

Szegedi Tudományegyetem
Informatikai Tanszékcsoport

Véleménydetekció a szociális médiában

Diplomamunka

Készítette:
Hangya Viktor
programtervező informatikus
MSc szakos hallgató

Témavezető:
Dr. Farkas Richárd
egyetemi adjunktus

Szeged
2014

Tartalomjegyzék

Feladatkiírás	4
Tartalmi összefoglaló	5
Bevezetés	6
1. Információkinyerés	8
1.1. Természetesnyelv-feldolgozás	8
1.2. Véleménydetekció	10
1.3. Szociális média	11
1.3.1. Termék értékelő szolgáltatások	12
1.3.2. Twitter	12
2. Adatbányászati módszerek	14
2.1. Felügyelt módszerek	14
2.1.1. Naïve Bayes osztályozó	14
2.1.2. Maximum entrópia osztályozó	15
2.2. Felügyelet nélküli módszerek	16
2.2.1. Látens Dirichlet allokáció	16
2.3. Kiértékelési mértékek	17
3. Dokumentumszintű véleménydetekció	20
3.1. Előfeldolgozás	20
3.2. Jellemzőkinyerés	21
3.3. Eredmények	22
4. Entitásorientált véleménydetekció	27
4.1. Jellemzőkinyerés	27
4.2. Eredmények	29

5. Aspektusorientált véleménydetekció	32
5.1. Nyelvtani elemzés alapú jellemzők	32
5.2. Egyéb jellemzők	34
5.3. Szavak jelentésének meghatározása	35
5.4. Eredmények	35
6. Kapcsolódó munkák	39
7. Összefoglalás	41
Irodalomjegyzék	45
Nyilatkozat	46
Köszönetnyilvánítás	47

Feladatkiírás

Egyre erősödő trend, hogy társadalmunk a szociális médián (blog, termékértékelő oldalak, twitter, facebook stb.) keresztül mondja el véleményét, közöl információt. Ennek az adatnak a döntő többsége folyó szöveges formában áll elő és mennyiségét tekintve elképesztően nagy (big data). Másrészt a felhasználók által közölt vélemény/információ igen értékes vállalkozásoknak és minden döntéshozónak. A téma olyan újszerű gépi tanulási algoritmusok kidolgozása és tesztelése, amelyek a szociális média szövegbányászatához kapcsolódnak.

Tartalmi összefoglaló

Téma megnevezése

A dolgozatban különböző véleménydetekciós módszerek kerülnek bemutatásra, melyek célja szöveges dokumentumokban található vélemények polaritásának meghatározása.

Feladat megfogalmazása

Nap mint nap hatalmas mennyiségű szöveges adatot hozunk létre, mely feldolgozásához automatikus módszerekre van szükség. A dolgozatban véleménydetekciós módszerek három fajtája kerül bemutatásra, melyek során rendre globális, adott entitásra, illetve adott aspektusra vonatkozó vélemények polaritását kell meghatározni.

Megoldási mód

A kitűzött feladatok különböző gépi tanulás alapú algoritmusok használatával kerültek megoldásra. A kidolgozott módszerek során, a szöveges adatokból olyan jellemzők kerültek kinyerésre, melyek alapján a tanuló algoritmusok képesek meghatározni a szövegekben levő vélemények polaritását.

Alkalmazott eszközök

A módszerek Java programozási nyelven kerültek megvalósításra a MALLET természetesnyelv-feldolgozó eszköz segítségével.

Elért eredmények

Az eljárások három nemzetközi konferenciához kapcsolódó verseny keretein belül létrehozott adatbázisok segítségével lettek tesztelve, melyeken kimagasló eredményeket sikerült elérni.

Kulcsszavak

természetesnyelv-feldolgozás, véleménydetekció, gépi tanulás, adatbányászat, big data

Bevezetés

Az internet széleskörű elterjedésével az elmúlt néhány évben a szociális média népszerűsége jelentős mértékben megnövekedett, naponta több százmillió mondatnyi szöveget írnak le a felhasználók. Ezekben a szövegekben elmondják véleményüket különböző entitásokról, például termékekről, szolgáltatásokat nyújtó cégekről vagy hírességekről, de számos egyéb általános témáról is. Ezen adatok egyfajta visszacsatolást jelentenek az entitások számára, így ezek megfelelő felhasználása nagy segítség abban, hogy igazodni tudjanak ügyfelek igényeihez és fent tudjanak maradni vagy akár növelni tudják piaci részesedésüket. Ennek a hatalmas mennyiségű adatnak (BigData) az emberi feldolgozása lehetetlen. Olyan módszerekre van szükség, amelyek automatikus véleménydetekciót végeznek, azaz egyrészt képesek kiszűrni a megfelelő információt az adatfolyamokból, másrészt a szükségleteknek megfelelően képesek értelmezni azokat. Napjainkban számos kutatást folytatnak ebben a témakörben. Közkedvelt feladat például különböző termékek értékelésének automatikus feldolgozása, részvényárak változásának előrejelzése, választási eredmények megbecslése.

Munkám célja a szociális médiából származó szövegek gépi tanulás alapú elemzése véleménydetekciós szempontból. Ehhez szükség van a szövegek megfelelő előfeldolgozására, mely során a kapott szövegeken olyan átalakításokat kell végezni, hogy azok alkalmasak legyenek a további feldolgozásra. Ilyen előfeldolgozási lépések például a szövegek helyesírás-ellenőrzése vagy a rövidítések feloldása. Az adatok előkészítése után azt is el kell dönteni, hogy a kapott szövegrész hordoz-e számunkra releváns információkat, tehát egyfajta szűrést is kell végezni. Például a következő mondat negatív üzenetet hordoz, de nem BMW-vel kapcsolatos véleményt fejez ki: *Hosszasan szomorkodtam a BMW ülésén*. Egy dokumentum akár több témakörben is hordozhat információt. Nem mindegy, hogy a teljes dokumentumról szeretnénk átfogó képet kapni, vagy részletekbe menően az egyes témakörök alapján szeretnénk információt kinyerni. Dolgozatomban ismertetek három eljárást, melyek különböző szempontok szerint nyernek ki véleményeket. Az első egyfajta globális módszer, mely a teljes dokumentumot vizsgálja. A második egy adott entitásra vonatkozó véleményt nyer ki a szövegből. A harmadik ennél is tovább megy, az entitás egyes tulajdonságaira, aspektusaira vonatkozóan gyűjt véleményeket. A kidolgo-

zott eljárások nemzetközi gépi tanulási versenyeken is előkelő helyezéseket értek el és az eredményekből négy konferenciacikk is született [13, 14, 15, 28].

A dolgozat hátralevő része a következőképpen épül fel. Az első fejezetben bemutatásra kerül az információkinyerés, a hozzá kapcsolódó tudományos területek és az adatforrások. A második fejezetben a témakörben használható technikák kerülnek ismertetésre, azok elméleti háttere és alkalmazhatóságuk. A harmadik, negyedik és ötödik fejezetek tartalmazzák munkám során megoldott feladatokat, a kidolgozott módszereket és az elért eredményeket. A dolgozatban bemutatott feladatokhoz és módszerekhez hasonló kutatások a hatodik fejezetben olvashatók. Végül az hetedik fejezet a dolgozat tartalmát összefoglalja.

1. fejezet

Információkinyerés

Az információkinyerés szöveges állományokban rejlő ismeret detektálásának tudománya. Célja az adatokban levő információk kinyerésén túl, az azokban rejlő összefüggések megtalálása is, melyek számos esetben csak rejtve vannak jelen. A fejezetben ismertetésre kerülnek azon tudományágak, melyek információkinyeréssel kapcsolatosak.

1.1. Természetesnyelv-feldolgozás

A számítástechnika és a nyelvészet tudományágak egy közös területe a természetesnyelv-feldolgozás [17], melynek egyik célja az emberek és számítógépek közötti interakció segítése. A tudományág múltja egészen az 1950-es évekig nyúlik vissza. Az első kutatások között jelent meg az ún. *Turing-teszt*, mely a számítógép embertől való megkülönböztethetőségével foglalkozott. Azóta a kutatások számos irányt vettek, mint például gépi fordítás, szövegek értelmezése vagy generálása. A természetesnyelv-feldolgozás során a feldolgozandó szöveg lehet írott vagy hallott, egyedüli kritérium, hogy valamely emberi nyelven legyen.

A felhasznált technikákat tekintve szerteágazó tudományág. Régebben az alacsony számítási kapacitás miatt inkább a szabály alapú rendszerek domináltak. Ilyen rendszerek *ha-akkor* típusú elágazások tömkelegét tartalmazták, melyeket főként nyelvészek állítottak össze. Ezen szabályok megalkotása azonban időigényes, valamint csak egy adott típusú és nyelvű szövegre alkalmazható. A számítási kapacitás növekedésével lehetőség nyílt statisztikai alapú módszerek használatára. Így került előtérbe a mesterséges intelligencia és gépi tanulás, melyek nagyban segítik a feladatok megoldását és univerzalitását. Az eddig használt szabályokat ezen módszerek képesek automatikusan észlelni és felhasználni. A szabályok megtanulásához azonban szükség van ún. *tanító adatbázisra*, melyek hasznos mintákat tartalmaznak az adott cél eléréséhez. A gépi tanuló algoritmusok két fő csoportba oszthatóak, ezek a felügyelt és felügyelet nélküli algoritmusok. Felügyelt algo-

ritmus használata esetében olyan adatbázisra van szükség, melyben az egyes tanító példák jelölve vannak. Például minden film értékelésre adott, hogy az pozitív vagy negatív véleményt hordoz. Ennek az adatbázisnak a felhasználásával az algoritmus olyan jellemzőket tud felismerni, melyeket felhasználva el tudja dönteni jelöletlen példákra is a hordozott információk polaritását. Felügyelet nélküli tanítás során az adatbázis jelöletlen adatokat tartalmaz. Ezen módszerek általában klaszterezést végeznek, azaz adott számú csoportba sorolják az adatokat. Ilyen feladat például a szövegek témakör alapú csoportosítása. Ezen algoritmusok mindegyike esetében meg kell tudnunk mondani, hogy mennyire jól oldottuk meg a kitűzött célt. Ehhez egyrészt egy teszt adatbázisra van szükségünk, melyen vizsgálni tudjuk a kidolgozott algoritmust. Másrészt egy kiértékelési metrikát is definiálnunk kell, mellyel mérni tudjuk a kapott eredmények pontosságát.

Az alábbiakban a teljesség igénye nélkül tekintsünk át röviden néhány területet, melyekkel a tudományág foglalkozik. Ezek egy része közvetlenül valamilyen valós problémát old meg, mint például a gépi fordítás, míg másik része olyan információkat nyer ki, melyek komplexebb feladatok megoldásához szükségesek.

- **Morfológiai elemzés** – Feladat az egyes szavakat szegmentálása morféimákra és toldalékokra. Morféma a legkisebb jelentéssel bíró egység egy nyelvben. A feladat nehézsége az adott nyelv morfológiai gazdagságától függ. Például a magyar morfológiailag gazdag nyelv, ellenben az angol.
- **Koreferencia feloldás** – Adott mondatba vagy szövegrészletben el kell dönteni, hogy mely szavak utalnak ugyanarra az entitásra. Vegyük a következő példa szöveget: *István jól teljesített. Ő nagyon ügyes.* A második mondat esetében detektálnunk kell, hogy az *Ő* személyes névmás Istvánra utal.
- **Kérdés megválaszolás** – Az előző kettő feladat eredményeit legtöbb esetben komplexebb feladatok megoldásához használják fel, mint például kérdések megválaszolására. A feladat során adott kérdésre keresünk választ. Ekkor értelmezni kell a kérdést és szöveges választ kell rá adnunk. Tehát a megoldás része a szöveggenerálás is.
- **Beszéd szegmentálás** – Adott egy hanganyag, melyen egy vagy több személy adott nyelven szöveget generál. Feladat, hogy a szöveg alapján partíciónáljuk a hanganyagot szóhatároknál.
- **Beszéd felismerés** – A feladat bemenete ismét egy hanganyag, melyet írott szöveggé kell alakítanunk. Megoldás során használhatók beszéd szegmentálási technikák, de nem feltétlenül szükséges. A feladat *AI-teljes*, ezért különböző heurisztikákat alkalmaznak.

- **Szöveg felolvasás** – Az előző feladat ellentéte, mely során írott szöveget kell felolvasni. Fontos a megfelelő hangsúlyozás és a szünetek betartása, ezért megoldás során a szöveg bizonyos szintű megértésére van szükség.

1.2. Véleménydetekció

A természetesnyelv-feldolgozás egyik aktívan kutatott területe a *véleménydetekció* [22]. Célja, hogy főként internetes forrásokból származó adatok segítségével nyomon kövesse az emberek álláspontját különböző entitásokra nézve. Egy ilyen rendszer kiépítéséhez szükség van egy adatforrásra, mely a szükséges információkat szolgáltatja. Napjainkban a *szociális média* jelentős népszerűsége tett szert, szinte bármilyen témakörben tartalmaz információkat. Ez okból kifolyólag egy jó kiindulási alapot biztosít. Az adatok feldolgozására számos esetben gépi tanulási módszereket használnak, mivel ezek a technikák könnyedén adaptálhatók különböző szövegekre, nyelvekre.

A véleménydetekció a hétköznapi emberek és különböző cégek számára is hasznos lehet. Vegyük példának egy termék megvásárlásának menetét. A vásárlónak először is el kell dönteni, hogy mely terméket vásárolja meg a piacon fellelhetőek közül. A döntés segítéséhez jelentős mennyiségű információ található az interneten, például különböző értékelések, termékleírások. Ennek a rengeteg adatnak azonban hátránya is van. A legtöbb ember nem szeretne órákat tölteni azzal, hogy adott termékről véleményeket olvas, melyek számos esetben tartalmaznak ellentétes információkat is. Szükség van olyan eljárásokra, melyek képesek ezeket az adatokat összegezni és könnyen áttekinthető formára hozni. Másfelől ezen információk hasznosak lehetnek a termékek gyártói számára is. Egy új termék megtervezése előtt információt gyűjthetünk arról, hogy más hasonló termékeknek milyen jó, illetve rossz tulajdonságaik vannak. Ezeket szem előtt tartva egy a vásárlók által kedveltebb terméket tervezhetünk meg.

Az emberek álláspontjának kiderítéséhez olyan dokumentumokat kell gyűjteni, melyek a kérdéses témához kapcsolódnak. Miután rendelkezünk ilyen dokumentumokkal, a következő lépés, hogy eldöntsük a vélemény polaritását. Ily módon egy osztályozási feladathoz jutunk, mely során a szóban forgó dokumentumról el kell dönteni, hogy pozitív, negatív vagy semleges véleményt hordoz-e. Egy dokumentum véleménytartalmának kizárólagos vizsgálata azonban nem elegendő a feladat megfelelő megoldásához. Sok esetben csak egy adott entitásra nézve szeretnénk véleményeket kinyerni. Ilyenkor szűrni kell a dokumentumokat, azaz el kell dönteni, hogy a megfelelő entitásról hordoz-e információkat. Ezen felül, előfordulhatnak olyan dokumentumok is, melyek nem csak egy entitásról szólnak, sőt egy mondaton belül is szólhatunk akár több entitásról is. Ilyenkor ún. *entitás-orientált* véleménydetekciót kell végezni, mely során meg kell határozni az

adott entitásra vonatkozó szövegrészeket és csak ezen adatokat felhasználva kell az osztályozást elvégezni. A feladat azonban még tovább specializálható. Főként termék értékelésekre jellemző, hogy annak különböző tulajdonságairól írunk véleményeket. Ezeket felhasználva részletesebben is informálódhatunk. Ily módon nem csak azt tudjuk megmondani, hogy mely entitást kedvelik az emberek és melyeket nem, hanem azok jó és rossz tulajdonságait is összegezhethetjük.

1.3. Szociális média

Szociális média alatt olyan virtuális hálózatot értünk, amelyben résztvevő személyek adatokat hoznak létre és osztanak meg egymással [18]. A hálózat résztvevőin túl, a szociális média fenntartásában fontos szerepet játszanak olyan alkalmazások, melyek működéséhez az internetkapcsolat elengedhetetlen. Ezen alkalmazások lehetővé teszik az emberek számára különböző tartalmak, vélemények megosztását és megvitatását.

Számos tulajdonságában különbözik a hagyományos médiától. Minőségét tekintve nagyon változó. Jellemző, hogy a tartalmuk kevésbé vagy egyáltalán nem felügyelt, így azok felhasználói szabadon publikálhatnak különböző tartalmakat. Sokszor előfordul, hogy helyesírási vagy igazságtartalmukat tekintve alacsony minőségűek. Ez a tulajdonság nagyban nehezíti az ilyen adatok feldolgozását. Egy másik különbség az interaktivitásban rejlik. Míg a hagyományos médiában az információ áramlás főként egyirányú, de legalábbis a visszajelzések lassúak, addig a szociális médiában az információ közlés sokkal interaktívabb, sokszor valós időben zajlik. Másrészt a közölt információk is gyorsan módosíthatók, ellentétben például egy nyomtatott újságcikkkel. Ezért is fontos, hogy rendelkezünk olyan technológiákkal, melyek automatikusan és gyorsan tudják feldolgozni ezt az információ áradatot. Mindezen felül a hagyományos média elérése számos esetben költségesebb is. Előfizetés szükséges különböző folyóiratok, újságok vagy TV-csatornák megtekintéséhez. Ezzel szemben, a szociális média tartalmának elérése legtöbb esetben ingyenes, csak internet kapcsolatra van szükség. Még ha bizonyos tartalmak megtekintése előfizetéshez kötött is, azok egy része sok esetben ingyenesen elérhető. Feldolgozás szempontjából is előnyösebb a virtuális média, mivel ellentétben nyomtatott újságokkal vagy TV-adásokkal, ezek elérése számítógépekkel sokkal egyszerűbb.

A szociális média tartalma számos formában jelenhet meg, az egyszerű szöveges weboldalaktól kezdve egészen a virtuális valóságokig. Számos olyan weboldal létezik, melyek mindenféle tudásanyagot gyűjtenek és tesznek elérhetővé az internet használói számára. Egy ilyen oldal a *Wikipedia*, mely lehetővé teszi bárki számára, hogy tetszőleges témában közzéjön információkat. Ezen információk mások által akár módosíthatóak vagy bővíthetőek is, tehát egyfajta együttműködést tesznek lehetővé. Egy másik nagyon el-

terjedt szegmense a szociális médiának a blogok és mikroblogok. Ezek olyan webszolgáltatások, melyek segítségével a felhasználók saját élményeiket írhatják meg, általában rövidebb, szabadabb formában. Ismertebb blog szolgáltatás a *Tumblr*, illetve mikroblog a *Twitter*. Napjainkban nagy népszerűsége szert tevő elemek a közösségi hálózatok, amelyek segítségével a felhasználók ismeretségi kapcsolatokat adhatnak meg. Ezen felül kommunikálhatnak is egymással, akár mások számára publikus módon is. A legismertebb és legnépszerűbb ilyen szolgáltatás a *Facebook*. Ezen hálózatok kutatásával a gráfelméletben élen járók sokat foglalkoznak. Számos nem szöveg alapú típus is létezik, mint például a *Youtube*, mely lehetővé teszi videók megosztását, vagy a *Second Life*, mely egyfajta virtuális világot teremt. A következőkben tekintsünk át kettő olyan szolgáltatást, melyek tartalma segítségével véleménydetekció végezhető és melyeket munkám során felhasználtam.

1.3.1. Termék értékelő szolgáltatások

A termék értékelő szolgáltatásra tekinthetünk úgy, mint különböző entitásokról információt közvetítő szolgáltatások. Egyrészt lehetőséget nyújtanak arra, hogy a felhasználók véleményezzék az általuk ismert entitást, másrészt ezek a vélemények publikusan elérhetők mindenki számára. Ily módon az internetet használók, könnyedén információhoz juthatnak az őket érdeklő entitásról. Számos olyan oldal létezik, melyek kifejezetten termékek értékelését teszik lehetővé, például *Epinions* vagy *Amazon*. Ezen szolgáltatások azon kívül, hogy szöveges értékelések írásának lehetőségét nyújtják, sokszor termékek diszkrét skálán történő pontozását is lehetővé teszik, ezáltal áttekinthetőbbé téve az értékeléseket. Tudományos munkák kimutatták, hogy az ilyen értékelések nagyban befolyásolják a vásárlási szokásokat és az egyes termékek tekintélyét [12]. Ezért ezek a szolgáltatások nem csak a termékek vásárlóinak, de azok gyártóinak is információt szolgáltatnak. Egyrészt arról, hogy milyen pozitívumokkal és negatívumokkal rendelkeznek termékeik, illetve mit csinál a konkurencia jobban vagy rosszabbul. Az értékelések formájukat tekintve szabályosnak mondhatóak. Mivel legtöbb esetben ezeknek a szövegeknek a létrehozása nem felügyelt, így tartalmazhatnak apróbb formai hibákat. Ennek ellenére a természetesnyelv-feldolgozás más területein kidolgozott módszerek alkalmazhatóak ezen szövegekre, ami megkönnyíti feldolgozásukat.

1.3.2. Twitter

A Twitter egy mikroblog szolgáltatás, mely 2006-ban jött létre kisebb csoportok kommunikációjának segítése céljából. Azóta hatalmas népszerűsége szert tett, 2012-ben 500 millió felhasználót számlált, akik naponta 340 millió hozzászólást, más néven *tweetet*

hoztak létre. Ezek a tweetek maximum 140 karakter hosszúak lehetnek, melyek szöveget, URL-eket és különböző Twitter-specifikus jelöléseket tartalmazhatnak. Lehetőség van arra, hogy egy bejegyzést adott felhasználónak címezzük, ilyenkor a @ karakter után meg kell adnunk a kívánt felhasználó nevét, például @János. Ekkor a kiválasztott felhasználó számára kiemelten jelenik meg ez a hozzászólás, de a többi felhasználó számára továbbra is látható marad. Egy másik lehetőség, hogy adott témával lássuk el bejegyzésünket, ezt a # karakter után írt témanév jelöléssel tehetjük meg, például #BMW. Ez arra szolgál, hogy meg tudjuk tekinteni az összes adott témájú tweetet. Ezekén felül lehetőség van még mások által létrehozott tweetek ún. *retweetelésére*, melynek célja, hogy újra megosszunk őket. A tweetekhez hozzá is szólhatunk, ezáltal egyfajta diskurzust hozva létre, bizonyos téma alapján.

Mivel a felhasználók számos témakörben fejtik ki véleményeiket tweetek formájában, így a Twitter hatalmas mennyiségű adatot szolgáltat, mely felhasználható véleménydetekcióhoz. A szövegek nagy részének alapvető jellemzője azonban, hogy írásbeli formájuk ellenére beszélt nyelvi sajátosságokat mutatnak [28]. A szóbeli kommunikáció valósídejűségét és multimodalitását egyszerre törekszik megtartani, így az élmény megosztásának gyorsasága és az érzelm kifejezés jelentős szerepet játszik a szövegekben. A gyorsaságot ugyanis a gépelés gyorsításával lehet elősegíteni: többek között központosítás és nagybetűk hanyagolásával, rövidítésekkel, egybeírással, valamint félregépeléssel. A érzelm kifejezés megnyilvánulhat a nagybetűhasználatban, a betűhalmozásban és hangulatjelek használatában. Egy példa tweet:

@Peti90 LOL a tegnapi buli naggyon jó volt :D :D #házibuli

Látható, hogy ezen szövegek feldolgozása nem egyszerű, előfeldolgozási lépésekre van szükség ahhoz, hogy megfelelően kezelni tudjuk a különböző jelöléseket és írási hibákat.

2. fejezet

Adatbányászati módszerek

A fejezetben ismertetésre kerül néhány módszer, melyek felhasználhatók különböző adatbányászati feladatok, illetve véleménydetekció elvégzéséhez. Ezen technikák gépi tanulási módszereken alapulnak, melyek két fő csoportra oszthatók, a felügyelt és felügyelet nélküli módszerekre. Munkám során az első kategóriából két módszer összehasonlításra is került. Különböző módszerek teljesítményének összehasonlítására számos technika létezik, melyek közül néhány a 2.3 fejezetben kerül bemutatásra.

2.1. Felügyelt módszerek

Felügyelt tanuló algoritmusok olyan gépi tanulási módszerek, melyek feladata az $f : X \rightarrow Y$ függvény meghatározása megfigyelések alapján. Egy osztályozó algoritmus bemenete egy tanító adatbázis, pontosabban jelölt adatok, vagy másképp tanító példák egy halmaza, melyek alapján a szóban forgó függvény meghatározható. Minden tanító példa egy $X \in \mathbb{R}^n$ jellemző vektor és $y \in Y = \{y_1 \dots y_m\}$ címke páros, mely lehet például egy a mondat szavaiból álló jellemzővektor és a mondat polaritása mint címke. A jelölt adatok alapján meghatározott függvény feladata ezután, hogy korábban nem látott elemekhez egy $\hat{y} \in Y$ címkét rendeljen, azaz például jelöletlen mondatok esetében meghatározza azok polaritását. Számos osztályozó algoritmus létezik [9], melyek közül munkám során a *Naïve Bayes* és a *Maximum entrópia* osztályozókat alkalmaztam.

2.1.1. Naïve Bayes osztályozó

A *Naïve Bayes* osztályozó algoritmus a Bayes-döntésméleten alapul [32], és azzal az erős feltételezéssel él, hogy $X_1 \dots X_n$ jellemzők feltételesen függetlenek, feltéve Y . Ez a feltevés nagyban egyszerűsíti $P(Y|X)$ meghatározását. A Bayes-szabály szerint a keresett függvény

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \quad (2.1)$$

módon számíthatók ki, ahol $P(Y)$ ún. *a priori* eloszlás, mely könnyedén becsülhető a tanító példák címkéi segítségével. $P(X)$ elhagyható, mivel értéke független az adott Y címkétől. Meg kell határozni még a $P(X|Y)$ *likelihood* eloszlást, melyet a függetlenségi feltevést használva a következőképpen számíthatunk:

$$P(X|Y) = P(X_1 \dots X_n|Y) = \prod_{i=1}^n P(X_i|Y) \quad (2.2)$$

ahol a $P(X_i|Y)$ valószínűségek *maximum-likelihood* becsléssel határozhatóak meg szintén a tanító adatbázisból. A szükséges valószínűségek kiszámítása után egy jelöletlen adatra a címke meghatározása tehát:

$$\hat{y} \leftarrow \arg \max_{y_k} P(Y = y_k) \prod_{i=1}^n P(X_i|Y = y_k) \quad (2.3)$$

Az erős feltevés ellenére az algoritmus számos gyakorlati feladat megoldásában jól használható. Előnye más algoritmusokkal szemben, hogy gyors, és kis méretű tanító adatbázisokon is képes jól működni.

2.1.2. Maximum entrópia osztályozó

A *maximum entrópia* osztályozó [6], vagy más néven multinominális logisztikus regresszió egy olyan osztályozási algoritmus, mely a keresett posteriori függvényt közvetlenül parametrikus formában keresi. A logisztikus regresszió általánosított változata, mivel nem csak bináris osztályozást képes elvégezni. A keresett függvény parametrikus alakja:

$$P(Y = y_k|X) = \frac{e^{w_k^\top X}}{1 + \prod_{i=1}^m e^{w_i^\top X}} \quad (2.4)$$

ahol w_i az y_i címkéhez tartozó súlyvektor. Keressük tehát azon súlyvektorokat, melyek legjobban magyarázzák a szükséges feltételes valószínűségeket, azaz:

$$\hat{W} \leftarrow \arg \max_W P(Y|X, W) \quad (2.5)$$

A maximalizálás zárt formában nem végezhető el, numerikus eljárások használatára van szükség. Használhatók gradiens alapú módszerek, ahol a cél a meghatározott modell által vétett hibák csökkentése, melyet a hibafüggvény gradienseinek felhasználásával hajthatunk végre. A súlyok kiszámítása után egy jelöletlen adatra a címke meghatározása hasonlóan az eddigiekhez:

$$\hat{y} \leftarrow \arg \max_{y_k} P(Y = y_k | X, W) \quad (2.6)$$

A módszer természetesnyelv-feldolgozásban sűrűn használt eljárás. Szövegek polaritásának meghatározása esetén a kiszámított w súlyvektorok megadják, hogy az osztályozás során mely jellemzők voltak fontosak. Nagy abszolút értékű súlyt fognak kapni az olyan szavak, melyek valamely polaritásra utalnak, például *jó*, *rossz*, *drága*, stb. Előnye a Naïve Bayes algoritmussal szemben, hogy nem feltételez függetlenséget az egyes jellemzők között, futási idejét tekintve azonban lassabb, ami a numerikus paraméterbecslő eljárások egyik következménye.

2.2. Felügyelet nélküli módszerek

Ahogy azt az előzőekben láttuk, felügyelt módszerek esetében olyan adathalmazra van szükségünk, melyben az adatok címkézettek. Ezzel szemben *felügyelet nélküli* módszereknek nincs szükségük jelölt adatbázisra. Ennél fogva egyszerűbb jelöletlen adathalmazt összeállítani, viszont hátránya hogy nincs semmilyen visszajelzés, például hibafüggvény, mely alapján a módszerek az adatokat feldolgozhatják. Mégis a cél az, hogy rejtett mintázatokat fedjünk fel az adatokban. Felügyelet nélküli módszerek közé tartoznak különböző klaszterező eljárások, melyek feladata az adatok csoportosítása adott szempont szerint. Egy másik hasznos terület, adatok vizualizálása, mely megoldható klaszterezéssel vagy dimenzió csökkentő eljárásokkal is. Utóbbi célja, hogy az adatokat leíró jellemzővektorokat úgy alakítsuk át, hogy azok megjeleníthetők legyenek különböző koordináta rendszerekben. A következőkben ismertetésre kerül egy ilyen módszer, mely szövegek témakor szerinti csoportosítását teszi lehetővé.

2.2.1. Látens Dirichlet allokáció

Látens Dirichlet allokáció [8] (LDA) egy generatív modellezési technika, mely célja hogy felfedje nagy szöveges adatok mögött rejlő rejtett témaköröket. Mint minden generatív modellezési technika, az LDA is egy olyan Bayes modellt feltételez, mely az adott adathalmazt generálja. A modellben szereplő valószínűségeket kell becsülni az adathalmaz segítségével, annak érdekében, hogy a dokumentumokat témakörök szerint csoportosítani tudjuk.

Az alapgondolat, hogy minden egyes dokumentum több témakörhöz is tartozhat és az adott dokumentum szavai ezen témakörök függvénye. Egy dokumentum generálása tehát a következő lépésekből áll:

1. véletlenszerűen válasszunk egy N számot, mely a dokumentum szavainak mennyisége lesz
2. véletlenszerűen válasszunk egy θ témakör eloszlást, mely felhasználható arra, hogy meghatározzuk a dokumentum egyes szavai mely témakörbe tartoznak
3. minden w_i szóra ($i = 1 \dots N$)
 - (a) válasszunk véletlenszerűen egy z_i témakört θ eloszlás segítségével
 - (b) válasszuk ki a w_i szót a z_i témakör alapján

A fent ismertetett modell és az adathalmaz segítségével meg kell határozni az egyes valószínűségeket. Az értékek számítása azonban zárt formulával nem lehetséges, ezért ismét numerikus módszerek használatára van szükség, mint például *Expectation Maximization* vagy *rejtett Markov Lánc* alapú módszerek. A valószínűségek ismerete után megadható, hogy az egyes dokumentumok mekkora valószínűséggel tartoznak az egyes témakörökbe, illetve az is, hogy az egyes témakörök mely szavakból állnak. Ezen információk később felhasználhatók felügyelt tanuló algoritmusok jellemzőkészletének növelésére és ezáltal a pontosság javítására.

2.3. Kiértékelési mértékek

Különböző adatbányászati módszerek fejlesztése során szükség van azok minőségének mérésére, annak érdekében, hogy tudjuk, mennyire jó a rendszerünk, illetve hogy lássuk a fejlesztések hatását. Ezen felül arra is hasznos, ha össze tudjuk hasonlítani a kidolgozott rendszert más módszerekkel. Az ilyen mértékek használatához szükség van egy D teszt adatbázisra, melyek jelölt adatokat tartalmaznak. Ezen adatokat fel kell címkézni a használt tanuló algoritmus segítségével, így minden $X \in \mathbb{R}^n$ elemre lesz egy $y \in Y$ valós és egy $\hat{y} \in Y$ következtetett címke. Számos kiértékelési metrikát alkottak meg, melyek ezeket felhasználva minőségi értéket számítanak ki. A következőkben tekintsük át néhány módszert, melyeket az eredmények ismertetésénél használni fogunk.

Egy egyszerű mérték az ún. *pontossági* metrika, mely gyakorta használatos különböző adatbányászati algoritmusok eredményének mérésére. Megadja a jól osztályozott teszt adatok arányát. Számítása:

$$\text{pontosság} = \frac{|\{\text{jól osztályozott adatok}\}|}{N} \quad (2.7)$$

ahol N a teszt adatok száma és $|\cdot|$ egy halmaz számosságát meghatározó művelet. A mérték könnyen értelmezhető, viszont számos esetben kevés információt szolgáltat vagy

félrevezető lehet. Tekintsünk egy olyan esetet, amikor a teszhalmazban egyik címke aránya nagyobb a többiénél. Ekkor egy egyszerű megoldás, mely során mindig ezt a címkét választjuk eredményként, magas pontosságot érhet el, viszont a megoldás a gyakorlatban használhatatlan.

A fent említett probléma kiküszöbölésére használhatók a *precizitás* és *lefedettség* mértékek. Mindkét mérték címkékre van definiálva, azaz ha van pozitív, negatív és semleges címke, akkor mindegyikre számíthatóak az értékek. A precizitás megadja, hogy a pozitívnak következtetett dokumentumok közül valójában mennyi pozitív:

$$\text{precizitás} = \frac{|\{y_k \text{ címkéjű adatok}\} \cap \{y_k \text{ következtetett címkével ellátott adatok}\}|}{|\{y_k \text{ következtetett címkével ellátott adatok}\}|} \quad (2.8)$$

míg a lefedettség érték megadja, hogy a valójában pozitív adatok közül mennyire következtettünk pozitív címkét:

$$\text{lefedettség} = \frac{|\{y_k \text{ címkéjű adatok}\} \cap \{y_k \text{ következtetett címkével ellátott adatok}\}|}{|\{y_k \text{ címkéjű adatok}\}|} \quad (2.9)$$

Szokás a két mérték harmonikus közepét számítani, melyet *F-mérték*nek nevezünk.

Egy hasonló mérték a *megbízhatóság* és *érzékenység* metrikák [2]. Az előző mértékektől ezek annyiban térnek el, hogy dokumentum párokra vannak definiálva, úgy, hogy egyfajta rendezést vizsgál. Például az eddigi pozitív, semleges és negatív címkék rendezettségét adhatnak meg, ha rendre +1, 0 és −1 értékeket rendelünk hozzájuk. Az említett mértékek számítása:

$$\text{megbízhatóság} = \frac{|\{(i, j) \in D^2 : \text{rel}(y_i, y_j) = \text{rel}(\hat{y}_i, \hat{y}_j) \wedge y_i \neq y_j\}|}{|\{(i, j) \in D^2\}|} \quad (2.10)$$

$$\text{szenzitivitás} = \frac{|\{(i, j) \in D^2 : \text{rel}(y_i, y_j) = \text{rel}(\hat{y}_i, \hat{y}_j) \wedge \hat{y}_i \neq \hat{y}_j\}|}{|\{(i, j) \in D^2\}|} \quad (2.11)$$

ahol $\text{rel}(y_i, y_j)$ az i és j dokumentumok között fennálló reláció. Látható tehát, hogy ezen mértékek minden olyan dokumentumpárt jónak vesznek, melyek között megfelelő reláció áll fent. Például ha egyik dokumentum pozitív, a másik pedig negatív, a rendszerünk pedig semleges és negatív címkéket választott, akkor ezen mértékek a megoldást jónak számítják, ellentétben például a pontosság metrikával. A két értékből, hasonlóan a precizitás és lefedettség mértékekhez, számítható a harmonikus közepük.

Egy rendszer kiértékeléséhez tehát szükség van tanító és teszt adatbázisokra és tanuló algoritmusra. Legtöbb feladat esetén nincs előre definiálva, hogy a teljes adathalmaz mely részét használjuk tanításra és mely részét tesztelésre. Megtehetnénk, hogy kézzel felosztjuk az adathalmazt, de ilyen statikus felbontás ahhoz vezethet, hogy az adott teszt halmazon jól viselkedik a kidolgozott módszer, viszont későbbiekben új adatokon már nem. Ennek elkerülését célozza meg a *keresztvalidáció*, mely lényegében a teljes adathalmazon értékeli ki. Pontosabban, felosztja n diszjunkt halmazra az adatbázist és n darab futtatást végez. Az i . futtatás során az i . halmazon végzi az eredmények pontosságának mérését adott metrikával és a maradék $n - 1$ halmazt pedig tanításra használja. Ezáltal n darab teszteredmény keletkezik, melyek végső lépésben átlagos pontossági értékek számítására használhatók. Munkám során tízszeres kereszt validációt alkalmaztam, azaz $n = 10$ értéket használtam teszteléskor.

3. fejezet

Dokumentumszintű véleménydetekció

Az elkövetkező három fejezetben bemutatásra kerülnek a munkám során kidolgozott véleménydetekciós módszerek és azok működési hatékonysága különböző adathalmazokon. Az első módszer az ún. *dokumentumszintű* véleménydetekció, mely célja rövid üzenetek globális polaritásának meghatározása, azaz adott üzenethez pozitív, negatív vagy semleges címkét kell rendelni annak tartalma alapján. Egy adott üzenet esetében, nem áll rendelkezésre többlet információ, mint például hogy mely entitásról vagy annak mely tulajdonságairól van szó. Ebből kifolyólag a teljes üzenet tartalma felhasználandó a feladat megoldás során. A módszer [13] kidolgozása mellett kettő gépi tanulási algoritmust is összehasonlítottam, melyeket alább ismertetek.

A módszer kidolgozásának elején egyszerű szó előfordulás (unigram) alapú felügyelt tanuló algoritmusokat alkalmaztam. Azonban, gyorsan nyilvánvalóvá vált, hogy szükség van előfeldolgozási lépésekre, melyekkel az előforduló szavak többalakúságát ki lehet küszöbölni. Az unigram jellemzőkön kívül olyan jellemzőket is vizsgáltam, melyek utalhatnak az adott szöveg polarítására. Ezen jellemzőkkel sikerült növelni az algoritmusok pontosságát. A megvalósítás során a *MALLET* nevezetű JAVA-alapú természetesnyelv-feldolgozó eszközt használtam [23].

3.1. Előfeldolgozás

Az üzenetekben az egyes szavak számos formában előfordulhatnak, például nagy- vagy kisbetűs alakban, rövidítve stb [28]. Ezeket a különböző formákat a tanuló algoritmus nem képes egységesen kezelni, ami rontja a működésének hatásfokát. Ezen felül az üzenetek számos olyan jelölést tartalmaznak, melyek a polarításra nézve nem relevánsak. Ezeket szem előtt tartva a következő előfeldolgozási lépéseket hajtottam végre.

- A szavak többalakúságának kiszűrésére kisbetűs formára hoztam őket, majd lecseréltem azokat szótöveikre. Utóbbit *Porter* szótövező algoritmussal hajtottam végre.

- Az 1.3.2 fejezetben említésre kerültek a tweetekben használható jelölések. Ezek nem hordoznak polarításra vonatkozó információkat, ezért egységesítettem őket. Minden felhasználónév-ementést és webcímet a *[USER]*, illetve az *[URL]* szövegekre cseréltem.
- Témakör jelölések esetében a # jel törlésre került. Például a #vásárlás a vásárlás szöveggé lett alakítva, ezáltal ezek is hagyományos szavakként lettek kezelve.
- Az előző jelölésekkel szemben hangulatjelek használhatók érzelmek kifejezésére. Ebből kifolyólag az ilyen jeleket *POZITÍV_JEL* és *NEGATÍV_JEL* szövegekre cseréltem. Néhány pozitív érzelmet kifejező hangulatjel: :) , :-), :D , ;) , illetve negatív: :(, : ((, :- (
- Hasonlóan a második pontban említett jelekhez, számok és egyéb írásjelek sem fejeznek ki érzelmet, ezért ezeket is egységesítettem. Minden számot a *[NUMBER]* szövegre cseréltem és a következő karaktereket töröltem: ' " # \$ % & () * + , / ; < = > \ ^ { } ~
- Az olyan szavak is egyszerűsítésre kerültek, melyek karakter többszöröződéseket tartalmaztak, azaz ugyanazt a karaktert legalább háromszor írták egymás után. Például az *iggggggen* szóból átalakítás után az *iggen* szó keletkezett. Ez által az ilyen szavak is egységes alakúak lesznek, de továbbra is tartalmazzák ezt az érzelmi töltetet.
- Végző lépésben stopszó szűrést végeztem, mely során a polarításra nézve hasztalan szavak törölve lettek. Az ilyen szavakat sok esetben előre összeállított listák alapján szűrik. Megoldásom során olyan szavakat tekintettem stopszónak, melyek nagyon sűrűn fordultak elő az egyes üzenetekben: a , hogy

3.2. Jellemzőkinyerés

Az üzenetek előfeldolgozása után kinyerhetők a különböző jellemzők, melyeket a tanuló algoritmus felhasznál. Ezek egy része a szövegben előforduló szavak. Ezen felül egyéb olyan indikátorokkal kísérleteztem, melyek utalhatnak az egyes szövegek polarítására:

- Egyik ilyen, a szövegben előforduló szavak polarítása, melynek meghatározására a *SentiWordNet* [4] lexikont használtam fel. Ebben minden szó minden jelentésére definiálva van annak pozitív, negatív és objektív értéke. Egy szó több jelentéssel is rendelkezhet, például *dob*. Egy mondat feldolgozása során nem ismerjük az abban

levő többalakú szavak pontos jelentését, így minden ilyen szó esetén annak polaritás értékeit a jelentések értékeinek átlagolásával határoztam meg. Ezután egy szót pozitívnak, negatívnak vagy objektívnak tekintettem, ha az adott polaritásokhoz tartozó érték rendre legalább 0.3, 0.2 vagy 0.8. Ezeket a küszöbértékeket empirikus vizsgálatokkal határoztam meg. Ezután minden egyes üzenethez három további jellemzőt definiáltam, mégpedig az üzenetben levő pozitív, negatív és objektív szavak száma.

- Ehhez hasonlóan, az üzenetekben található rövidítéseket is polaritás alapján csoportosítottam. A rövidítések kezeléséhez ismét egy lexikont¹ használtam, mely egy rövid leírást ad ezekhez. A leírásokban szereplő szavak polaritása alapján meghatározható az egyes rövidítések polaritása is. Így két újabb jellemzőt hoztam létre, mely az üzenetekben szereplő pozitív, illetve negatív rövidítések száma. Ebben az esetben az objektivitás kezelését mellőztem, mivel ez az érték nem megbízhatóan detektálható a leírások alapján.
- Az emberekre jellemző, hogy ha valamilyen érzelmet szeretnének kifejezni, akkor a szavakat nagybetűsen írják, vagy karaktertöbbszöröződésekkel használják. Ennél fogva további két jellemző, az ilyen szavak száma egy üzenetben.

3.3. Eredmények

A kidolgozott módszer a *Semantic Evaluation Exercises* 2013-ban megrendezett nyílt nemzetközi verseny keretein belül lett tesztelve [31]. A verseny célja egyrészt, hogy egy olyan szöveges adatbázist hozzanak létre, mely lehetővé teszi információ kinyerő rendszerek fejlesztését és tesztelését. Másrészt, a mai technikákkal létrehozható rendszerek minőségét is fel szeretnék volna tární. A kiértékeléshez használt adatbázis Twitter- és SMS-üzenetekből áll. Az adatokat egy év leforgása alatt gyűjtötték, melyek különböző entitásokkal (személyek, termékek, események) kapcsolatosak. Ily módon a tanító adatbázis közel 10, 000 tweetet tartalmazott, míg a két teszt adatbázis 3, 813 és 2, 094 Twitter-, illetve SMS-üzenetet. A megoldásokat tekintve, lehetőség volt ún. zárt és nyílt verziók létrehozására. Az első esetben különböző természetesnyelv-feldolgozásban használatos eszköz mellett csak a kiadott adatbázisok voltak használhatóak, míg az nyílt megoldás során egyéb forrásból szerzett szövegek is. A feladat megoldásai a precizitás és lefedettség alapú F-mértékkel lettek kiértékelve.

Munkám során zárt és nyílt megoldást is létrehoztam, utóbbi esetén további tweetek kerültek letöltésre és automatikus címkézésre. A Twitter mikroblog-szolgáltatás lehetővé

¹www.internetslang.com

teszi üzenetek letöltését az ún. *Twitter Streaming API* szolgáltatásán keresztül. A letöltött üzenetek címkézését egy másik szolgáltatás, a *Twitter Sentiment*² segítségével végeztem el, mely az üzeneteket polaritásuk alapján a már ismertetett három osztályba sorolja. Fontos megjegyezni, hogy az így meghatározott címkék egy része a használt rendszer pontatlanságából kifolyólag hibás lehet. 16 órás adatgyűjtés során 350,000 címkézett üzenetet sikerült letölteni. Annak ellenére, hogy mindkét szolgáltatás nyújt bizonyos szintű nyelvi szűrőt, nagy mennyiségű nem angol szöveges adat is letöltésre került. A teljes adathalmazból a végső kísérletekhez 10,000 angol üzenetet tartottam meg, melynek címke eloszlása egyenletes.

Az alábbiakban láthatók a Naïve Bayes és maximum entrópia osztályozók által elért eredmények az egyes teszt adatbázisokon. Mindkét osztályozó esetén a zárt és nyílt verziók is kiértékelésre kerültek. A rendszerek pontossága a különböző technikák fokozatos bevezetésével kerülnek ismertetésre.

A 3.1 és 3.2 táblázatokban láthatóak az egyes osztályozók által elért F-értékek a különböző polarításokra. Ebben az esetben jellemzőként csak szavak voltak használva, és az említett előfeldolgozási lépések sem kerültek használatra. A legjobb értékeket a semleges osztály esetében sikerült elérni, míg a legrosszabbat a negatív esetén. Ennek oka egyszerű, a tanító-adatbázisban negatív tanító példából volt a legkevesebb. Látható, hogy a nyílt verzióhoz letöltött külső adatok növelték a pontosságot mindkét teszt adatbázis esetében. Mivel előfeldolgozási lépések még nem történtek, a jellemzőkészlet hatalmas, és az osztályozók nem képesek a megfelelő általánosításra. További tanító példák használatával azonban növelhető a pontosság. Másrésztől észrevehető az is, hogy a Naïve Bayes rosszabb értékeket produkál a maximum entrópia osztályozónál, mely a későbbiekben is megfigyelhető lesz.

	pozitív	negatív	semleges	átlag
twitter-zárt	0.59	0.09	0.65	0.34
twitter-nyílt	0.60	0.17	0.65	0.38
sms-zárt	0.46	0.16	0.63	0.31
sms-nyílt	0.47	0.38	0.53	0.42

3.1. táblázat: Naïve Bayes osztályozó által elért F értékek Twitter és SMS teszt adatokon, előfeldolgozási lépések nélkül, jellemzőként csak szavakat használva

²<http://www.sentiment140.com>

	pozitív	negatív	semleges	átlag
twitter-zárt	0.60	0.33	0.67	0.46
twitter-nyílt	0.60	0.40	0.66	0.50
sms-zárt	0.47	0.31	0.69	0.39
sms-nyílt	0.52	0.47	0.66	0.49

3.2. táblázat: Maximum entrópia osztályozó által elért F értékek Twitter és SMS teszt adatokon, előfeldolgozási lépések nélkül, jellemzőként csak szavakat használva

A 3.3 és 3.4 táblázatokban a továbbfejlesztett módszer eredményei láthatók, mely során már az előfeldolgozási lépések is alkalmazásra kerültek. Ezen lépések láthatóan javították az eredményeket mindkét osztályozónál, főként a Twitter-adatbázison futtatott zárt kiértékelés esetében. Ezen kívül megfigyelhető, hogy a többlet adat a nyílt kiértékelések során a Twitter-adatbázison rontották a pontosságot. Az indok erre, hogy az egyes osztályozók az előfeldolgozásnak köszönhetően már a kisebb számú tanító példán is jól tudtak általánosítani, és a többlet adat, aminek automatikus címkézése nem teljesen pontos, csak zavartak az osztályozásban.

	pozitív	negatív	semleges	átlag
twitter-zárt	0.65	0.32	0.67	0.48
twitter-nyílt	0.62	0.21	0.63	0.41
sms-zárt	0.56	0.27	0.72	0.41
sms-nyílt	0.52	0.35	0.66	0.43

3.3. táblázat: Naïve Bayes osztályozó által elért F értékek Twitter és SMS teszt adatokon, előfeldolgozási lépésekkel, jellemzőként csak szavakat használva

	pozitív	negatív	semleges	átlag
twitter-zárt	0.66	0.40	0.68	0.53
twitter-nyílt	0.61	0.42	0.64	0.51
sms-zárt	0.61	0.38	0.77	0.49
sms-nyílt	0.57	0.47	0.72	0.52

3.4. táblázat: Maximum entrópia osztályozó által elért F értékek Twitter és SMS teszt adatokon, előfeldolgozási lépésekkel, jellemzőként csak szavakat használva

Végül a 3.5 és 3.6 táblázatokban a végső rendszer eredményei láthatók, melyek a szavakon felül a fent említett egyéb jellemzőket is felhasználták. Ezen jellemzők a Naïve

Bayes osztályozó eredményein nem tudtak javítani, kivéve a zárt SMS kiértékelést. Ezzel szemben a maximum entrópia osztályozó ki tudta használni a többlet információt, így növekedtek az F-mértékek.

	pozitív	negatív	semleges	átlag
twitter-zárt	0.65	0.32	0.67	0.48
twitter-nyílt	0.62	0.17	0.79	0.39
sms-zárt	0.56	0.38	0.74	0.47
sms-nyílt	0.54	0.29	0.70	0.41

3.5. táblázat: Naïve Bayes osztályozó által elért F értékek Twitter és SMS teszt adatokon, előfeldolgozási lépésekkel, szavakat és más polarításra utaló jellemzőket használva

	pozitív	negatív	semleges	átlag
twitter-zárt	0.66	0.41	0.69	0.54
twitter-nyílt	0.63	0.43	0.65	0.53
sms-zárt	0.62	0.39	0.79	0.50
sms-nyílt	0.61	0.49	0.75	0.55

3.6. táblázat: Maximum entrópia osztályozó által elért F értékek Twitter és SMS teszt adatokon, előfeldolgozási lépésekkel, szavakat és más polarításra utaló jellemzőket használva

Az eredmények elemzésével leszűrhetjük, hogy rövid üzenetek osztályozása során az előfeldolgozási lépések jelentősége nagy, főként ha kevés tanító példa áll rendelkezésre. Ezen felül olyan jellemzők vizsgálatával, melyek az üzenetek polarítására utalnak tovább növelhetőek a pontossági mutatók. Összehasonlítottam két felügyelt gépi tanuló algoritmust, melyek közül a maximum entrópia osztályozó minden esetben jobban teljesített.

Végezetül a 3.7 táblázatban látható a teljesség igénye nélkül más csapatok eredménye is, ahol a saját rendszer az *SZTE-NLP* nevet viseli. Az alapmegoldást adó módszer, minden üzenethez a tanító adatbázisban legtöbbet előforduló címkét rendeli. A 40 nevezett csapatból 21. és 8. helyezést sikerült elérni a zárt illetve nyitott Twitter adatbázison, valamint a 13. és 3. helyezést a zárt illetve nyílt SMS adatbázison. A legjobb eredményt az *NRC-Canada* érte el, egy az itt kidolgozott módszerhez hasonló rendszerrel, mely során számos további jellemzőt, illetve SVM osztályozót alkalmaztak [24].

	twitter-zárt	twitter-nyílt	sms-zárt	sms-nyílt
NRC-Canada	0.69	–	0.68	–
teregram	0.64	0.64	–	0.59
SZTE-NLP	0.54	0.53	0.51	0.55
Uom	0.51	0.45	0.42	0.35
uottawa	0.42	–	0.40	–
Alapmegoldás	0.29	0.29	0.19	0.19

3.7. táblázat: Más módszerek Twitter és SMS teszt adatbázison elért átlagos F értékei.
Kidolgozott rendszer: SZTE-NLP

4. fejezet

Entitásorientált véleménydetekció

A következő véleménydetekciós feladat során adott entitás tekintélyének követését szeretnénk elvégezni Twitter-üzenetek segítségével. Első lépés eldönteni, hogy az adott dokumentum a cél entitásra vonatkozóan hordoz-e véleményt. Adott tehát egy üzenet és egy entitás, a feladat egy bináris osztályozás, mely megadja, hogy a kérdéses üzenet az entitásra vonatkozik-e vagy sem. A következő lépés az előző feladathoz hasonló, a szűrésen átesett és megfelelőnek bizonyult üzenetekre el kell dönteni, hogy azok pozitív, negatív vagy semleges véleményt hordoznak-e.

Az itt kidolgozott módszer [14], az előző feladat megoldásának egy továbbfejlesztése, illetve az aktuális feladathoz alakított változata. Szűréshez és polaritás meghatározáshoz is ugyanazt a módszert használtam fel, csak más paraméterezettséggel, melyekről a 4.2 fejezetben található pontosabb leírás. Az alkalmazott osztályozó algoritmus maximum entrópia, mivel az előzőek alapján ez jobban teljesít, mint a Naïve Bayes. Az előfeldolgozási lépések megegyeznek a 3.1 fejezetben tárgyaltakkal, így itt nem kerülnek újra ismertetésre.

4.1. Jellemzőkinyerés

A fejlesztések főként a tanuló algoritmusok által eddig használt jellemzők feladat-specifikus átalakítását illetve bővítését érintettek, így a következőkben ezeket tekintjük át.

- Az üzenetekben található szavak mellett fontosak lehetnek bizonyos szókapcsolatok is. Vegyük például a *nem szeretem* kifejezést, amit ha két külön szóként kezelünk, elveszítjük azt a többlet információt, hogy a pozitív töltetű *szeretem* szó tagadva van. Ebből kifolyólag kísérleteztem szókettesekkel és szóhármassokkal (bigram és trigram), ami során kiderült, hogy a bigramok hasznosak az osztályozások szempontjából, míg a trigramok jelentősége nem szignifikáns, viszont a futási időt

jelentősen növelik. Ezért a módszerben újabb jellemzőkként a bigramok is használatra kerültek.

- Mivel a feladat entitásorientált osztályozás végzése, fontos az ezekkel kapcsolatos információk felhasználása is [15]. A szűrés során az entitás hiánya a szövegben sokszor jelezheti, hogy az üzenet valami mással kapcsolatos. Ezen felül inkább polaritás detektálásnál hasznos annak ismerete, hogy az emberek mely entításokat milyen mértékben kedvelik. Így két újabb jellemzőt alkottam meg, egyik egy bináris mutató, mely az entitás nevének jelenlétét/hiányát jelzi, másik pedig maga az entitás azonosítója. Az utóbbi jellemző és a tanító adatbázis segítségével egyfajta apriori tudás nyerhető ki az egyes entítások kedveltségét illetően.
- Fontos figyelembe venni az egyes szavak távolságát az adott entitás említésétől. Minél közelebb van egy szó az entitáshoz, annál nagyobb annak valószínűsége, hogy az adott szó az entításra vonatkozik. Tekintsük a következő példát:

A tegnap esti Omega koncert örületes volt, csak ne kellett volna ma ennyire korán kelnem. :(

melyben az említett entitás az *Omega*. A mondat az entításra vonatkozóan pozitív hangulatú, ami abból is látható, hogy a *örületes* szó közel van az *Omega* említéséhez, míg a mondat végén szereplő negatív hangulatjel távol. Ebből kifolyólag, súlyozott jellemzővektorok használatára tértem át. Az egyes szavakat a következők szerint súlyoztam:

$$\frac{1}{e^{\frac{1}{n}|i-j|}} \quad (4.1)$$

ahol n a mondat hossza, i és j az aktuális szó és az entitás említésének pozíciója.

- Kísérletet tettem az üzenetek témakör szerinti csoportosítására, mely segítségével az alkalmazott gépi tanulási algoritmus meghatározhatja az egyes témakörök átlagos polaritását. Ehhez a 2.2.1 fejezetben ismertetett *latent Dirichlet allokációt* alkalmaztam, mely minden üzenetre meghatározza, hogy az egyes témakörökbe mekkora valószínűséggel tartozik. A keresendő témakörök számát előre meg kell adni az algoritmusnak, így ezek optimális számának meghatározása további empirikus vizsgálatokat igényeltek. A témakör valószínűségeket mint újabb jellemzőket felhasználva, tovább sikerült javítani a módszer pontosságán.

Összesítve tehát a következő jellemzőkészlettel dolgoztam: unigramok és bigramok (távolság alapján súlyozva), szó- és rövidítéspolaritások, karaktertöbbszöröződést tartal-

mazó és nagybetűs szavak száma, entitás jelenléte a szövegben, üzenet mely entitáshoz tartozik és témakör információk.

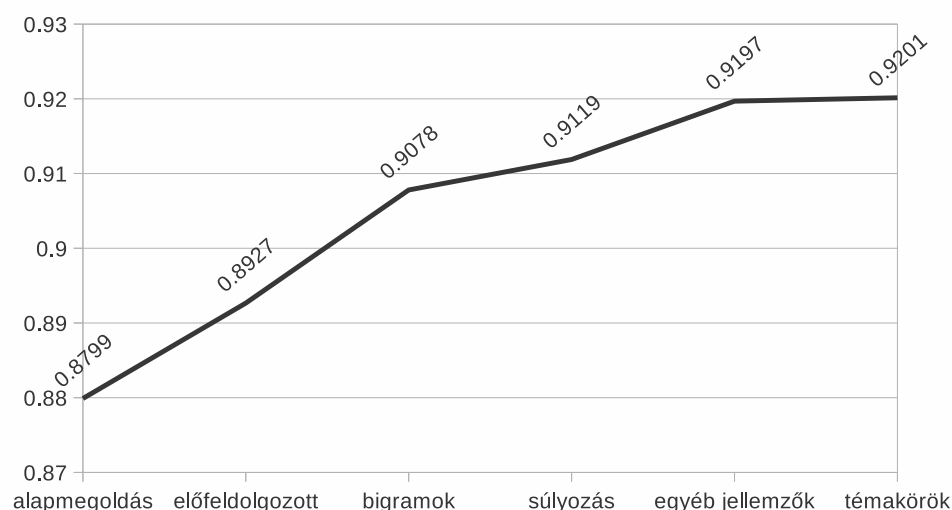
4.2. Eredmények

A kifejlesztett módszer az *An evaluation campaign for Online Reputation Management Systems* nevezetű nyílt nemzetközi verseny keretein belül létrehozott adatbázison lett kiértékelve [3]. A szervezők által biztosított adathalmaz 80%-a angol, míg a maradék spanyol nyelvű tweeteket tartalmazott. A tweetek 61 entitás alapján lettek begyűjtve, melyek az autó, bank, egyetem vagy zene témakörök egyikéhez tartoznak. Az üzenetek gyűjtését fél évig végezték, így entitásonként 700 tanító példa és 1500 teszt példa keletkezett. Ezen felül 50, 000 további címkézetlen üzenetet is rendelkezésre bocsájtottak, melyek az LDA futtatásához használtam fel. Az előző feladattal ellentétben a kiértékelés zárt volt, azaz csak a rendelkezésre álló üzeneteket lehetett használni. A rendszerek kiértékeléséhez a pontosság és a megbízhatóság és szenzitivitás alapján számított F-mérték metrikák voltak használatosak.

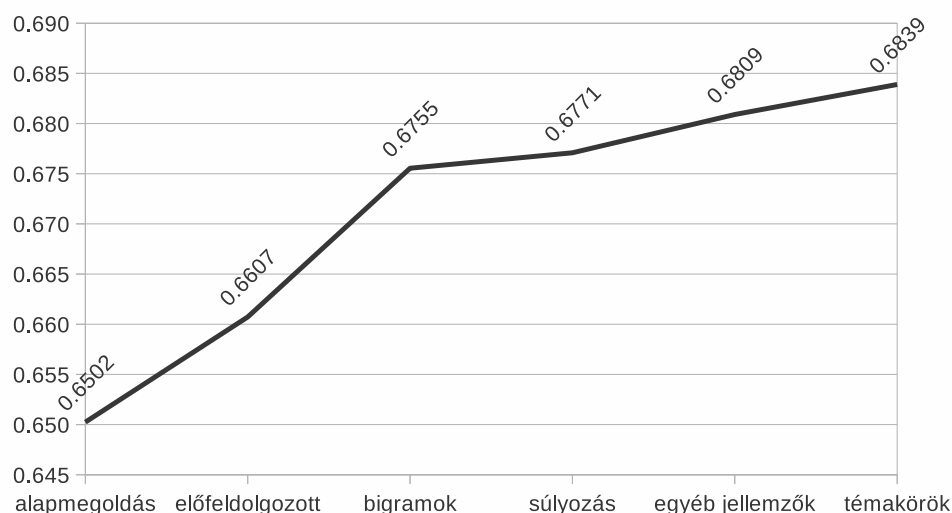
A 4.1 és 4.2 ábrákon láthatóak a szűrő és polaritás detektáló módszerek eredményei pontosság metrikával mérve a teszt adathalmazon. Az ábrák bemutatják, hogy hogyan javulnak az eredmények újabb és újabb jellemzők hozzáadásával. Az *alapmegoldás*nak nevezett verzió kizárólag unigram jellemzőket alkalmazott előfeldolgozási lépések nélkül. A következő lépés előfeldolgozási módszerek alkalmazása, amik szemmel láthatóan mindkét feladat esetében nagyban javították az elért eredményeket. Bigramok használatával tovább javultak az eredmények, főként a polaritás meghatározása során. A *súlyozás* nevezetű rendszerben bevezetésre kerültek a súlyozott jellemzővektorok, azaz az eddig használt jellemzők súlyozva lettek az entitástól vett távolsággal. Az *egyéb jellemzők* rendszerben az előző fejezetben bemutatott további jellemzők (a témakör-információkat kivéve) is használatra kerültek, melyek a szűrés feladatnál bizonyultak hasznosabbnak. Végül az utolsó verzióban 50 témakör detektálásából adódó információk is felhasználásra kerültek, melyek a szűrés feladatnál nem voltak szignifikánsak, ellentétben a polaritás meghatározással.

A 4.1 táblázatban láthatók az eredmények a témakörök számának függvényében, a tanító adatbázison tízszeres keresztvalidációval mérve. Az itt használt rendszer az összes eddig ismertetett jellemzőt és technikát alkalmazta. Ahogy az az előző diagramon is látható volt, a szűrés feladat esetében ez a jellemző nincs nagy hatással a pontosság metrikára. Megfigyelhető, hogy mindkét feladat esetében az F-mértéknél akár 1%-os, míg a pontosság esetében csak tizedes javulások voltak elérhetők.

Végül tekintsük át a hivatalos eredményeket ismét a teljesség igénye nélkül, melyek



4.1. ábra. Szűrés feladat pontossága a teszt adatbázison, x tengelyen a megoldás különböző verziói, kezdetben csak szavak használat, majd előfeldolgozás, bigramok, távolság alapú súlyozás, polarításra utaló jellemzők és LDA által meghatározott 50 témakör



4.2. ábra. Polaritásdetekció pontossága a teszt adatbázison, x tengelyen a megoldás különböző verziói, melyek egyre több és több technikát alkalmaznak

a 4.2 és a 4.3 táblázatokban találhatók. Ahogy az előző feladatnál a saját eredmények most is a *SZTE-NLP* név alatt találhatók meg. A verseny hivatalos kiértékeléshez több megoldást is lehetőség volt beküldeni. A szűrés feladat során 50 témakör használatával sikerült a legjobb eredményeket elérni. A metrika, mely alapján a csapatok rangsorolását végezték, ennél a feladatnál az F-mérték volt. Ez alapján a második helyezést sikerült elérni, viszont a pontossági metrika alapján első helyezett lett a kidolgozott rendszer. A polaritás meghatározása esetében a rangsorolást adó metrika a pontosság volt. Ebben

témakörök	szűrés pontosság	szűrés F	polaritás pontosság	polaritás F
0	0.919	0.380	0.680	0.368
20	0.920	0.387	0.680	0.370
50	0.919	0.391	0.682	0.375
100	0.920	0.388	0.683	0.379

4.1. táblázat: Eredmények különböző számú témakör használatával a tanító adatbázison mérve tízszeres keresztvalidációval

az esetben a 20 témakör használata adta a legmagasabb értékeket, így mindkét metrika alapján első helyezést sikerült elérni.

Az eredményeket megvizsgálva észrevehető, hogy mindkét feladat esetében a megfelelő előfeldolgozási lépések szignifikáns javulást hoznak. Ezen felül a különböző szókapcsolatok és egyéb polarításra utaló jelek tovább növelhetik a rendszerek működésének minőségét. Mivel a feladatok entitás alapú vizsgálat volt, így az ezen információkat kihasználó technikákat érdemes használni. Ezeken felül kísérleteztem témakörök meghatározásával, mely a polaritás detekció során hasznos többlettudást adnak.

Csapat	pontosság	megbízhatóság	szenzitivitás	F-mérték
POPSTAR	0.91	0.73	0.45	0.49
SZTE-NLP	0.93	0.60	0.44	0.44
LIA	0.87	0.66	0.36	0.38
UNED ORM	0.88	0.43	0.38	0.34
Alapmegoldás	0.87	0.49	0.32	0.33

4.2. táblázat: Más szűrő módszerek eredményei a teszt adatbázison, rangsorolás az F-mérték szerint

Csapat	pontosság	megbízhatóság	szenzitivitás	F-mérték
SZTE-NLP	0.69	0.48	0.34	0.38
POPSTAR	0.64	0.43	0.34	0.37
LIA	0.65	0.50	0.15	0.19
UNED ORM	0.62	0.36	0.10	0.15
Alapmegoldás	0.58	0.32	0.29	0.30

4.3. táblázat: Más polaritás detektáló módszerek eredményei a teszt adatbázison, rangsorolás a pontosság szerint

5. fejezet

Aspektusorientált véleménydetekció

Ahogy az a bevezetőben említésre került, bizonyos esetben nem elegendő az adott entitásra vonatkozót globális vélemény kinyerése. Fontos lehet azok egyes tulajdonságaira (aspektusaira) vonatkozó különböző polaritású információk detektálása. Könnyedén megeshet, hogy egy étterem értékelése során az étellel meg vagyunk elégedve, viszont a pincérek kedvességével már nem. A következőkben bemutatásra kerül erre a célra kidolgozott módszer, mely során adott egy mondat és egy abban szereplő aspektus, mely polaritását kell meghatározni. A feladat megoldása során figyelembe kell venni azt a tényt, hogy egy mondaton belül adott entitás akár több aspektusát is említhetjük. Vegyük a következő egyszerű példát:

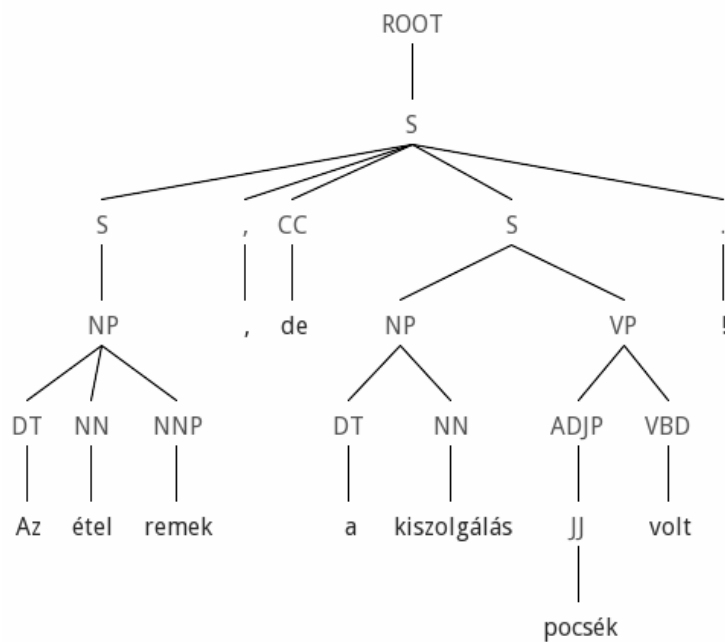
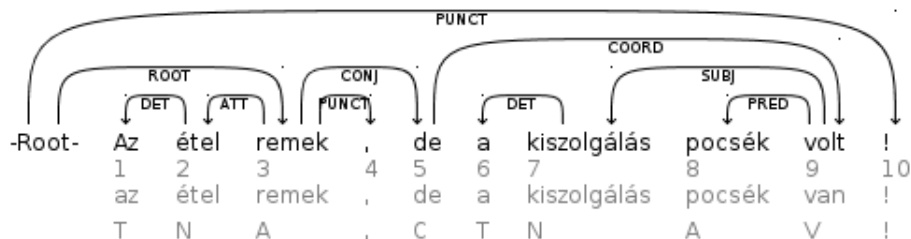
Az új autóm fogyasztása kiváló, de a végsebessége igen alacsony.

A mondatban a *fogyasztás* és a *végsebesség* az *autó* két említett tulajdonsága, melyekkel kapcsolatban pozitív, illetve negatív véleményt fejtettünk ki. Látható, hogy szükség van a mondat megfelelő felbontására, és ezen szövegrészek külön kezelésére, annak érdekében, hogy az egyes aspektusokról a megfelelő véleményt nyerhessük ki.

A módszer kidolgozása során továbbra is maximum entrópia osztályozó eljárás került alkalmazásra. A lentebb részletezett adathalmaz ellentétben az eddigiekkel nem Twitter-üzenetkből áll, hanem annál jóval szabályosabb szövegekből. Az egyes mondatok részsövegekre bontását különböző nyelvi elemzők segítségével végeztem el.

5.1. Nyelvtani elemzés alapú jellemzők

Szövegek nyelvtani elemzése során célunk a mondatokon belül, a szavak közti kapcsolatok azonosítása. Két különböző szintaktikai elemző egy *függőségi* és egy *konstituencia elemző* kimenetelét is felhasználtam. Mindkét módszer felügyelt tanuló algoritmusokon alapul. Függőségi elemzés (5.1 ábra) során egy ún. irányított függőségi fa képezhető.



a fában legfeljebb öt hosszú irányítatlan úton elérhető az adott aspektusra vonatkozóan tekintetem. Hasonló megfontolásból adódik egy másik jellemző, mely már nem aspektus specifikus, de a polarításra nézve fontos. Már említésre került, hogy az egyes szavak polarítása meghatározható lexikon segítségével. Ha egy szó feje pozitív, illetve negatív polarítású, egy újabb jellemzőt hozok létre ennek indikálására. Egy szó polarításának meghatározására továbbra is a SentiWordNet lexikon lett használva, de egy pontosabb módszerrel, mely alább részletesebben bemutatásra kerül.

Kísérletet tettem arra is, hogy ne csak bizonyos szavak kerüljenek felhasználásra nyelvi elemzők segítségével, hanem tagmondatok is. Ehhez a konstituenselemző eredményeit használtam fel, oly módon, hogy a teljes fa csak egy részét használtam fel. A részfa kiválasztásának módszere a következő. Kezdetben a részfa gyökere az adott aspektust tartalmazó levél. Ezek után a fát bővítjük, oly módon, hogy a következő gyökér elem az aktuális szülője legyen. Ezt addig végezzük, míg az alábbi feltételek mindegyike teljesül:

- A részfának kötelezően tartalmaznia kell a szóban forgó aspektust
- Legalább öt levéllel kell rendelkeznie a túl rövid tagmondatok kiszűrése érdekében
- Nem tartalmazhat más aspektust, kivéve ha az ötnél közelebb távolságra van a kérdéses aspektustól. w_i és w_j szavak közötti távolság $|i - j|$, ahol i és j az egyes szavak pozíciója a mondatban.
- Az aktuális gyökér elem nem tagmondatot jelöl a fában (PP vagy S)

A megfelelő részfa megtalálása után, a benne szereplő levelek alapján unigramokat és bigramokat használtam mint jellemzők. Az adatbázis segítségével, melyen a kísérletek folytak, polaritást jelző szavak egy csoportja is kigyűjtésre került, χ^2 statisztika segítségével. Az eljárás olyan szavakat válogat ki, melyeknek pozitív vagy negatív véleményt tartalmazó mondatokban való előfordulása magas. Ezáltal egy újabb jellemző a kiválasztott részfában előforduló, polaritást jelző szavak száma. Az előzőtől független módszer segítségével [21] egy adott részfa teljes polarítását is meghatároztam, melyet újabb jellemzőként használtam.

5.2. Egyéb jellemzők

A nyelvtani szerkezetet kihasználó jellemzőkön felül, klaszterezéssel előállítható jellemzők is használatra kerültek. Az aspektusokat témakör alapján 10 csoportba soroltam, az LDA-hoz hasonló módszerrel. A módszer során minden egyes aspektushoz tartozik egy

jellemző vektor, mely azon szavakat tartalmazza, melyek az adott aspektussal egy mondatban előfordulnak. Így előáll egy $n \times m$ méretű mátrix, ahol n az aspektusok, m pedig a különböző szóalakok száma. A könnyebb kezelhetőség érdekében a mátrixot dimenzióredukációs technikával $n \times 50$ méretűre alakítottam. Erre a célra *szingulárisérték-felbontást* alkalmaztam. Az így előálló kisebb méretű mátrixot *k-közép* eljárással 10 csoportba soroltam, melyek az egyes témaköröket reprezentálták. A *k-közép* eljárás során koszinusz távolság lett alkalmazva. Ezután jellemzőként használtam fel a kategóriát, melybe az adott aspektus tartozik.

A fent említett jellemzőkön túl az eddig ismertettek is felhasználásra kerültek. Ezek egyrészt unigramok és bigramok súlyozva az aspektustól vett távolságtól. Fontos megjegyezni, hogy a teljes mondat alapján készített ilyen típusú jellemzők különböznek azoktól, melyek a kiválasztott rész konstituensfa alapján lettek létrehozva. Továbbá minden unigram és bigram esetén azok tagadása is jelezve lett. Egy mondatban a negált szavak meghatározása kézzel gyűjtött tagadó szavak (*nem*, *sem*) és a tagadást megszüntető jelek (*"*, *"*, *de*, *és*) segítségével történt. Mindezen felül a adott mondatban szereplő pozitív, negatív és objektív szavak száma is jellemzőként lett használva.

5.3. Szavak jelentésének meghatározása

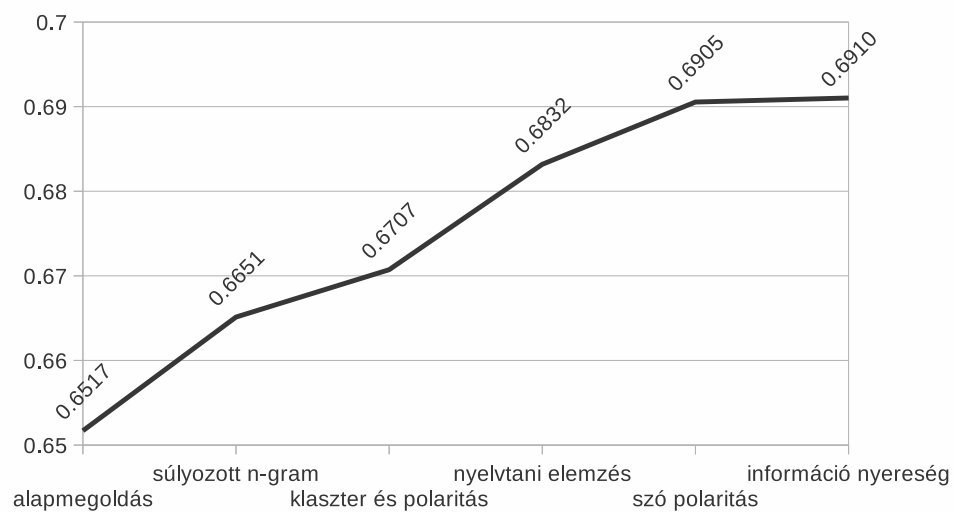
Ahogy az már említésre került, az egyes szavak polaritását meghatározó módszer finomításra került. Az eddigiekben egy szóhoz tartozó polaritás értékek a jelentések megfelelő értékeinek átlagaként lett kezelve. Kísérletet tettem az egyes szavak jelentésének meghatározására, és így az átlagolás kiküszöbölésére. A módszer megértéséhez tekintsünk egy példamondatot. *Ez a körte finom és édes*, melyben a *körte* szó jelentését keressük. A *SentiWordNet* lexikon segítségével meg kell határozni az összes lehetséges jelentést, melyek most legyenek a *gyümölcs* és *izzó*. Ezután azt a jelentést kell választani, ami az adott környezetben legnagyobb valószínűséggel előfordul, tehát *gyümölcs finom* vagy *izzó finom*. Ezeket a valószínűségeket internetről gyűjtött nagy dokumentumhalmaz segítségével becsülhetjük [10]. Miután meghatároztuk az adott szóra a jelentést, a szóban forgó lexikon segítségével megkaphatjuk annak polaritás értékeit is.

5.4. Eredmények

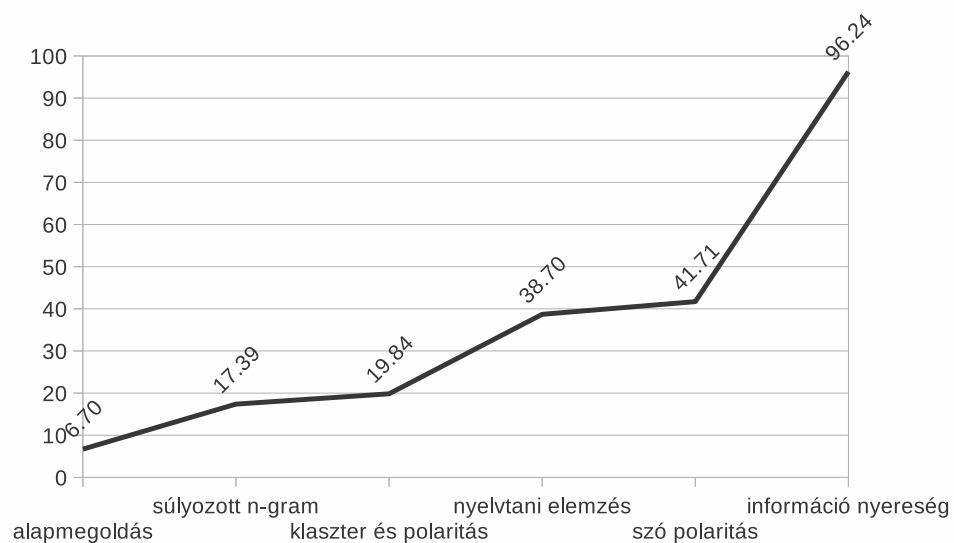
A kidolgozott módszer a *Semantic Evaluation Exercises* 2014-ben megrendezett nyílt nemzetközi verseny [26] keretein belül lett kiértékelve. A szervezők által létrehozott adathalmaz angol nyelvű termék értékeléseket tartalmaz, 3,000 mondatot éttermekről és ugyanennyit laptopokról. Minden egyes mondat esetében adott a benne szereplő aspektu-

sok egy listája, melyekre adott a polaritásuk. A már megszokott 3 polaritás érték mellett egy negyediket is alkalmaztak, mely az *ellentmondó*, mely olyan aspektusok esetében lett használva, melyekre pozitív és negatív vélemény is érkezett. Így a teljes adatbázis 6, 051 jelölt aspektust tartalmaz.

A 5.3 és 5.4 ábrákon láthatóak az elért eredmények. Az első ábrán a módszer által elért pontossági metrika értékei láthatók tízszeres keresztvalidációval mérve a tanító adatbázison, míg a második a felügyelt tanuló algoritmus futási idejét másodpercenként. A diagramok x tengelyén ismét a rendszerek láthatók, melyek egyre több és több technikát alkalmaznak. Az *alapmegoldás* a teljes mondat alapján kinyert unigram és bigram jellemzőket használja súlyozatlanul. Egyetlen előfeldolgozási lépést alkalmaztam, mégpedig a szavak kisbetűsítését. Mivel az itt használt szövegek szabályos formájúak, melyek esetében az előfeldolgozó lépések nem játszanak nagy szerepet, sőt számos esetben ronthatják is az eredményeket. Ennek oka, hogy a különböző nyelvtani elemző rendszerek megfelelő működéséhez szükség van az egyes szavak ragozott formájára, például a mondat tárgyának megtalálásában nagy szerepet játszik a szó t végződése. Az így elért pontosság 65%, a futási idő pedig 10 másodperc alatt van. A következő rendszer a *súlyozott n -gram*, ahol már az eddigi jellemzők súlyozásra is kerülnek. A *klaszter és polaritás* rendszer, már felhasználja az aspektusok témakör szerinti csoportosításából adódó információkat és a polarításra utaló szavak számát is. A következő lépésben (*nyelvtani elemzés*) a függőségi és konstituensfa alapján kinyerhető jellemzők kerültek felhasználásra. A *szó polaritás* magába foglalja a polaritás lexikonból kinyerhető jellemzőket is, tehát minden eddig bemutatott technikát. Az így elért eredmény 69%, a futási idő pedig 41 másodperc. Az utolsó rendszerben mindezen felül, információnyereség-statisztikán alapuló jellemzőkiválasztás is tesztelésre került, mely során a nagyságrendileg 116, 000 különböző jellemzőből a 2, 000 legkevésbé releváns nem került használatra. Látható, hogy a pontosság javulása nem szignifikáns, de a futási idő nagyban megnőtt. Érdeemes megemlíteni még, hogy a teljes jellemzők készlet kinyerése az adathalmazból 10 perc.



5.3. ábra. Pontossági értékek a tanító adatbázison tízszeres keresztvalidációval. X tengelyen a módszerek, melyek egyre több technikát alkalmaznak.



5.4. ábra. Az egyes módszerek futási ideje másodpercben.

Hasonlóan az előző feladatokhoz, egy 1, 788 jelölt aspektust tartalmazó teszt adatbázis segítségével összehasonlítások történtek más megoldások eredményeivel. A versenyre nevezett 30 csapatból 3. és 6. helyezést sikerült elérni a laptop, illetve étterem teszt adatbázisokon. Ismét a teljesség igénye nélkül az értékek a 5.1 táblázatban láthatók külön az éttermekkel és laptopokkal kapcsolatosan.

Csapat	étterem	laptop
DCU	0.81	0.70
SZTE-NLP	0.75	0.67
USF	0.73	0.64
Alapmegoldás	0.61	0.51

5.1. táblázat: Más módszerek eredményei a pontosság metrikával mérve az étterem és laptop teszt adatbázisokon

6. fejezet

Kapcsolódó munkák

Az elmúlt néhány évben az emberek véleményének kinyerése szöveges tartalmakból aktívan kutatott területté vált annak gazdasági jelentősége miatt. Megmutatták, hogy valóban hasznos információk nyerhetők ki a szociális médiából, melyek korrelációban állnak a kérdőívek által gyűjtött adatokkal [25]. Ezenkívül azt is felfedezték, hogy az emberek megbíznak mások (akár idegenek) termékekről alkotott véleményben, és ezáltal befolyásolhatók vásárlási szokásaik [11].

A rendelkezésre álló adatok segítségével számos témakörben gyűjthetünk információkat. A termékek népszerűségének nyomon követése egy jelentős részét képezi a véleménydetekciós kutatásoknak. Ezek alapjául a legtöbb esetben mikroblog bejegyzések szolgálnak. Megmutatták, hogy a bejegyzések 19%-a említést tesz valamilyen termékről, és ezek 20%-a tartalmaz szubjektív véleményt is [16]. Ehhez szorosan kapcsolódó feladat a részvények árainak követése. A pontos árak meghatározása nehéz lenne, de relaxálható egy kétszintű feladattá, ahol az árak változásának irányát kell meghatározni [29]. A termékek népszerűségének követésén felül számos más területen kiaknázzhatóak ezek a módszerek. Nyomon követhető különböző entitások, mint például hírességek vagy politikai pártok támogatottsága is. Ehhez egy egyszerű módszer lehet az adott entitást említő bejegyzések számának meghatározása [27]. Azonban egy bejegyzés tartalmának polaritását is figyelembe kell venni, annak érdekében, hogy a megfelelő döntést hozzuk. A szövegek polaritásának kizárólagos meghatározásán túl, komplexebb vizsgálatokat is végezhetünk. Szabályos szövegeken jól használhatóak olyan módszerek, melyek különböző szemantikai kapcsolatokat is fel tudnak fedni. Ezeket felhasználva meghatározható a konkrét vélemény egy szövegben, annak birtokosa, de akár a témakör is, melyben a véleményt alkották [20].

A számítástechnika gyors fejlődésének köszönhetően megnyílt a lehetőség számításgépes módszerek széles körű használatára. Egyre inkább előtérbe kerülnek az adatbányászati, gépi tanulás alapú módszerek. Mivel legtöbb esetben osztályozási feladatról

beszélünk, használhatunk felügyelt vagy felügyelet nélküli tanuló algoritmusokat. Empirikus vizsgálatok alapján ezek a módszerek jól működnek rövid bejegyzéseken is. Ha rendelkezünk tweetek egy adathalmazával, melyben minden szöveghez adott annak polaritása, akkor egy felügyelt tanuló algoritmus jó választás a feladat megoldására [1]. Az eredmény nagyban függ a szövegekből kibányászott jellemzőktől, melyek alapján az algoritmus megtanulja a különböző mintázatokat. Ilyen jellemzők lehetnek egészen egyszerűen csak a dokumentumban levő szavak, de akár összetettebbek is, mint például az egyes szavak szófajai vagy nyelvtani szerepük a mondatokban. Az ilyen módszerek egyik hátránya, hogy nagy mennyiségű tanító példára van szükségük a megfelelő általánosításra, melyek kézi összeállítása költséges lehet. Bizonyos esetekben tanító példákat automatikusan is gyűjthetünk más rendszerek segítségével [5], viszont ilyenkor számolni kell az adatok bizonyos mértékű pontatlanságával. Ahogy az már említésre került, véleménydetekció feladata, hogy időben kövesse az emberek véleményét különböző entitásokra nézve. Statikus tanító adatbázisok használata esetén egyrészt elveszítjük ezt az időbeliséget, másrészt nem kezeljük trendek változását sem. Ha rendelkezünk időbélyegekkal az egyes adatokra, akkor ezt felhasználva szimulálhatjuk az időbeliséget, például úgy, hogy az egyes dokumentumok osztályozásánál csak a korábbi adatokat használjuk fel [7].

Számos olyan adatbázist hoztak létre, melyek felhasználhatók véleménydetekciós kutatásokhoz. Ezekben az adatbázisokban levő szövegek különböző részletezettséggel lehetnek jelölve. Amennyiben a feladat, hogy a teljes dokumentum, illetve az abban levő entitás alapján nyerünk ki véleményeket dokumentum szintű jelölés is elegendő [31, 3]. Számos esetben azonban szükség van mondatok részletesebb jelölésére, például termékek tulajdonságaira vonatkozó információk indikálásra [26], vélemények és azok közlőinek vagy célpontjainak meghatározására [30, 19]. Ezen adatbázisok publikusan elérhetőek és felhasználhatóak különböző módszerek összehasonlítására.

Dolgozatomban olyan felügyelt és felügyelet nélküli módszereket mutattam be, melyek a fent említett technikák alapötleteit is felhasználva osztályozzák a dokumentumokat polaritásuk alapján. A módszerek megfelelő működéséhez előfeldolgozási és szűrési feladatokat is el kellett végezni a még pontosabb eredmények eléréséhez.

7. fejezet

Összefoglalás

A dolgozatban bemutatásra került a véleménydetekció mint kutatási terület és annak néhány gyakorlati alkalmazása. A kidolgozott módszerek segítségével látható volt, hogy nagyon fontos a dokumentumok megfelelő előkészítése, főként ha azok közel valós idejű kommunikációból származnak [28]. Három, egyre részletesebb információ kinyerési feladat segítségével prezentálásra kerültek a különböző szinteken használható technikák. Dokumentum szintű információ kinyerés során egyfajta globális véleményre vagyunk kíváncsiak, mely során a teljes dokumentumot felhasználjuk azok polaritásának kiderítéséhez [13]. Entitás orientált véleménydetekció során [14, 15], egy nulladik lépésként, érdemes a dokumentumok szűrését elvégezni, melyre a polaritás detekcióhoz hasonló módszerek használhatók. Láttuk, hogy ebben a lépésben az adott témakör, melybe a dokumentum tartozik nem releváns, ellentétben a polaritás meghatározásával. Annak vizsgálata azonban, hogy mely szavak tartoznak a kérdéses entitáshoz, már igenis fontos. Tovább részletezve a feladatot, az egyes entitások különböző tulajdonságaira vonatkozó információk kinyerésével is próbálkozhatunk. Bemutattam egy módszert termékek tulajdonságaihoz tartozó jó és rossz vélemények automatikus meghatározására, melyben nagy hangsúlyt kaptak az egyes tulajdonságokra vonatkozó szövegrészek azonosítása mondatok szintaktikai elemzése alapján.

A szociális média hatalmas adatmennyiséget szolgáltat az ilyen kutatások elvégzéséhez. Azonban sokszínűsége miatt, az információk számos formában jelenhetnek meg. A dolgozatban egy a napjainkban közkedvelt forrás, a Twitter került használatra. Az innen származó ún. tweetek nagyon zajos szövegek, aminek egyik következménye, hogy számos, a természetesnyelv-feldolgozásban használatos eszköz csak nehezen alkalmazható elemzésükre. Egy sokkal szabályosabb információ forrás a termékértékelő oldalak. Ilyen szövegek kezelésére már nagyobb pontossággal alkalmazhatóak ilyen eszközök, melyek felhasználási módjára is láthattunk példát a dolgozatban. A kidolgozott módszerek a tesztadatbázisokon kimagasló eredményeket értek el.

Irodalomjegyzék

- [1] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. Sentiment Analysis of Twitter Data. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 30–38, June 2011.
- [2] E Amigó, J Gonzalo, and F Verdejo. Reliability and sensitivity: Generic evaluation measures for document organization tasks. Technical report, Tech. rep., UNED, 2012.
- [3] Enrique Amigó, Jorge Carrillo de Albornoz, Irina Chugur, Adolfo Corujo, Julio Gonzalo, Tamara Martín, Edgar Meij, Maarten de Rijke, and Damiano Spina. Overview of RepLab 2013: Evaluating Online Reputation Monitoring Systems. In *Fourth International Conference of the CLEF initiative*, 2013.
- [4] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 2010.
- [5] Luciano Barbosa and Junlan Feng. Robust Sentiment Detection on Twitter from Biased and Noisy Data. In *Poster volume*, Coling 2010, pages 36–44, August 2010.
- [6] Adam L. Berger, Stephen Della Pietra, and Vincent J. Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71, 1996.
- [7] Albert Bifet and Eibe Frank. Sentiment Knowledge Discovery in Twitter Streaming Data. 2010.
- [8] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.
- [9] Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pages 161–168. ACM, 2006.

- [10] Hakan Ceylan and Rada Mihalcea. An efficient indexer for large n-gram corpora. In *ACL (System Demonstrations)*, pages 103–108. The Association for Computer Linguistics, 2011.
- [11] Chrysanthos Dellarocas. The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management science*, 49(10):1407–1424, 2003.
- [12] Wenjing Duan, Bin Gu, and Andrew B Whinston. Do online reviews matter?—an empirical investigation of panel data. *Decision Support Systems*, 45(4):1007–1016, 2008.
- [13] Viktor Hangya, Gabor Berend, and Richárd Farkas. Szte-nlp: Sentiment detection on twitter messages. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 549–553, 2013.
- [14] Viktor Hangya and Richárd Farkas. Filtering and polarity detection for reputation management on tweets. In *Working Notes of CLEF 2013 Evaluation Labs and Workshop*, 2013.
- [15] Viktor Hangya and Richard Farkas. Target-oriented opinion mining from tweets. In *Cognitive Infocommunications (CogInfoCom), 2013 IEEE 4th International Conference on*, pages 251–254. IEEE, 2013.
- [16] Bernard J. Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury. Twitter Power: Tweets as Electronic Word of Mouth. In *Journal of the American society for information science and technology*, pages 2169–2188, 2009.
- [17] Daniel Jurafsky and James H Martin. An introduction to natural language processing, computational linguistics, and speech recognition, 2000.
- [18] Andreas M Kaplan and Michael Haenlein. Users of the world, unite! the challenges and opportunities of social media. *Business horizons*, 53(1):59–68, 2010.
- [19] Jason S. Kessler, Miriam Eckert, Lyndsie Clark, and Nicolas Nicolov. The 2010 icwsm jdpa sentiment corpus for the automotive domain. In *4th International AAAI Conference on Weblogs and Social Media Data Workshop Challenge (ICWSM-DWC 2010)*, 2010.
- [20] Soo-Min Kim and Eduard Hovy. Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, pages 1–8. Association for Computational Linguistics, 2006.

- [21] Angeliki Lazaridou, Ivan Titov, and Caroline Sporleder. A bayesian model for joint unsupervised induction of sentiment, aspect and discourse representations. In *Proceedings of ACL*, 2013.
- [22] Bing Liu. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167, 2012.
- [23] Andrew Kachites McCallum. Mallet: A machine learning for language toolkit. <http://mallet.cs.umass.edu>, 2002.
- [24] Saif M Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. *arXiv preprint arXiv:1308.6242*, 2013.
- [25] Brendan O’Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, May 2010.
- [26] Maria Pontiki, Dimitrios Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the International Workshop on Semantic Evaluation, SemEval ’14*, 2014.
- [27] Erik Tjong Kim Sang and Johan Bos. Predicting the 2011 Dutch Senate Election Results with Twitter. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 53–60, April 2012.
- [28] Viktor Varga, Vilmos Wieszner, Viktor Hangya, Veronika Vincze, and Richard Farkas. Magyar nyelvű webes szövegek számítógépes feldolgozása. In *X. Magyar Számítógépes Nyelvészeti Konferencia*, pages 327–331, 2014.
- [29] Tien Thanh Vu, Shu Chang, Quang Thuy Ha, and Nigel Collier. An experiment in integrating sentiment features for tech stock prediction in twitter. In *24th International Conference on Computational Linguistics*, page 23, 2012.
- [30] Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210, 2005.
- [31] Theresa Wilson, Zornitsa Kozareva, Preslav Nakov, Sara Rosenthal, Veselin Stoyanov, and Alan Ritter. SemEval-2013 Task 2: Sentiment Analysis in Twitter. In

Proceedings of the International Workshop on Semantic Evaluation, SemEval '13, June 2013.

[32] Harry Zhang. The optimality of naive bayes. *A A*, 1(2):3, 2004.

Nyilatkozat

Alulírott Hangya Viktor programtervező informatikus MSc szakos hallgató, kijelentem, hogy a dolgozatomat a Szegedi Tudományegyetem, Informatikai Tanszékcsoport Számítógépes Algoritmusok és Mesterséges Intelligencia Tanszékén készítettem, programtervező informatikus MSc. diploma megszerzése érdekében.

Kijelentem, hogy a dolgozatot más szakon korábban nem védtem meg, saját munkám eredménye, és csak a hivatkozott forrásokat (szakirodalom, eszközök, stb.) használtam fel.

Tudomásul veszem, hogy diplomamunkámat a Szegedi Tudományegyetem Informatikai Tanszékcsoport könyvtárában, a helyben olvasható könyvek között helyezik el.

Szeged, 2014. május 16.

.....

aláírás

Köszönetnyilvánítás

Ezúton szeretnék köszönetet nyilvánítani témavezetőmnek Dr. Farkas Richárdnak a dolgozat készítéséhez nyújtott segítségéért. Jelen kutatási eredmények megjelenését az „Infokommunikációs technológiák és a jövő társadalma (FuturICT.hu)” című, TAMOP-4.2.2.C-11/1/KONV-2012-0013 számú projekt támogatta. A projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg.