeztem, és ez a modell ekkor is a második helyezést érte el. Valamint abban a tekintetben is jó helyen végzett, hogy összesen hány egyént sorolt rossz osztályba. Noha az SVM összességében nézve csak 5 embert „rontott el”, de a fentebb is említett valószínűsíthető túlillesztés miatt, új, eddig nem látotta adatokon nem biztos, hogy ilyen jól teljesítene. A BNB a maga 9 félreosztályozásával itt is második helyezett lett.

Classification report

	precision	recall	f1-score	support
1	0.50	0.89	0.64	9
2	0.97	0.80	0.88	40
accuracy			0.82	49
macro avg	0.73	0.84	0.76	49
weighted avg	0.88	0.82	0.83	49

3. ábra: A BNB modell osztályozási jelentése

A modell mikor egy személyt férfiként jósol meg az esetek 50 százalékában helyesen jár el. A nőként való jóslás esetében 3 százalékban hibázik. A férfi osztály elemeinek 89 százalékát, míg a nők esetében 80 százalékát jósolja a megfelelő egységbe.



4. ábra: A BNB modell keveredési mátrixa

A keveredési mátrix jól szemlélteti, hogy az algoritmus a teszthalmazba került 9 férfi esetében 8-at jósol helyesen férfinak, és 1-et tévesen nőnek. Míg a 40 nő esetében 32-t prediktált nőnek, 8-at pedig férfinak.

Nem csak a BNB modell, hanem az összes többi futtatott módszer esetében általánosságában elmondható, hogy a férfiak pontosabb prediktálása a nők helyes jóslási eredményének csökkenésével - és fordítva - járt együtt.

Megvizsgáltam az emberek posztjait, akiket a modell rossz kategóriásba sorolt be. Nem találtam olyan általános, minden egyénnél megjelenő domináns témákat, melyeket be lehetne azonosítani a férfiként klasszifikált női felhasználók esetében, sem a nőként klasszifikált férfiak esetében. Egyesével megnézve az egyes eseteket, a férfi személy esetében lehet olyan szavakat találni, mely alapján kissé érthetőbbé válik, hogy miért rossz osztályba került. Elolvasva a felhasználó posztjait, számunkra is nehézkes lehet eldönteni, hogy éppen milyen nemű emberről van szó. A nőként való azonosítás felé billenthetik a mérleg nyelvét a zölddel kiemelt szavak, amik jellemzően inkább nők által használt kifejezések. A felhasználó posztjai ömlesztve és kisbetűsítve, de eredeti formájukban szerepelnek. Egyedül azokat a posztokat távolítottam el, ahol esetlegesen benmaradt egy-egy megemlített ismerőse.

'kipróbáltam a tesco önkiszolgáló kasszáját és ezzel nyerhetek naponta egy 5000 forintos vásárlási utalványt vagy egy nagybevásárlást 100 000 forint értékben próbáld ki te is klikk ide a korábbi tromos napok kiemelkedően sikeres szereplőjeként alakításod és szépséged örökre emlékezetes számomra gratulálok kislányodhoz további sikereket kívánok 72 gyerekjáték a szedd magad pluszjel edd magad akció igazi élmény egészséges hasznos időtöltés az egész családnak minden korosztálynak a wellnessre pályázom ki a galambot szereti rossz ember nem lehet nagyon jó úticélok tiszta vízből szódavíz erre jó a sodastream várom a sorsolást'

5. ábra: Példa nőként klasszifikált férfi felhasználóra.

A rosszul besorolt nők esetében nehézkes kiemelni olyan kifejezéseket, melyek a félreklasszifikálást befolyásolhatták. A 8 hölgy bejegyzéseit vizsgálva mögöttes oknak talán az emelhető ki, hogy némelyikük előszeretettel oszt meg híreket, álláshirdetéseket, tudományos/ismertető cikkeket, melyekben lehetséges, hogy nehezebb nyelvhasználati mintákat találni és ez félrevezetheti a tartalmi alapon döntő osztályozókat.

6 Összegzés

Diplomamunkám keretén belül azt jártam körbe, hogy a rendelkezésemre álló magyar nyelvű Facebook posztokból álló korpuszomon belül meg lehet-e különböztetni férfiakat és nőket nyelvhasználatuk alapján.

Elsőként röviden ismertettem az írott beszélt nyelv fogalmát és bemutattam a jellegzetességeit. Ezt követően definiáltam a Natural Languege Processing, szövegbányászat és felügyelt tanulás fogalmát, melyen belül kitértem a klasszifikációra, és az osztályozó modellek illesztésének lépéseiről is áttekintést nyújtottam. Mindezek után a teljesség igénye nélkül összegeztem az általam legrelevánsabbnak gondolt korábbi tanulmányok eredményeit. Kitérve arra, hogy a szerzők milyen adatforrásokból indultak ki, és hogy az alkalmazott statisztikai módszerek segítségével találtak-e empirikus bizonyítékot a nemek közti nyelvhasználat különbözőségének meglétére, és ha igen, milyen eltéréseket véltek felfedezni.

A szakirodalmi áttekintést az általam végzett kutatás lépéseinek és eredményeinek bemutatása követte. A Facebook posztokból álló korpuszomon az adatelőkészítési folyamatok után négy fajta osztályozó algoritmus teljesítményét vizsgáltam. A Support Vector Machine, Logisztikus Regresszió, Random Forest és Naiv Bayes modellek esetében ismertettem azok elméleti hátterét, valamint kitértem arra is, hogy melyek azok az elemek, melyek kapcsán különbség fedezhető fel a modellek működési mechanizmusa között. A modellillesztés folyamatának bemutatása során kitértem azokra a hiperparaméterekre, melyeket a legjobb kombinációk megtalálása végett hangoltam. Célom volt egyetlen egy legjobbnak kikiáltható modell megtalálása, melynek segítségével a legsikeresebben szét tudom választani a nők és férfiak osztályát. Az ezzel elért eredményeket ismertettem, kitérve arra is, hogy a tévesen prediktált egyének esetében felfedezhető-e valamilyen mintázat.

A futtatott modellek által elért eredmények összeségében nem voltak rossznak mondhatóak, a nemek közötti nyelvhasználat különbség a Facebookon is megjelenik. Ám a korábbi strukturálatlan szövegeken végzett azonos témájú kutatásokkal való teljeskörű összehasonlítás több okból sem lehetséges. Legfőképp amiatt, hogy a megelőző elemzések kiegyensúlyozott adatbázisokkal való operálása végett accuracy-t használtak a modellek teljesítőképességének értékelésére. Míg én, az előbbi ellentetjének fennállása miatt összetett mutatók segítségével

értékeltem az egyes módszereket. Másrészt korábban angol nyelvű (Facebook) szövegeket felhasználva végeztek felméréseket, tudomásom szerint magyar Facebook posztokon alapuló nyelvhasználati kutatást még nem hajtottak végre.

A kapott eredmények önmagukban is hasznosíthatóak, ám a nem valószínűségi kényelmi minta miatt általánosíthatóságuk korlátozott. Önálló felhasználásuk mellett ideális kiindulópontot jelenthetnek egy olyan gyakorlatban hasznosítható kutatáshoz, melyben esetlegesen a szerzőket más demográfiai változók mentén is kategorizálják (pl.: iskolai végzettség, életkor, társadalmi státusz) és amely aztán a személyre szabott marketing, vagy a digitális bűnüldözés és kiberbiztonság területén alkalmazható.

Diplomamunkámnak ez már nem képezte részét, de jövőbeni további munkálatokként érdemes lehetne megnézni a legjobbnak választott modell univerzalitását, hogy lehetséges-e a kapott eredményeket a konkrét mintán kívül adott műfajon belül más mintára is általánosítani. Esetlegesen – ha a későbbiekben ilyen adatok is elérhetővé válnak – megvizsgálni, hogy a társadalmi státusz vagy foglalkozás bevonásával lehetséges lenne-e választékosabb modell kiépítése, mely segítené a nem és a nyelvhasználat kapcsolatának mélyebb megértését. Ezen kívül a keletkezett modelleket érdekes lehetne olyan szempontból is megvizsgálni, hogy feature vektorokkal kiegészítve – mint például használt emoticonok halmozódása, a leggyakoribb szófajok, vagy a személyes névmások aránya – az illeszkedésük javítható lenne-e.

7 Irodalomjegyzék

Andó, É. (2010). E-nyelv – Netbeszéd. Az elektronikus kommunikáció nyelvi jellemzői. In Tudományos Közlemények, 2010/23.

Aragón, M. E., & López-Monroy, A. P. (2018). A Straightforward Multimodal Approach for Author Profiling. Notebook for PAN at CLEF 2018. CLEF 2018 Evaluation Labs and Workshop -- Working Notes Papers. Padova: CEUR-WS.org

Argamon, S., Koppel, M., Fine, J., & Shimoni, A. R. (2003). Gender, genre, and writing style in formal written texts.

Babbie, E. (2003). A társadalomtudományi kutatás gyakorlata. Balassi Kiadó, Budapest.

Balázs, G. (2005). Az internetkorszak kommunikációja. In Balázs Géza - Bódi Zoltán: Az internetkorszak kommunikációja - Tanulmányok. Budapest, Gondolat - INFONIA.

Biber, D., Conrad, S. & Reppen, R. (1998). Corpus linguistics: Investigating language structure and use. Cambridge, England: Cambridge University Press.

Boda, M. (2007). Az internet hatása a XX-XXI. század kommunikációjára és nyelvhasználatára. A Budapesti Kommunikációs Főiskola negyedéves folyóirata. V/1.

Bódi, Z. (2004). A világháló nyelve. Internetezők és internetes nyelvhasználat a magyar társadalomban. Budapest, Gondolat Kiadó.

Bradley, P. H. (1981). The folk linguistics of women's speech: An empirical examination. Communication Monographs.

Breuer, J., Kmetty, Z., Haim, M., Stier, S. (2021). User-focused approaches for collecting Facebook data in the "post-API age. Kézirat.

Brownlow, S., Rosamon, J. A. & Parker, J. A. (2003). Gender-linked linguistic behavior in television interviews. Sex Roles.

Buda, Zs. (2011). Az internet hatása a nyelvhasználatra. Fiatalok fogalmazás- és kifejezőkészsége az internethasználattal összefüggésben. In Tudományos Közlemények, 2011/26.

De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. In: Seventh international AAAI conference on weblogs and social media.

Dovidio, J. F., Brown, C. E., Heltman, K., Ellyson, S. L. & Keating, C. F. (1988). Power displays between women and men in discussions of gender-linked tasks: A multichannel study. Journal of Personality & Social Psychology.

Eckert, P., & McConnell-Ginet, S. (2003). Linking the linguistic to the social. In P. Eckert, & S. McConnell-Ginnet, Language and Gender. Cambridge: Cambridge University Press.

Eichstaedt, J. C., et al. (2018). Facebook language predicts depression in medical records. Proceedings of the National Academy of Sciences 115.44 (2018): 11203-11208.

Érsok, N. Á., (2003). Írva csevegés – virtuális írásbeliség. In Magyar Nyelvőr 2003/1.

Fink, C., Kopecky, J., & Morawski, M. (2012). Inferring Gender from the Content of Tweets: A Region Specific Example. Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media. Association for the Advancement of Artificial Intelligence.

Garimella, A. & Mihalcea, R. (2016). Zooming in on Gender Differences in Social Media. In Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning. New York: Springer.

Herring, S. C. (2000). Gender differences in CMC: Findings and implications. Computer Professionals for Social Responsibility Journal, 18/1.

Huffaker, D. & Calvert S. (2005). Gender, Identity, and Language Use in Teenage Blogs. Journal of Computer-Mediated Communication, 10/2.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013). An introduction to statistical learning. New York: Springer.

Karvalics, L. Z. (2012). Információs kultúra, információs műveltség - egy fogalomcsalád. In Információs Társadalom – 12/1.

Kovács, E. (2014). Többváltozós adatelemzés. Budapest: Typotex Könyvkiadó.

Kubik, Gy. B. (2016). Osztályozás: felügyelt tanulási módszerek. In Sebők, M. (2016). Kvantitatív szövegelemzés és szövegbányászat a politikatudományban.

Laczkó, M. (2007). Napjaink tizenéveseinek beszéde szóhasználati jellemzők alapján. In Magyar Nyelvőr, 2007/2.

Lakoff, R. (1973). Language and Woman's Place. Language in Society.

Lakoff, R. (1975). Linguistic Theory and the Real World. New York: Harper Colophon Books.

Lee, C. (2003). How Does Instant Messaging Affect Interaction Between the Genders? Stanford, CA: The Mercury Project for Instant Messaging Studies at Stanford University.

Liddy, E. D. (2001). Natural Language Processing. In Encyclopedia of Library and Information Science, 2nd Edition. New York: Marcel Decker Inc.

Markov, I., Gómez-Adorno, H., & Sidorov, G. (2017). Language- and Subtask-Dependent Feature Selection and Classifier Parameter Tuning for Author Profiling. CLEF 2017 Evaluation Labs and Workshop -- Working Notes Papers, 11-14 September, Dublin, Ireland. Dublin: CEUR-WS.org.

Mehl, M. R. & Pennebaker, J. W. (2003). The sounds of social life: A psychometric analysis of students' daily social environments and natural conversations. Journal of Personality & Social Psychology.

Minya, K. (2003). Mai magyar nyelvújítás – Szókészletünk módosulása a neologizmusok tükrében a rendszerváltástól az ezredfordulóig. Budapest: Tinta Könyvkiadó.

Molnár, Cs. (2016). Osztályozás (klasszifikáció). In Sebők, M. (2016). Kvantitatív szövegelemzés és szövegbányászat a politikatudományban.

Molnár Zs., & Molnárné, C. C. (2009). A tiszta magyar nyelv kincseinek megőrzése és tanítása. Budapest: Püski Kiadó.

Molnos, A. (2003). Jövőnkért, a magyar nyelv ügyében. Debrecen: Lélektani Szaknyelv Megújításáért közhasznú alapítvány.

Mulac, A. & Lundell, T. L. (1986). Linguistic contributors to the gender-linked language effect. Journal of Language & Social Psychology.

Mulac, A. & Lundell, T. L. (1994). Effects of gender-linked language differences in adults' written discourse: Multivariate tests of language effects. Language & Communication.

Mulac, A., Lundell, T. L. & Bradac, J. J. (1986). Male/female language differences and attributional consequences in a public speaking situation: Toward an explanation of the gender-linked language effect. Communication Monographs.

Mulac, A., Seibold, D. R. & Farris, J. L. (2000). Female and male managers' and professionals' criticism giving: Differences in language use and effects. Journal of Language & Social Psychology.

Nádasdy, Á. (2001). Használjunk minél több idegen szót. In: Magyar Narancs, 2001/08/02.

Naveenkumar, D. S. R., Kiran, M. K., Reddy, K. T., & Raju, V. S. (2015). Applying NLP Techniques to Semantic Document Retrieval Application for Personal Desktop. In: Emerging ICT for Bridging the Future-Proceedings of the 49th Annual Convention of the Computer Society of India (CSI). Springer, Cham.

Németh, R., Katona, E. R., Kmetty Z. (2019). Az automatizált szövegelemzés perspektívája a társadalomtudományokban. In Szociológiai szemle 30/1.

Németh, R., Koltai, J. (2021). The Potential of Automated Text Analytics. In Social Knowledge Building in Pathways Between Social Science and Computational Social Science.

Newman, M. L., Groom, C. J., Handelman, L. D. & Pennebroke, J. W. (2008). Gender Differences in Language Use: An Analysis of 14,000 Text Samples. In Discourse Process.

Nguyen, D., Gravel, R., Trieschnigg, D., Meder, T. (2013). How Old Do You Think I Am?: A Study of Language and Age in Twitter. The 7th International AAAI Conference on Weblogs and Social Media.

Rakovics, M. (2016). Adattudomány jegyzet.

Rangel, F. & Rosso, P. (2013). Use of Language and Author Profiling: Identification of Gender and Age.

Ray, S. (2017). Understanding Support Vector Machine (SVM) algorithm from examples (along with code)
https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/
Utoljára megtekintve: 2021. április 12.

Sap M., Park G., Eichstaedt J. C., Kern M. L., Stillwell D., Kosinski M., Ungar L. H., Schwartz H. A. (2014). Developing Age and Gender Predictive Lexica over Social Media. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)

Savicki, V. (1996). Gender language style and group composition in internet discussion groups. Journal of Computer-Mediated Communication, 2/3.

Schler, J., Koppel, M., Argamon, S., Pennebaker, J. (2006). Effects of Age and Gender on Blogging. Papers from the 2006 AAAI Spring Symposium AAAI.

Sík Zoltán (2001). Gyermekek az információs Bábelben. In Gabos, E. (2000). A média hatása a gyermekekre és fiatalokra II. Budapest: Nemzetközi Gyermekmentő Szolgálat Magyar Egyesülete.

Szita, I., Szepesvari, Cs. (2010). Model-based reinforcement learning with nearly tight exploration complexity bounds. Omnipress.

Szokolszky, Á. (2004). Kutatómunka a pszichológiában.

Tan, Pang-Ning, Steinbach, M., Kumar, V. (2011). Bevezetés az adatbányászatba. Panem Kft.

Thomson, R. & Murachver, T. (2001). Predicting gender from electronic discourse. British Journal of Social Psychology.

Tikk, D. (2007). Szövegbányászat. Budapest: Typotex Kiadó.

Vincze, V., Üveges, I., Szabó, M. K. & Takács, K. (2021). A magyar beszélt és írott nyelv különböző korpuszainak morfológiai és szófaji vizsgálata. In XVII. Magyar Számítógépes Nyelvészeti Konferencia.

Weatherall, A. (2002). Gender, language, and discourse. London: Routledge.

Weaver, S. D. & Gahegan, M. (2007). Constructing, visualizing, and analyzing a digital footprint. In Geographical Review, 97/3.

Zafra, M. F. (2019). Text Classification in Python.
https://towardsdatascience.com/text-classification-in-python-dd95d264c802
Utoljára megtekintve: 2021. április. 12.

Zhang, C., Zhang, P. (2010). Predicting gender from blog posts. Technical Report. USA: University of Massachusetts Amherst.

Zhang H. (2004). The Optimality of Naive Bayes.